Information Theory for Risk-based Water System Operation

Steven Weijs

Information Theory for Risk-based Water System Operation

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. ir. K.C.A.M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen

op vrijdag 1 april 2011 om 12:30 uur

door

Steven Vincent WEIJS

civiel ingenieur
geboren te Groningen

Dit proefschrift is goedgekeurd door de promotor:
Prof. dr. ir. N. C. van de Giesen

Samenstelling promotiecommissie:

Rector Magnificus	voorzitter
Prof. dr. ir. N. C. van de Giesen	Technische Universiteit Delft, promotor
Prof. Dr. rer.nat. Dr.-Ing. A. Bárdossy	Universität Stuttgart
Prof. dr. ir. D. Koutsoyiannis	National Technical University of Athens
Prof. dr. ir. H. H. G. Savenije	Technische Universiteit Delft
Prof. dr. D. P. Solomatine	UNESCO-IHE
Dr. ir. P. J. A. T. M. van Overloop	Technische Universiteit Delft
Dr. ir. F. Pianosi	Politecnico di Milano

This research was performed at the section Water Resources Management, faculty of Civil Engineering & Geosciences of TU Delft and has been financially supported by Delft Cluster.

Keywords: operational water management, control, optimization, information theory, probabilistic forecasts, risk.

cover: A medal design by Gottfried Wilhelm Leibniz included in a letter in 1697 to Rudolph August, Duke of Brunswick, celebrating his discovery of the binary system: "for all to spring from nothing, a oneness suffices". The version used is from Rudolf Nolte in 1734. The second, turbulent figure is a fractal from the Mandelbrot set.

Published by VSSD, Delft, The Netherlands
ISBN 978-90-6562-264-8

This thesis was written using LyX and LaTeX
Printed in The Netherlands

See `www.hydroinfotheory.net` for more background information and updates.

To my parents

In a predestinate world, decision would be illusory;
in a world of a perfect foreknowledge, empty,
in a world without natural order, powerless.
Our intuitive attitude to life implies
non-illusory, non-empty, non-powerless decision. ...
Since decision in this sense excludes both perfect foresight and anarchy in nature,
it must be defined as choice in face of bounded uncertainty.
– (George Shackle, Decision Order and Time in Human Affairs, 1961)

Preface

Information is a quantity, measurable in bits. Any piece of information can be expressed in zeros and ones and any calculation can be performed by a simple computer reading only zeros and ones. Gottfried Wilhelm Leibniz was justifiedly excited when in 1697 he discovered the binary system and designed the medal for the Duke of Brunswick, depicted on the front cover, to celebrate his discovery: "For all to spring from nothing, a oneness suffices"

While the research leading to this thesis started as a study of water system operation under uncertainty, it soon became an investigation into the role of information in that process. Because information and uncertainty are quantities measured in the same unit, the topic essentially remained the same, but with a more positive sound to it. I admit that information theory in a broad sense has now come close to being my pet theory of the universe, which might of course bias my view on its importance. This is also visible in the contents of this thesis, whose central four chapters focus on the application of information theory.

> *The law that entropy always increases, holds, I think, the supreme position among the laws of Nature. If someone points out to you that your pet theory of the universe is in disagreement with Maxwell's equations - then so much the worse for Maxwell's equations. If it is found to be contradicted by observation - well, these experimentalists do bungle things sometimes. But if your theory is found to be against the second law of thermodynamics I can give you no hope; there is nothing for it but to collapse in deepest humiliation.* - Arthur Stanley Eddington, The Nature of the Physical World (1928)

In support of my pet theory, note that the second law of thermodynamics, celebrated in the above quote, can be interpreted information-theoretically as a loss of information into microscopical degrees of freedom that we usually cannot observe or control. Thermodynamics can therefore be seen as a specific application of information theory. The second law also defines the direction of time, which inevitably runs out before all questions are answered.

New unanswered questions are also an inevitable by-product of answers, making curiosity an addiction. This thesis therefore ends with an ellipsis rather than a period, because I do not think its submission will cure my addiction. I would also like to apologize for the unanswered questions that might feed the curiosity of its readers.

Summary

Risk-based water system operation can be formulated as a problem of rational decision making with incomplete information, which can be approached using the interlinked fields of probability, decision and information theory. This thesis presents a perspective from information theory, by focusing on selected issues for which this theory can help understanding how information flows or should flow from observation to decision.

Water system operation is the task of finding a sequence of decisions in time to influence a water system in order to optimally benefit from it and minimize the risks, using real-time information. An example is the operation of a hydropower reservoir, where the daily releases should maximize the power production benefits, but balance this against flood risks downstream by offering sufficient flood-storage, given forecasted inflows and present state.

Uncertainty in forecasts plays an important role in the operation of water systems and negatively affects the performance of their operation. Conversely, information reduces uncertainty and therefore has a value for operation. Information theory offers a rigorous framework to study information and uncertainty as quantities, but its application is not widespread in water resources management. In this thesis, the risk based operation of water systems is studied, with a specific focus on the role of information in this process. Because probabilistic forecasts are often the interface between observations, model outcomes and decisions, their evaluation is a key component of studying the flow of information. The information-theoretical perspective results in a number of new practical techniques for the evaluation of forecasts, merging information from different sources, and using the information in sequential decision processes.

The overall aim of this thesis is to develop methods for optimal risk based water system operation. During the research, Shannon's information theory was found to be a central framework to study risk and decision making in the context of modeling and water system operation procedures. This results in the following main research questions:

1. How does information play a role in optimal operation of water systems?
2. How can information be exploited optimally in decisions?

The nature of the research questions requires a theoretical approach to the problem. This thesis combines fundamental results from the interlinked fields of information theory, control theory and decision theory and applies them to problems in water resources management, while maintaining an integrated view on the nature of information.

The outcomes of the research can be subdivided into two levels. Firstly, at the conceptual level, the research tries to make a contribution to the debate about uncertainty analysis philosophy and methods, which are important topics in current hydrological literature. Secondly, at the applied level, the thesis presents a number of practical methods and recommendations.

The conceptual contributions result from taking an information-theoretical perspective on a number of open problems in hydrology and water resources management, notably the evaluation of probabilistic forecasts and the calibration of models. It is shown why forecasting should be seen as a communication process, where information is transferred from the forecaster to the user, in order to reduce the uncertainty of the latter. The quality of such forecasts can be measured using the expected Kullback-Leibler divergence from the observations to the forecasts. This "divergence score" can be interpreted as the remaining uncertainty of the user, after he has received the forecast. A new mathematical decomposition of the divergence score is found that can be interpreted as "the remaining uncertainty is equal to the initial uncertainty, minus the correct information, plus the wrong information." The correct information is bounded by the information in the predictors, while the wrong information must be minimized in the model calibration process. The decomposition is also used to demonstrate why deterministic forecasts are theoretically unacceptable and should be replaced by probabilistic forecasts.

Furthermore, the roles of parsimony and purpose of models are analyzed in the context of model calibration. Algorithmic information theory is used to shed light on the connection between science and data compression and the link between the amount of information and model complexity is discussed in this context. It is argued that the purpose of a model should play no role in the choice of calibration objective, because it can both lead to a reduction in the information that reaches the model from the data and to the model learning from information that is not existing.

When information in probabilistic predictions is used to support sequential decision processes, such as finding the sequence of optimal releases from a reservoir, the complex time-dynamics of information become important. In that case, the information that will become available in the future has an influence on the current optimal decision. Although the information itself is not available for the current decision, it is possible to take into account the fact that information will become available and benefits future decisions. Because doing this explicitly is computationally intractable, approximate solutions are necessary that estimate the future value of water, which must be balanced against the benefits of immediate use.

The more practical contributions of this thesis include: a method to generate long lead time weighted ensemble streamflow forecasts, based on information from El-Niño; a score for verification of probabilistic forecasts that rewards maximum information extraction from predictors; guidelines for optimization horizons in controller design for sequential decision processes; and a method to study the increase in value of water due to more informative predictions.

Samenvatting

Risicogestuurd operationeel waterbeheer is te formuleren als een zoektocht naar rationele beslissingen, gegeven onvolledige informatie. Hierbij kunnen de onderling verbonden disciplines kansrekening, beslistheorie en informatietheorie worden gebruikt. Dit proefschrift beoogt een informatie-theoretisch perspectief te bieden, door een aantal problemen te belichten waarvoor informatietheorie een bijdrage kan leveren aan het begrip en de verbetering van de informatiestroom van observatie naar beslissing.

Operationeel waterbeheer behelst het vinden, gebruikmakend van actuele informatie, van een reeks opeenvolgende beslissingen, die een watersysteem zodanig beïnvloeden dat er maximaal van geprofiteerd kan worden, terwijl risico's zo klein mogelijk worden gehouden. Een voorbeeld is het aansturen van gemalen in poldersystemen, waarbij een voldoende hoog waterpeil moet worden gehandhaafd voor onder andere de landbouw, maar inundatie moet worden voorkomen door op tijd te pompen, daarbij rekening houdend met de huidige systeemtoestand en anticiperend op eventueel voorspelde neerslag.

Onzekerheden in voorspellingen hebben een belangrijke invloed op het operationeel waterbeheer en verminderen de beheersbaarheid van watersystemen. Omgekeerd geldt ook dat informatie onzekerheid terugdringt en daardoor een waarde krijgt voor waterbeheer. Informatietheorie biedt een wiskundig kader voor informatie en onzekerheid als meetbare grootheden, maar is nog weinig toegepast in het waterbeheer. In dit proefschrift wordt het risicogestuurd beheer van watersystemen onderzocht, waarbij de nadruk ligt op de rol die informatie hierbij speelt. Omdat kansvoorspellingen vaak de verbinding vormen tussen waarnemingen, modeluitkomsten en beslissingen, speelt hun evaluatie een sleutelrol bij het onderzoeken van de informatiestromen. De informatie-theoretische blik resulteert in nieuwe methoden voor het evalueren van voorspellingen, het samenbrengen van informatie uit verschillende bronnen en het gebruik van informatie in sequentiële beslissingsprocessen.

Het achterliggende doel van dit onderzoek is het ontwikkelen van methoden voor optimaal risicogestuurd waterbeheer. Tijdens het onderzoek bleek de informatietheorie van Shannon een centraal kader voor het onderzoek naar risico's en beslisproblemen in de context van modelleren en operationeel waterbeheer. Dit levert de volgende onderzoeksvragen op:

1. Welke rol speelt informatie bij optimaal operationeel beheer van watersystemen?
2. Hoe kan informatie optimaal worden benut voor beslissingen?

De aard van de onderzoeksvragen vraagt om een theoretische onderzoeksaanpak. Dit proefschrift combineert een aantal fundamentele resultaten uit de informatietheorie, meet- en regeltechniek en beslistheorie en past deze toe op problemen in het waterbeheer, terwijl de aard van informatie steeds referentiekader en rode draad blijft.

De resultaten zijn te onderscheiden in twee abstractieniveaus. Op het eerste, conceptuele niveau levert dit onderzoek een bijdrage aan de discussie over methoden en filosofie van onzekerheidsanalyse van modellen. Deze discussie speelt een grote rol in de hedendaagse hydrologische wetenschappelijke literatuur. Op het tweede, toegepaste niveau, presenteert dit proefschrift een aantal praktisch toepasbare methoden en aanbevelingen.

De conceptuele bijdragen volgen uit een informatie-theoretische kijk op een aantal openstaande problemen in de hydrologie en het waterbeheer, met name de evaluatie van kansvoorspellingen en de kalibratie van modellen. Er wordt aangetoond waarom het doen van voorspellingen als een communicatieproces beschouwd moet worden, waar informatie wordt overdragen van de voorspeller op de gebruiker, om diens onzekerheid te verminderen. De kwaliteit van zulke voorspellingen kan worden gemeten met de verwachtingswaarde van de Kullback-Leibler divergentie van de waarnemingen tot de voorspellingen. Deze zogenaamde "divergence score" is te interpreteren als de resterende onzekerheid van de gebruiker, nadat deze de voorspelling heeft ontvangen. Er wordt een nieuwe wiskundige decompositie van de *divergence score* gepresenteerd, die in woorden kan worden geïnterpreteerd als "de resterende onzekerheid is gelijk aan de initiële onzekerheid, min de juiste informatie, plus de onjuiste informatie." De juiste informatie is gelimiteerd tot de informatie die aanwezig is in de voor de voorspelling gebruikte gegevens, terwijl de onjuiste informatie moet worden geminimaliseerd in de modelkalibratie. De decompositie wordt ook gebruikt om aan te tonen waarom deterministische voorspellingen theoretisch gezien onacceptabel zijn en vervangen moeten worden door kansvoorspellingen.

Verder wordt gekeken naar de rol van parsimonie en van het doel van een model in de context van modelkalibratie. Algoritmische informatietheorie wordt gebruikt om de analogie tussen wetenschap en datacompressie te verhelderen en vanuit deze context wordt het verband tussen de hoeveelheid informatie in gegevens en de complexiteit van modellen beschouwd. Er wordt betoogd dat het doel van een model in principe geen rol zou mogen spelen in de kalibratie, omdat dit tot een verlies van informatie uit de data kan leiden of juist tot het leren van niet bestaande informatie.

Wanneer de informatie uit kansvoorspellingen wordt gebruikt om sequentiële beslisprocessen zoals reservoirbeheer te ondersteunen, komt de complexe tijd-dynamica van informatie in beeld. In dat geval beïnvloedt de informatie die in de toekomst beschikbaar zal komen de optimale waarde van de huidige beslissing. Hoewel deze informatie zelf nog niet beschikbaar is voor die beslissing, is het wel mogelijk rekening te houden met het feit dat toekomstig beschikbare informatie toekomstige beslissingen verbetert. Omdat het qua rekenkracht niet haalbaar is om dit expliciet te doen, zijn benaderingen nodig om de toekomstige waarde van water in te schatten, om deze af te wegen tegen de opbrengsten van onmiddellijk gebruik.

De meer praktisch georiënteerde bijdrage van dit proefschift omvat: een methode om gewogen ensemble-voorspellingen op seizoen-tijdschaal te doen op basis en informatie over "El Niño"; een score ter evaluatie van kansvoorspellingen die het maximaliseren van de informatie beloont; ontwerprichtlijnen voor optimalisatie-horizons van regelaars voor sequentiële beslisprocessen; en een methode om de waardetoename van water als gevolg van informatievere voorspellingen in te schatten.

Contents

Preface vii

Summary ix

Samenvatting xi

Contents xiii

List of Figures xix

List of Tables xxi

1 Introduction 1
- 1.1 *Why operate water systems?* . . . 2
- 1.2 *Uncertainty, rationality and risk in decision making* . . . 3
  - 1.2.1 Uncertainty . . . 3
  - 1.2.2 Rationality . . . 4
  - 1.2.3 Risk . . . 4
  - 1.2.4 Risk-based water system operation is rational . . . 5
- 1.3 *Information reduces uncertainty* . . . 6
- 1.4 *The value of information* . . . 7
- 1.5 *Objectives and research questions* . . . 7
  - 1.5.1 The broader perspective . . . 7
  - 1.5.2 Open questions . . . 8
  - 1.5.3 Research objectives . . . 8
  - 1.5.4 Research questions . . . 9
- 1.6 *Thesis outline* . . . 9

2 Risk-based water system operation 11
- 2.1 *Introduction* . . . 11
  - 2.1.1 Water system operation as a mathematical optimization problem . . . 12
  - 2.1.2 Formulation of a control problem . . . 12
  - 2.1.3 Solution techniques . . . 14
- 2.2 *An example: a lowland drainage system* . . . 16
  - 2.2.1 Delfland . . . 17
  - 2.2.2 The Delfland decision support system . . . 18
  - 2.2.3 Objective of control . . . 18

2.3 *Off-line optimization of operation* . . . 20
2.3.1 Example: the "regelton" . . . 21
2.3.2 Disadvantages compared to online optimization . . . 23
2.4 *Online optimization of operation by model predictive control* . . . 23
2.5 *Uncertainties affecting the Delfland system* . . . 24
2.5.1 Uncertainties . . . 24
2.5.2 Data driven inflow model . . . 25
2.6 *Relevant time horizons for uncertainty and optimization* . . . 26
2.6.1 Time horizons relevant for prediction and control . . . 27
2.6.2 The time horizons for the Delfland system . . . 28
2.6.3 Results . . . 33
2.7 *Certainty equivalence* . . . 34
2.8 *Multiple model predictive control (MMPC)* . . . 35
2.9 *The problem of dependence on future decisions* . . . 36
2.10 *Summary and Conclusions* . . . 37

3 Uncertainty or missing information defined as entropy 39
3.1 *Uncertainty and probability* . . . 39
3.2 *The uncertainty of dice: Entropy* . . . 40
3.3 *Side information on a die: Conditional Entropy* . . . 43
3.3.1 Conditioning, on average, reduces entropy . . . 43
3.3.2 Conditional entropy . . . 44
3.4 *Mutual Information and Relative Entropy* . . . 45
3.5 *Rolling dice against a fair and ill-informed bookmaker* . . . 47
3.6 *Interpretation in terms of surprise* . . . 49
3.6.1 Surprise and meaning in a message . . . 50
3.6.2 Can information be wrong? . . . 50
3.7 *Laws and Applications* . . . 51
3.7.1 Information cannot be produced from nothing . . . 51
3.7.2 Information never hurts . . . 52
3.7.3 Given the information, uncertainty should be maximized . . . 52
3.7.4 Applications of information theory . . . 52
3.8 *Relation to thermodynamics* . . . 53
3.9 *Applications of information theory in water resources research* . . . 55

4 Adding seasonal forecast information by weighting ensemble forecasts 57
4.1 *Introduction* . . . 57
4.1.1 Use of ensembles in water resources . . . 58
4.1.2 Weighted ensembles . . . 60
4.1.3 Previous work on adding forecast information to climatic ensembles by weighting . . . 61
4.2 *Information, Assumptions and Entropy* . . . 62
4.2.1 The principle of maximum entropy . . . 63
4.3 *The Minimum Relative Entropy Update* . . . 63
4.3.1 Rationale of the method . . . 63

4.3.2 Formulation of the method . . . 64
4.4 *Theoretical test case on a smooth sample and comparison to existing methods* 65
4.4.1 Results in a theoretical test case . . . 67
4.4.2 Discussion . . . 73
4.4.3 Conclusions from the theoretical test case . . . 76
4.5 *Multivariate case* . . . 77
4.6 *Application to ESP forecasts* . . . 77
4.6.1 Seasonal forecast model . . . 79
4.6.2 Results . . . 80
4.6.3 Discussion . . . 84
4.6.4 Conclusion . . . 85
4.7 *Conclusions and recommendations* . . . 85

5 Using information theory to measure forecast quality 87
5.1 *Introduction* . . . 87
5.2 *Definition of the divergence score* . . . 89
5.2.1 Background . . . 89
5.2.2 Definitions . . . 90
5.2.3 The divergence score . . . 90
5.2.4 Decomposition . . . 91
5.2.5 Relation to Brier score and its components . . . 94
5.2.6 Normalization to a skill score . . . 94
5.3 *Relation to existing information-theoretical scores* . . . 96
5.3.1 Relation to the ranked mutual information skill scores . . . 96
5.3.2 Equivalence to the Ignorance score . . . 97
5.3.3 Relation to information gain . . . 98
5.4 *Generalization to multi-category forecasts* . . . 98
5.4.1 Nominal category forecasts . . . 98
5.4.2 Ordinal category forecasts . . . 99
5.4.3 The Ranked divergence score . . . 101
5.4.4 Relation to Ranked Mutual Information . . . 101
5.4.5 Information and useful information . . . 102
5.5 *An example: rainfall forecasts in the Netherlands* . . . 103
5.6 *Generalization to uncertain observations* . . . 106
5.6.1 Introduction . . . 106
5.6.2 Decomposition of the divergence score for uncertain observations . . . 108
5.6.3 Expected remaining uncertainty about the truth: the cross-entropy score . . . 109
5.6.4 Example application . . . 111
5.6.5 Discussion: divergence vs. cross-entropy . . . 113
5.7 *Deterministic forecasts cannot be evaluated* . . . 114
5.7.1 Deterministic forecasts are implicitly probabilistic (information interpretation) . . . 115

5.7.2 Deterministic forecasts can still have value for decisions (utility interpretation) . . . 117
5.8 *Conclusions* . . . 118

6 Some thoughts on modeling, information and data compression 121
6.1 *Introduction* . . . 122
6.1.1 The principle of parsimony . . . 122
6.1.2 Formalizing parsimony . . . 123
6.2 *Algorithmic information theory, complexity and probability* . . . 124
6.2.1 The Bayesian perspective . . . 125
6.2.2 Universal computability . . . 126
6.2.3 Kolmogorov complexity, patterns and randomness . . . 127
6.2.4 Algorithmic probability and Solomonoff induction . . . 127
6.2.5 Computable approximations to automated science . . . 128
6.3 *The divergence score: prediction, gambling and data compression* . . . 129
6.3.1 Dependency . . . 131
6.4 *A practical test: "Zipping" hydrological time series* . . . 132
6.4.1 Data and Methods . . . 133
6.4.2 Results . . . 136
6.4.3 Compressing with hydrological models . . . 139
6.4.4 Discussion and conclusions . . . 140
6.5 *Prediction versus understanding* . . . 141
6.5.1 Hydrological models approximate emergent behavior . . . 141
6.5.2 Science and explanation as data compression? . . . 142
6.5.3 What is understanding? . . . 144
6.6 *Modeling for decisions: understanding versus utility* . . . 145
6.6.1 Information versus utility as calibration objective . . . 145
6.6.2 Locality and philosophy of science: knowledge from observation . . . 145
6.6.3 Utility as a data filter . . . 147
6.6.4 Practical example . . . 150
6.7 *Conclusions and recommendations for modeling practice* . . . 151

7 Stochastic dynamic programming to discover relations between information, time and value of water 155
7.1 *Introduction* . . . 155
7.2 *Stochastic dynamic programming (SDP)* . . . 156
7.2.1 Formulation of a typical hydropower reservoir optimization problem using SDP . . . 157
7.2.2 Computational burden versus information loss . . . 158
7.3 *Example case description* . . . 159
7.3.1 Simulation and re-optimization . . . 160
7.4 *Optimization and the value of water* . . . 162
7.4.1 Water value in the example problem . . . 163
7.5 *Interdependence of steady state solution and real-time control* . . . 165
7.6 *How predictable is the inflow? - entropy rate and the Markov-property* . . . 167

7.7 *The influence of information on the marginal value of water* 168
7.8 *Sharing additional benefits of real-time information between stakeholders* . 170
7.9 *Reinforcement Learning to approximate value functions* 172
7.10 *Conclusions and recommendations* 174

8 Conclusions and recommendations **175**
8.1 *Conclusions at the conceptual level* 175
8.1.1 The nature of information 175
8.1.2 The flow of information 176
8.1.3 The value of information 177
8.1.4 The necessity of probabilistic predictions 177
8.1.5 Information theory as philosophy of science 178
8.2 *Methodological contributions and recommendations for practice* 179
8.2.1 On risk based water system operation 179
8.2.2 On weighted ensemble forecasts 179
8.2.3 On the evaluation of probabilistic forecasts 179
8.2.4 On performance measures for model inference 180
8.2.5 On optimization of reservoir release policies 180
8.3 *Limitations and recommendations for further research* 181
8.3.1 Going from discrete to continuous 181
8.3.2 Expressing prior information 181
8.3.3 Merging information theory with statistical thermodynamics 182
8.3.4 An integrated information-theoretical framework from observation to decision 182
8.3.5 Problem solved, but the solution is a problem 183

References **185**

A Equivalence between MRE-update and pdf-ratio solutions for the normal case **199**

B The decomposition of the divergence score **201**

C Relation divergence score and doubling rate in a horse race **203**

Acknowledgements **205**

About the author **207**

Publications **209**

List of Figures

1.1 Control increases flexibility of the Q-h relation of a weir 2
1.2 Information is the reduction in uncertainty 6

2.1 A schematic representation of a "polder-boezem" system 17
2.2 Variability of water levels within the Delfland boezem canals 20
2.3 Setpoints and simulated water levels for water butt optimized off-line . . . 22
2.4 Schematic representation of model predictive control (MPC) on a real system 24
2.5 Rainfall-runoff model found by linear system identification 26
2.6 Increasing uncertainty in inflow modeled as second order Markov chain . . 28
2.7 Two methods for comparison of observed and forecast rainfall 30
2.8 Time-decreasing skill of Delfland rainfall forecasts 30
2.9 MPC simulations for different prediction horizons with perfect forecasts . . 32
2.10 MPC simulation with real imperfect forecasts 33
2.11 MPC performance against prediction horizon with perfect and actual forecasts 34
2.12 Schematic representation of MMPC. 36
2.13 Two MMPC formulations with different assumptions on future information availability 38

3.1 Expected number of questions to know the outcome of a die 41
3.2 Observing the dice can yield a further conditioning model. 44
3.3 Venn diagrams depicting the relations between information measures . . . 46

4.1 Various climatic and conditional ensembles related to ESP forecasting . . . 59
4.2 Equally weighted ensemble members can represent a nonuniform density. This density can be changed by shifting or by weighting the ensemble traces. 60
4.3 Resulting ensemble weights for μ_1=3, σ_1=0.5 and resulting empirical CDF 67
4.4 Resulting ensemble weights for μ_1=4, σ_1=0.5 and resulting empirical CDF 67
4.5 Resulting ensemble weights for μ_1=3, σ_1=1.2 and resulting empirical CDF 68
4.6 Resulting ensemble weights for μ_1=4, σ_1=1.2 and resulting empirical CDF 68
4.7 Pareto-front showing trade-off between lost information and lost uncertainty 74
4.8 Weights and CDF for MRE-update with skewness constraint 76
4.9 Bubble plots of weights for MRE-update in the multivariate case 78
4.10 Bivariate kernel density estimate of the joint distribution of the ENSO index (November-February) and the average streamflow (April-August). . . 81
4.11 An example of the kernel density estimate used in the pdf-ratio method for the year 2003 (hindcast mode). 81

4.12 Ensemble weights for 2003, plotted against the average streamflow of the each trace from April to September (hindcast mode). 81
4.13 Forecast CDFs for the different weighting methods 83
4.14 The RPSS for the weighted ensemble forecasts by the different weighting methods for the whole forecast period (hindcast mode).. 83
4.15 The RPSS for the weighted ensemble forecasts by the different weighting methods for the period from 1970 (forecast mode, starting from 20 traces). 83

5.1 The components of the divergence score as additive bars 92
5.2 Comparison of the uncertainty component in Brier and divergence scores . 95
5.3 Comparison of the resolution component in Brier and divergence scores . . 96
5.4 Comparison of the reliability component in Brier and divergence scores . . 97
5.5 Skill against lead-time for Dutch probability of precipitation forecasts . . . 104
5.6 Sensitivity of the components for rounding of the forecast probabilities . . 105
5.7 The relation between Brier and divergence scores is not monotonic. 107
5.8 The additive component bars for the case of uncertainty in observations . 111
5.9 Uncertain binary observation for Gaussian measurement error 112
5.10 Decomposition for uncertain observations applied to KNMI data 113

6.1 Temporal dependence reduces the uncertainty of a guess 132
6.2 The effect of quantization on the hydrological signal 134
6.3 The effect of lossy compression on the signal. It introduces a noise but is not significantly more than the quantization noise. 138
6.4 Compression vs. signal to noise ratio for JPG on hydrological data 138
6.5 Local vs. non-local scores: DS vs. (C)RPS 146
6.6 Schematic representation of how a model learns from data. Information reaches the model through three routes, and can be filtered through a utility function. 148
6.7 Model behavior in validation of utility-calibrated versus information-calibrated parameters 150

7.1 Schematic illustration of the value-to-go calculation in SDP 158
7.2 The storage volume-area-head relation for the reservoir and the storage discretization using the Savarenskiy scheme. 161
7.3 The discretization of the flow in 5 equiprobable classes for each month . . . 161
7.4 Reservoir behavior (level, releases, spills, power) for the simulated policy. . 162
7.5 Immediate water value and opportunity cost as a function of the release. . 164
7.6 Water values from the re-optimization for the different months of the year. 165
7.7 The reservoir behavior as a function of the month of the year. 165
7.8 Marginal values of water as a function of storage, inflow and month. 166
7.9 Mutual information due to temporal dependence in the inflow process. . . 169
7.10 Shift in Pareto front and negotiated solution due to real-time operations. . 171
7.11 The agent-environment interaction in reinforcement learning. 173

List of Tables

3.1 Joint distributions expressing side information for two models of a die . . . 47

4.1 Overview of the methods and types of forecasts that are compared in chapter 4 . . . 66
4.2 Resulting tercile probabilities for the four methods compared . . . 68
4.3 Resulting mean and standard deviation for the various methods . . . 69
4.4 Resulting relative entropy for the different methods . . . 69
4.5 The resulting skill score for the different methods. . . . 84

6.1 Analogy science ⇔ data compression and physical systems ⇔ computation 122
6.2 Optimal code lengths proportional to minus the log of events' probability . 130
6.3 Compression performance for well-known compression algorithms on various time series . . . 137
6.4 Information-theoretical and variance statistics and compression results for rainfall-runoff modeling . . . 139
6.5 Validation and calibration results for information and utility objectives . . 151

7.1 Characteristics of the hydropower reservoir toy model. . . . 160

Chapter 1

Introduction

> *"Only a fool tests the depth of the water with both feet."*
> - African proverb

Water systems are all around us: our ancestors emerged from one; we drink from them; we feed from them; we fear to be submerged in them; we ship goods over them; we store heat in them; we generate power from them; we drain them; we replenish them; we rinse in them; we pollute them; we try to harness them and preserve them for our children, who, like us, are water systems themselves.

Although water, which plays such an important role in our lives and life in general, lends itself perfectly to long-winding and poetic descriptions, this thesis will try to solidify some of its fluidity and will look at water mostly in dry and mathematical terms. This by no means indicates an under-appreciation of the beauty that water enables and possesses. On the contrary, appreciation of complex systems and their many interrelated subsystems that all have their functions is only enhanced by understanding them. Understanding amounts to linking observations of quantities in the water system by describing them in a compact and precise manner that enables making sensible predictions. The system-view is useful because it helps to formalize the thinking about behavior, interaction with the surroundings, interactions with us, boundaries, modeling, division into subsystems, inputs and outputs to the system. This facilitates making predictions, which are of paramount importance in both science and engineering. In science predictions are the only way to test our theories and in engineering they enable decisions that improve our quality of life.

This thesis is specifically focused on water systems that can be influenced by humans and that are interesting for us to influence. Examples are artificial lakes, rivers, aquifers, polder systems and irrigation systems. All these systems have in common that we can exert an influence on them, for example through dams, dikes, wells, pumps and weirs. The construction and installation of these structures helps harness these systems and make them behave in a way that is beneficial to mankind. Once in place, the functionality of these structures can be enhanced by operating them in an optimal manner according to our objectives.

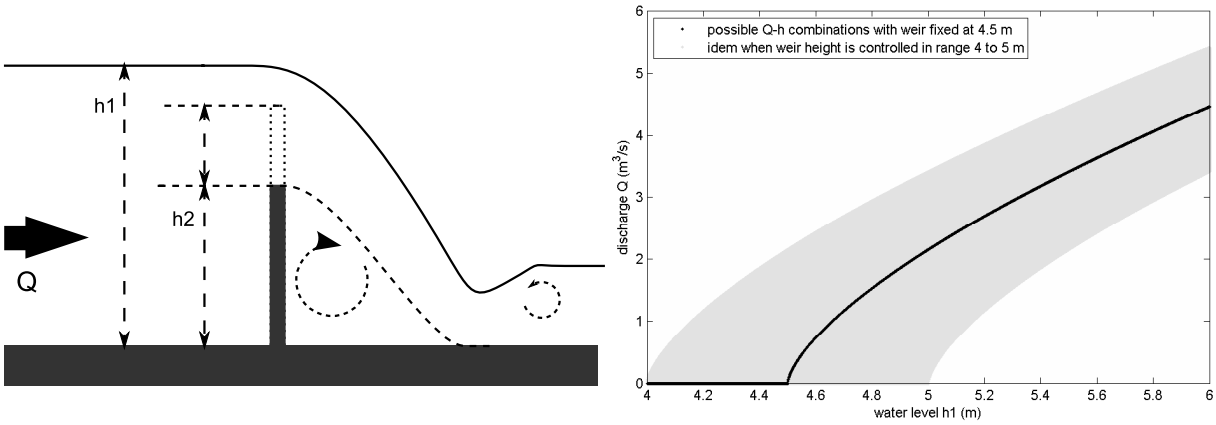


Figure 1.1: The relation between discharge and water level is more flexible if a controllable weir is used. The gray area shows the possible combinations for a controllable weir, while the black line gives the combinations attainable with a fixed weir.

1.1 Why operate water systems?

In contrast to structural measures, such as the construction of new infrastructure like dams, weirs and pumps, operational measures do not aim to *alter* the water system permanently. Instead, they are *using* existing infrastructure, to have a temporary influence on the system's behavior to cope with the specific situation at hand. Switching on a pumping station in a polder or releasing water from a hydropower reservoir are examples of operational measures. As with many classifications, the distinction is not clear-cut and depends on the time-scale of reference.

Often, water system operation is needed, in addition to structural measures, to achieve a good and cost-effective performance of the water system. This need arises from the fact that, apart from our influence, the water system is influenced by external forcings, such as the weather, which typically vary in time. Therefore, the best response will often also be time-varying. Although it is possible to achieve a time-varying response by structural measures (a weir, for example, "reacts" on a higher water level by letting more water pass), operation provides much more flexibility (extra water can be released by lowering the weir when needed, even when the water level is still low). The key difference is that operational measures can vary in time, depending on objectives that change in time and on information beyond the local and present situation. The processing of this information into an action or decision is extremely flexible. See for example figure 1.1, where the relation between discharge and upstream water level for a weir is plotted for a free flowing weir. The downstream water level is considered to be low enough not to influence the flow. The equation that describes the flow is

$$Q = k\,(h_1 - h_2)^{\frac{3}{2}} \tag{1.1}$$

where Q is the flow over the weir, k is a constant and h_1 and h_2 are the water levels indicated in figure 1.1. The range of possible flows for a given water level or the range of possible water levels for a given flow are far larger for a controllable weir.

Structural and operational measures complement each other. A well-designed dam, built at the right location, enhances the possibilities to benefit from the water system, but only if it is thoughtfully operated. Conversely, the room for operation is very small if the structures through which the water system can be influenced do not have sufficient capacity. Therefore, optimal decisions about operational measures have to take into account the constraints posed by the infrastructure. Furthermore, the effects of future operation need to be considered already at the stage where structures are planned and designed. Because of this interdependency, decisions on structural and operational measures interact.

This thesis is purely concerned with decisions on operational measures, referred to as water system operation. The design of the structures and the range of possible actions to control the water system will be considered as given and not as part of the optimization problem. The focus will be on water system operation under uncertainty and the role information plays in that process. Later in this thesis, the close link between information and uncertainty will be further analyzed. The next sections give a short introduction in the related concepts of information, uncertainty and risk in the context of finding optimal decisions, such as optimal actions for operating the structures to influence a controlled water system.

1.2 Uncertainty, rationality and risk in decision making

1.2.1 Uncertainty

Almost every decision in real life is influenced by uncertainty. In some cases this uncertainty is more obvious than in others. A famous example of a decision under uncertainty is the decision to take an umbrella or not, based on the weather forecast. In case it is forecast that it will rain with certainty, the obvious decision is to take an umbrella. In another case, when the sky is blue and the air pressure is high, it would be a better decision not to carry the useless umbrella around (assuming pure water management and no fashion objectives). In many cases, however, the decision maker does not know with certainty whether it will rain. In such a case, his decision depends on how much he values remaining dry, how annoying carrying the umbrella is and his estimate of the probability that it will rain.

An example in which the uncertainty is less obvious is the decision of a Ph.D. student to travel to university by bicycle. Suppose it is simply the best mode of transportation according to this Ph.D. student's preferences. No uncertainty seems to affect his choice. However, if he knew beforehand that he was to get a flat tire, which is always possible, he would prefer to go on foot. Again, the optimal decision depends on the outcome of an uncertain event. The apparent absence of uncertainty stems from the fact that the probability of a flat tire is very small compared to the threshold probability above which the student would prefer walking. A large amount of new information is therefore necessary to convince the student to leave his bicycle at home.

In water resources management, decision makers are often faced with decisions under uncertainty. An example is the daily release of water at a hydropower dam, given the incomplete information about future inflow into the reservoir. If the future inflow into the lake will be high, the optimal current release will also be high, because future spills are likely to occur if the water is not used now and the capacity of the generators is limited. If the future inflows are low, on the other hand, a better decision would be to keep the reservoir filled, so power is produced at a higher reservoir level elevation, yielding more power per unit of water volume. The decision maker does not know the future inflow and the best he can do is to use all information available at the time of the decision to choose the release that he expects to be best. Because the water manager is not able to influence the inflow, it makes little sense to judge these decisions in hindsight, based upon information on the actual inflow that turned out to occur.

Uncertainty, although in some occasions more than in others, plays a role in all decisions, adversely affecting their quality. There are two things we can do about that. In the first place we try to reduce uncertainty by obtaining information. This information can result in understanding and predictions that remove some of the initial uncertainty. In the second place, we have to deal with the remaining uncertainty by making rational decisions.

1.2.2 Rationality

Even in case we are excellent decision makers, the presence of uncertainty at the time of decision means that actions are not always optimal in hindsight, once more information has become available. Therefore, it is important to make a distinction between decisions that are right in hindsight and decisions that are right, given the information available at the time the decision was made. The latter are often referred to as rational decisions.

By definition, rational decisions maximize the expected fulfillment of the decision makers' objective from the perspective of the information available to him at the time. This thesis focuses on rational decisions in the context of water system operation. To limit the scope, the objective of the operator is assumed to have been correctly formalized and to reflect the objectives of society. Questioning the objectives is outside the scope of this thesis.

1.2.3 Risk

Risk is usually associated with adverse uncertain events, such as floods and nuclear accidents. However, uncertainty with respect to positive events can also be formulated in terms of risk by considering the failure to attain a potential benefit as an adverse event. Irrigation risk, for example, can be defined in terms of the probabilities of the water supply falling short to produce the maximum agricultural yield. In this risk, both the probability and the magnitude of the reduction in crop yield play a role. Generally, risk is defined as a function of probabilities and consequences.

A widely applied quantification of this definition is the engineering definition of risk:

$$\text{Risk} = \text{probability of an event} \times \text{consequences of that event.}$$

In this thesis, this definition of risk is used. Because probability is dimensionless, the risk is expressed in the same unit as the consequences. Often these are expressed in monetary terms, rendering them readily applicable in cost-benefit analyses. In the decision whether or not to build a dike, for example, construction and maintenance costs are weighed against the benefit of decreased flood-risk. This can be a very challenging task, given the fact that both the costs (e.g. ecological damage, destruction of the landscape) and the benefits (e.g. saving human lives, preventing emotional shock) are hard to express in monetary terms.

In addition to monetary values, a more general measure of costs and benefits within the framework of decision theory is the concept of utility. Utility cannot be defined directly, but is measured through preferences of a decision maker. If a decision maker prefers A over B, then A is supposed to have a higher utility than B for that decision maker. At first sight, there seems to be a circular definition in the concept of utility, since it *is used to define* the rational choice but is also *defined* in terms of choices (preferences). However, when applied to a set of decisions, the concept can be used to check whether a decision maker's preferences are rational (i.e. self-consistent) and to make him conscious of his preferences. Once the utility function of a decision maker has been captured in a quantitative objective, the best way to maximize utility are rational decisions taking into account all available information. When the information is incomplete, this entails taking into account uncertainty.

1.2.4 Risk-based water system operation is rational

One way to cope with uncertainty is to always be on the safe side. A decision can for instance be based on a worst case scenario, no matter how small its probability. An example from the context of structural water management measures is the construction of dikes. One could argue dikes should be designed to withstand the "maximum possible flood", because flooding leads to damage and loss of lives. However, flood protection also comes at a cost. Not only the monetary cost of dikes, but also the damage to environment and landscape has to be considered. These costs have to be balanced by the benefits of the increased flood protection. A worried citizen could of course argue that the loss of human lives can never be outweighed by such costs, but it is easy to point out that resources spent on dikes cannot be spent on, for example, healthcare. The potential loss of lives as a result of cutting funds there makes it clear that there is no evident "safe side" in this decision problem.

Risk-based decision making is rational decision making, given that a rational agent values an uncertain outcome as a linear function of its probability. If a decision problem is accurately formulated in terms of objectives (utilities), possible decisions and available information, the decision that minimizes risk is the best one. Other results from decision theory prescribe that preferences for decisions must be transitive and complete (Peterson, 2009). Transitivity means that preferences cannot be circular: if A is better than B and B is better than C, C cannot be better than A. Completeness means that different options

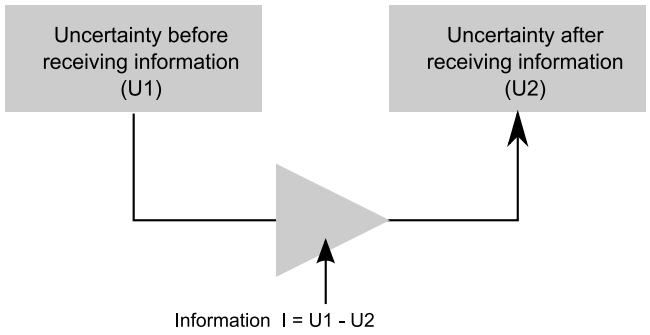


Figure 1.2: Information is the reduction in uncertainty

are always comparable: a decision must always be made, even if the decision is to postpone action until more information is available.

If for two rational decision makers the objectives and available information are equal, the decisions will also be equal and conditionally optimal in terms of expected utility. In practice, this situation never occurs, because of three reasons. Firstly, prior information is often implicit and never the same for two persons. Secondly, the objectives are often not entirely the same and partly implicit (hidden agendas). Thirdly, human decision makers are only partly rational. Specifically, rational decision making is sometimes avoided to prevent explicit moral choices. For example, in flood protection, the risks involve a combination of monetary value and human lives. A clear statement of the objectives together with a systematic treatment of information and uncertainties would prevent circular preferences and irrational choices. However, such an explicit formulation of objectives can only be done on the basis of moral choices. Because these are difficult to agree on, the objectives are often kept partly implicit. Some less harmful and very entertaining examples of predictably irrational behavior can be found in Ariely (2008).

1.3 Information reduces uncertainty

Uncertainty in a decision problem can be associated with incomplete information. The less information is available about the outcome of some event, the more uncertain we are about it, and this will be reflected in uncertainty about the best decision. Within the framework of information theory, which will be used extensively in this thesis and will be introduced in chapter 3, uncertainty is defined as missing information and information is defined as a reduction in uncertainty. This relation is schematically represented in figure 1.2. In the information-theoretical framework, uncertainty is defined as a unique measurable quantity defined in terms of probabilities. The exact definitions of these concepts will be introduced in chapter 3 and some extensions to these definitions will be presented in chapter 5.

For now, it must be stressed that uncertainty in this definition influences risk and value of decisions, but is not equivalent to it. Uncertainty depends only on how much is known or believed about some event, but not on what that event is. To put it differently: there is a difference between uncertainty and risk. Imagine a coin toss that is used for deciding

who will get coffee to fuel the research in some university office. Now imagine that same coin toss executed by some morbid dictator who uses it to make a decision on starting a nuclear war for fun. In both cases the uncertainty about the outcome of the toss is the same, but the risk associated with that outcome is far higher in the latter case. The reason that some people might perceive the second coin toss to be associated with more uncertainty is the higher uncertainty about the future state of the world as a whole. The uncertainty about the outcome of the coin toss itself, however, remains the same.

1.4 The value of information

Intuitively, most people would agree that more information leads to better decisions. A water system that is better known and monitored, is easier to control. Uncertainty adversely affects the quality of decisions and information reduces this negative influence by resolving uncertainty. For the person making decisions or the groups whose objectives he represents, information therefore has a value.

The value of information can be defined as the expected value of the outcome given a decision using that information minus the expected value of the outcome with a decision based on the prior information. An example is the increase in hydropower revenue due to improved inflow predictions. According to Hamlet et al. (2002), including information about El Niño Southern Oscillation into a more flexible operating strategy of the Columbia river hydropower system would lead to an expected increase in expected annual revenue of $153 million in comparison with the status quo. Generally, inflow predictions are important for hydropower production and better forecasts or better ways to make decisions based on these forecasts are an important topic in water resources research. Summarizing, it can be observed that information acquires value through its use in decisions.

1.5 Objectives and research questions

1.5.1 The broader perspective

The overall objective of water resources management is to make water systems behave in a way that is beneficial from a human perspective. Of course this is easier said than done. First of all, we need to define beneficial, and this leads to a number of questions that are progressively more difficult to answer: In what ways can we benefit from the water system? Who should benefit? Should we benefit now or later? Should we benefit at the cost of reducing benefits for future generations? Is there a reason to preserve ecosystems or the earth beyond the human benefits? Is survival of the human race the ultimate objective? Why are we here in the first place? Are we actually here?

The preceding series of questions makes it clear that it is in the end impossible to satisfactorily define beneficial. However, just as it is possible to live life without knowing the

ultimate purpose, it is possible to manage water resources without an all-encompassing definition of beneficial. We just do what we think is best, considering our incomplete information and simplified, derived objectives.

Information is beneficial to water management. Firstly, the degree of achievement of a derived objective can never be decreased by obtaining information about the workings of the system, provided the information is true and processed in a correct fashion. Secondly, more information about the water system helps to better define the objectives and possibilities to control the system. In other words, it helps posing the right optimization problem. Often, defining the problem and objectives correctly involves participation of stakeholders and negotiation (Soncini-Sessa et al., 2007). This process helps to derive the immediate objectives from more abstract underlying objectives. Ultimately, however, the underlying objectives reduce to moral statements and information cannot tell us what we should or should not strive for. In fact, it could very well be that all purpose is just an emergent behavior of our evolution.

This thesis deals with the role and benefits of information in the first sense. It is supposed that the objective has been defined and that it is clear how the water system can be influenced. The question then remains how to convert the information about the system and its surroundings into a decision that will optimize the objectives. Some tools that facilitate this process are developed within this research, often using simplified representations of the real decision problem. The main results are therefore general methodologies rather than solutions to specific real-world problems. Notwithstanding the somewhat theoretical character, the application of these methodologies can be part of finding such solutions.

1.5.2 Open questions

Water system operation continuously deals with decisions under uncertainty. This uncertainty is reduced by information from forecasts, which are based on models, which are based on observations. The remaining uncertainty must be taken into account in decisions. Because new decisions for the operation of water systems are taken continuously, there is a complex interaction between the current decision, future decisions and new information that becomes available in between subsequent decisions. The flow of information appears to be a key component in the whole process of going from observations to decisions.

In water system operation, this process is usually not viewed from the perspective of information being a quantity that flows through this process. Yet, such a perspective could provide new insights in these complex relations. Studying how to maximize the information that infiltrates our models and how it percolates into our decisions might be as important as studying the flow of water itself.

1.5.3 Research objectives

The objective of this thesis is to develop a framework to study the role of information in the context of water system operation. Because uncertainty plays an important role

in most water systems, it is important that the framework takes into account both the information that is available and the remaining uncertainty. Studying information in risk based water system operation requires describing the interrelations between objectives, decisions, uncertainty, measurements, models, forecasts, probability, value or utility, time and information.

Apart from offering another view on existing methodologies, a self-consistent framework can also serve to reflect on them and suggest improvements. Therefore the search for the framework is expected to yield several concrete and practical recommendations about existing methodologies and possibly some alternatives to these methodologies. Ideally, these methodologies should be defendable from a self-consistent framework.

1.5.4 Research questions

The main questions that need to be addressed in the framework that is to be developed in this research is:

- How does information play a role in optimal operation of water systems?
- How can this information be exploited optimally?

To answer these questions, both water system operation and the nature of information have to be investigated. A formal framework for quantifying information can be used to study its flow through the processing of observations into decisions. Because decisions are usually not just based on raw observations, but also on predictions made by models, both predictions and models are expected to play an important role in optimal risk based decisions. Ultimately, the information that enters the decisions, should improve their quality and therefore possesses some value. An important question is how this value can be maximized, i.e. how to exploit information optimally.

Special attention will be given to probabilistic forecasts, as they form summaries of available information, but also represent the remaining uncertainty. Providing the right information for risk based decisions might be seen as the task of producing a probabilistic forecast that is in some sense optimal. Having optimal forecasts requires that the information processing from observation to decision, which includes the employed models, is optimized. The evaluation of probabilistic forecasts is therefore an important part of the problem that will be studied in this thesis, because it allows some diagnosis of the information flow.

1.6 Thesis outline

This thesis focuses on several aspects of information in its course from observations to decisions. In chapter 2, the context of risk based water system operation is explored using a case-study. A framework for relevant time-horizons that describes how information about the future can affect current decisions is presented. It it also shown why uncertainty needs to be taken into account to arrive at optimal decisions. Shannon's formal theory

of information and uncertainty is introduced in chapter 3, along with some additional interpretations that will facilitate the understanding of the methods in the next chapters. This chapter also gives a short overview of some of the other applications of information theory within water resources management and hydrology. Chapter 4 presents the "minimum relative entropy update" (MRE-update), a method to add seasonal forecast information to an existing ensemble of historical time series by attaching weights to them. The information-theoretical foundation of the method ensures that not more information is added than is present in the forecast, but also not less. The information-perspective is also used to analyze what happens in the existing methods that this method aims to complement or replace.

Probabilistic forecasts, such as resulting from the MRE-update, are sometimes claimed to be problematic in terms of their evaluation and acceptance by decision makers. In chapter 5, a radically opposite view is presented, where it is claimed that probabilistic forecasts are the only forecasts that actually contain information in a strict sense. In this pivotal chapter, an information-theoretical analogy of a well-known decomposition of an existing score for the quality of probabilistic forecasts (the Brier score) is found. In light of this result, forecasting can be seen as a communication process in which information is transferred to the user to reduce his uncertainty. The presented framework for the evaluation of forecasts defines the relations between climatic uncertainty, correct information, wrong information and remaining uncertainty in this context and also explicitly distinguishes between useful information and pure information.

The evaluation of predictions, as treated in chapter 5, is of critical importance to the process of inference of models and therefore to both science and engineering. The quality of a model is determined by the quality of its predictions of unseen data. Chapter 6, building on the insights from the previous chapter, presents information as a central concept in the philosophy of science. Some thoughts on the link between the principle of parsimony, data compression and algorithmic information theory are presented. As an example application, common data compression algorithms are used to "ZIP" hydrological time series to estimate their information content. Some philosophical implications of the deep theories of algorithmic information theory, which sees science as data compression, are also discussed.

Chapter 7 returns to the narrower scope of information in the context of optimal water system operation. Stochastic dynamic programming is used to study the value of water when being allocated under uncertain conditions. It is shown why the complex dynamics of information play a role in both the value of water and in optimal decisions. A possible way forward is presented in the form of a more empirical approach to estimate optimal decisions under these complex information-dynamics. Chapter 8 summarizes the conclusions on both the conceptual and the applied level and proposes a few potentially interesting future research avenues.

Chapter 2

Risk-based water system operation

"The policy of being too cautious is the greatest risk of all."
- Jawaharlal Nehru

2.1 Introduction

This chapter[1] presents the formulation of an example risk based water system operation problem. Furthermore, a framework is introduced that describes how uncertainties about future events influence the current decision. The typically Dutch water system of the Delfland storage canals is used as an illustration of these concepts and the practice of water system operation in general. For optimization of water system operation, an online and an off-line approach is distinguished and illustrated. For the online approach, the Model Predictive Control (MPC) formulation for the Delfland system is analyzed. The background about the problem formulation and solution techniques is presented in sections 2.2-2.4, while section 2.5 concerns uncertainties affecting the water system.

Uncertainties influence the operation of the Delfland system in various ways. The uncertainties are present in the measurements, models, and predictions. The uncertainties in the predictions have an important time-dimension. On the one hand, uncertainties tend to grow with increasing lead time. On the other hand, the importance of future events for the present decisions decreases with lead time. Section 2.6 presents a framework for time horizons relating these effects and draws some conclusions for controller design. The time horizons are determined empirically for a simplified schematization of the Delfland system.

1. based on:
   - Weijs, S.V., van Leeuwen, P.E.R.M., van Overloop, P.J., van de Giesen, N.C. Effect of uncertainties on the real time operation of a lowland water system in The Netherlands. *IAHS publication 313 IUGG Perugia*, 2007
   - Weijs, S.V. Information content of weather predictions for flood-control in a Dutch lowland water system. *4th International Symposium on Flood Defense: Managing Flood Risk, Reliability and Vulnerability*, Toronto, Ontario, Canada, 2008
   - Overloop, P.J. van, Weijs, S.V., Dijkstra, S.J. Multiple Model Predictive Control on a drainage system, *Control Engineering Practice*, 16:531-540, 2008

Using the concept of certainty equivalence (section 2.7), it is found that uncertainties in the predictions need to be considered explicitly in the decision making process to make optimal risk-based decisions. A multiple model extension for MPC is proposed for this task (section 2.8). A further analysis of the Multiple Model Predictive Control methodology (section 2.9) reveals that future decisions influence the current decision. Therefore, also the information that will be available for future decisions is important for the current decision. Two MMPC formulations are presented that represent the two extreme assumptions of no new information and perfect new information. More realistic representations are dealt with in chapter 7, analyzing stochastic dynamic programming (SDP) formulations of this problem.

2.1.1 Water system operation as a mathematical optimization problem

When water systems are operated, usually some objective is involved, depending on the functions of the water system and the preferences of the operator and the organization he works for. A reasonable requirement for operation is that it strives to optimally meet these objectives, given the constraints posed by the water system and other requirements (e.g. legal norms). Examples of objectives are maximal profit, minimal danger, and keeping the water level within predefined limits. This last objective can also be formulated as a constraint, which can be combined with other objectives, like minimizing pumping costs. Generally, objectives and constraints are interchangeable mathematically, although in our logical perception, it would be awkward to formulate adherence to physical law as an objective instead of a constraint. Optimal water system operation can therefore be regarded as the optimization problem of finding the choice for all actions within our control that optimizes the objective while satisfying the constraints. Such a problem is here referred to as the *control problem*.

2.1.2 Formulation of a control problem

Solving a control problem requires defining a system, controls, objectives, constraints, measurements and disturbances. The system to be controlled is usually formulated in discrete time (Eq. 2.1).

$$\mathbf{x}_{t+1} = f\left(\mathbf{x}_t, \mathbf{u}_t, \mathbf{d}_t\right) \tag{2.1}$$

where vector $\mathbf{x}_t$ is the state of the system, $\mathbf{u}_t$ the vector of control actions and $\mathbf{d}_t$ the vector of disturbances at time t. The function f describes the behavior of the system under influence of the controls $\mathbf{u}_t$ and the disturbances $\mathbf{d}_t$. This can be a nonlinear function, but in control theory the system is often assumed to be linear for computational reasons, leading to the state-space formulation

$$\mathbf{x}_{t+1} = A\mathbf{x}_t + B_u\mathbf{u}_t + B_d\mathbf{d}_t \tag{2.2}$$

in which matrix A describes the autonomous behavior of the system, B_u the influence of the controlled inputs, and B_d the influence of the known disturbances on the system.

The controls $\mathbf{u}_t$ on a system are often limited by physical or other constraints. For example, if $\mathbf{u}_t$ represents pump flows, the capacity may be limited by a vector of maximum pump capacities $\mathbf{u}_{t,\max}$. There may also be constraints on the states. If state i represents the total volume of water stored in a reservoir, for example, that state cannot become negative, leading to the constraint $[\mathbf{x}_t]_i \geq 0$, where $[\mathbf{x}_t]_i$ is element i of the state vector.

Apart from satisfying the constraints, water systems are usually controlled with additional objectives, which can be expressed in quantities that need to be maximized or minimized. Identifying these quantities is not always straightforward, as they reflect the definition of desirable behavior, which inevitably leads to questions about whose desires are most important. The objective function can be interpreted as reflecting the utility of a particular system behavior and can therefore be formulated and analyzed using economic theory. The objective can then be satisfied by maximizing

$$J = g(\mathbf{x}_t, \mathbf{u}_t, \mathbf{d}_t) \tag{2.3}$$

where g is the objective function and J the objective function value. This function usually does not only account for immediate benefits, but needs to take into account future behavior of the system as well[2]. Therefore, the objective function that actually should be minimized is

$$J = g(\mathbf{x}_{t..T}, \mathbf{u}_{t..T}, \mathbf{d}_{t..T}) \tag{2.4}$$

where T is the end of the time horizon of interest. For example, when a reservoir is used for hydropower production, the largest immediate benefits are obtained by using the turbines at full capacity. This, however, may lead to a rapid depletion of the reservoir, which is not in the interest of overall long term benefits. Keeping the reservoir levels higher results in higher power yields per unit of water that flows through the system.

In case the function g for the benefits can be disaggregated in time, the objective function in Eq. 2.4 can be written as a sum of the benefits in individual timesteps

$$J = \sum_{t=1}^{T} g(\mathbf{x}_t, \mathbf{u}_t, \mathbf{d}_t) \tag{2.5}$$

which allows the use of certain special solution techniques for finding the optimum of J (see subsection 2.1.3).

Measurements from the system can be a function of the state, decision and disturbance

$$\begin{aligned} \mathbf{y}_t &= h(\mathbf{x}_t, \mathbf{u}_t, \mathbf{d}_t) \\ &= C\mathbf{x}_t + D_u\mathbf{u}_t + D_d\mathbf{d}_t \text{ for a linear system} \end{aligned} \tag{2.6}$$

which can mean that not all states are directly observable. All information a controller can receive about the system is in the measurements. This can pose important challenges

2. Note that the most basic objective that emerges from our evolution, reproduction, also needs to optimize survival over a future time period in order to optimize reproduction probability. Optimizing over a finite time horizon is therefore very natural.

for the design of a control system. This chapter deals mostly with states that are directly observable.

The above state-space formulation of a control problem is just one of the possibilities. In other formulations, only the objective is considered and all other equations, including model, are seen as constraints. Ultimately, every optimization problem can in theory be reduced to an objective and a number of decision variables.

2.1.3 Solution techniques

When the control problem has been formulated, there are several techniques to find the optimal solution. Some of these techniques find analytical solutions and others approach the solution numerically. Several techniques make assumptions or require the problem to be cast into a predefined form that does not exactly match the original problem. In this section, the solution techniques that are used in this thesis will be briefly introduced.

Global optimization algorithms

When optimizing an objective function by manipulating two decisions that are real numbers (e.g. adjustable weir flows), the problem can be visualized as finding the highest point in a mountain landscape. One option to find such a point is to test each point in the landscape and compare it with the highest point so far. This strategy is referred to as exhaustive optimization or brute force optimization. Especially when there are many decision variables to optimize, this strategy becomes infeasible due to the high computational cost, e.g. exponentially increasing time of computation.

Global optimization algorithms reduce this computational burden significantly, by using efficient strategies to sample the search space, using results from previous samples to identify promising solutions. These algorithms are often biologically inspired, because in living organisms, populations and ecosystems, optimality often emerges from simple rules. Examples are *ant colony optimization* (Dorigo and Stützle, 2004), *particle swarm optimization* (Vesterstrøm and Thomsen, 2004) and evolutionary algorithms like *differential evolution* (Storn and Price, 1997) and its self-adapting variant (Brest et al., 2006). Some of these algorithms have been combined in meta-algorithms like *AMALGAM* (Vrugt and Robinson, 2007) that let several algorithms run in parallel and adapt their relative importance based on their performance. Apart from applications to model parameter estimation (see for example Duan et al. (1992); Shoemaker et al. (2007)) and design problems (e.g. Dandy et al. (1996); Savic and Walters (1997); Abebe and Solomatine (1998); Solomatine (1999)), evolutionary algorithms have also been applied in control problems in for example Rauch and Harremoës (1999); Merabtene et al. (2002); Huang and Hsieh (2010); Koutsoyiannis and Economou (2003).

The global optimization algorithms are sometimes also referred to as black box optimization algorithms, because they do not need any information about the problem they

optimize. They are especially valuable for problems that are difficult to solve (e.g. high-dimensional search space, many local optima), because the more efficient class of gradient based search algorithms fail to find the global optimum in those cases. Any structure that might exist in the search space is left for the algorithm to discover. This makes global optimization algorithms less competitive on problems with a clear structure, where thinking carefully during the problem formulation might reveal a problem that is far easier to solve. However, given the low cost of computer power compared to brain power (even Ph.D. students'), solving such problems using global optimization might still be the most economical solution.

Dynamic Programming (DP)

Dynamic Programming is a solution technique that makes use of the structure in a particular class of problems. One type of problem in this class are sequential decision processes, where decisions have to be taken repeatedly and each decision influences the state of the world for the next decision, and the objectives can be disaggregated in time (like in Eq. 2.5). While the original problem has many decision variables (e.g. a hydropower reservoir release for each timestep in the planning horizon) that have to be solved simultaneously, DP allows splitting the problem in a series of simpler problems, that can be solved one by one. This is achieved by disaggregating the problem in time, making use of the Bellman principle of optimality (Bellman, 1952), which states that "An optimal policy has the property that whatever the initial state and initial decisions are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decisions". This leads to the Bellman equation

$$H_t(\mathbf{x}_t) = \min_{\mathbf{u}_t} \left\{ g_t(\mathbf{x}_t, \mathbf{u}_t, \mathbf{d}_t) + H_{t+1}\left(\mathbf{x}_{t+1}\right) \right\} \tag{2.7}$$

in which $H_t(\mathbf{x}_t)$ is the optimal cost-to-go function in timestep t, evaluated at state $\mathbf{x}_t$, and g_t is the step cost associated with the transition from $\mathbf{x}_t$ to $\mathbf{x}_{t+1}$. The disturbance $\mathbf{d}_t$ is assumed to be deterministically known in this case. Equation 2.7 can be solved recursively, going backwards in time. At each stage, only the step cost g_t and the cost-to-go at the next timestep H_{t+1} (calculated in the previous iteration) need to be considered. For certain functional forms of $H_t(\mathbf{x}_t)$ and g_t, the equations can be solved analytically to yield an optimal policy $\mathbf{u}_t = f(\mathbf{x}_t)$, but for most problems the solution has to be found numerically and discretization of the state vector $\mathbf{x}_t$ is required. Discretization makes $H_t(\mathbf{x}_t)$ a look-up table, of which all values have to be calculated one by one. Also the policy f then becomes a lookup table consisting of the optimal control actions for each state. Finding the table is referred to as *off-line* optimization (see section 2.3), although it is also possible to re-optimize the policy online. When the number of states in $\mathbf{x}_t$ increases, the computational cost grows exponentially. This is referred to as "the curse of dimensionality". Yakowitz (1982) provides a review of applications of dynamic programming to water resources planning, which date back to Hall and Buras (1961), and the application in the form of Stochastic Dynamic Programming (SDP), which is able to deal with uncertainties and has now largely replaced deterministic DP. In chapter 7, SDP will be applied to a reservoir operation problem and analyzed in terms of information flows and the value of water.

Model Predictive Control

Model Predictive Control (MPC) is an approximate solution of the Bellman equations for a finite time horizon (Bertsekas, 2005), usually for the case where the model is linear and the objective functions are quadratic. It has the advantage that no discretization of the states is necessary and it does not suffer from the curse of dimensionality. Instead of finding the optimal action for each possible state, the problem is solved "online" over a receding horizon, calculating a sequence of control actions $\mathbf{u}_{t...t+h}$ that is optimal for the initial state $\mathbf{x}_t$ and predicted sequence of disturbances $\mathbf{d}_{t...t+h}$. For each timestep the optimization is repeated and the first control action $\mathbf{u}_t$ is executed. The objective function that is optimized in each timestep is

$$\min_{\mathbf{u_{t..t+h}}} J = \sum_{i=0}^{h} \mathbf{x}^T(t+i|t) Q \mathbf{x}(t+i|t) + \mathbf{u}^T(t+i|t) R \mathbf{u}(t+i|t) \tag{2.8}$$

where Q and R are the weight matrices for the penalty on states and control actions, respectively, and i is a counter for the timesteps in the optimization horizon of length h. The solution of this quadratic objective function by MPC can explicitly take into account the linear model (Eq. 2.2) and possible constraints on the states $\mathbf{x}_{t...T}$ and control actions $\mathbf{u}_{t...T}$.

Model predictive control has been applied in various studies of optimal water system operation, for systems ranging from irrigation channels to the centralized water management of the main water systems in the Netherlands (van Overloop et al., 2010a,b) or emergency flood areas (Barjas Blanco et al., 2010). The Ph.D. thesis by van Overloop (2006) presents a number of applications and further references. Recent advances include the application of MPC to the control of both water quality and quantity simultaneously (see for example Xu et al. (2010)). The next section describes the practical details of a real world water system control problem, where MPC has been implemented in a decision support system (DSS) that is used in daily practice.

2.2 An example: a lowland drainage system

The western part of The Netherlands is mainly situated below sea level. Most of the water systems there are divided into hydrological units called "polders", in which a certain water level is maintained, depending on the land use and functions assigned to the area (Schuurmans et al., 2002; Lobbrecht et al., 1999; van Andel et al., 2010). For the drainage of these areas, the responsible water boards are mainly depending on pumping stations. The evacuation of the drainage water usually takes place in two stages, as depicted in Fig. 2.1. The water is first pumped to the storage belt ("boezem"), a system of canals and lakes connecting the pumping stations of different polders to large pumping stations at the coast or large rivers. Here the water is further elevated in a second step. The boezem serves both for transport and for storage of drainage water. The water level is usually higher than the surface level in most of the surrounding polders, but lower than the water

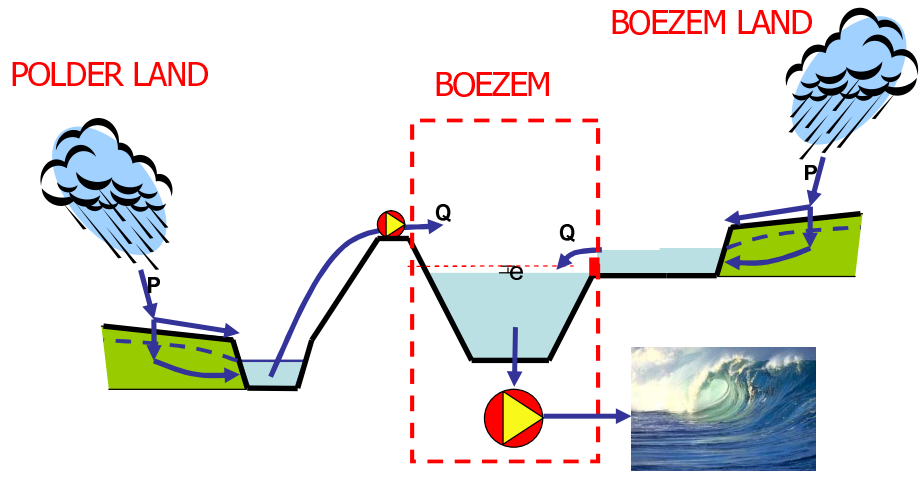


Figure 2.1: A schematic representation of the "polder-boezem" systems that characterize the western part of The Netherlands. The "polder land" drains onto the main "boezem" canals through pumping stations, limiting the peak flows. The higher lying "boezem land" can cause higher peak flows into the system. The water is eventually pumped into the sea (see bottom).

level in the surrounding outer waters. These typically Dutch water systems are referred to as "polder-boezem" systems. In this chapter, the water system of the Delfland area is used as an example.

2.2.1 Delfland

Delfland is the water board responsible for the south-western part of the province of South Holland. With 1.4 million inhabitants on 410 km^2, the Delfland region is one of the most densely populated areas in The Netherlands. The area is characterized by a high concentration of economical value. The water system has important functions for drainage, water supply for agriculture, navigation and recreation. The eastern part consists mainly of polder areas while the western part also has more elevated areas, draining into the boezem canals directly. Greenhouses cover a large area, resulting in a very fast runoff process. The total area of the boezem canal system is 7.3 km^2 (about 1.8 percent of the total area). The water level in these canals is usually kept close to the target level of -0.42 m below mean sea level. The tolerable range for extreme conditions is between -0.60 m and -0.30 m below sea level. If large amounts of precipitation are expected, the water level is temporarily lowered. The total capacity of the polder pumping stations discharging on the canals is around 50 m^3/s, but due to the fast runoff from the higher lying areas, the total inflow can easily exceed 100 m^3/s during heavy rainstorms. The main pumping stations discharge the water from the boezem canals to the North Sea and the artificially canalized tidal river "Nieuwe Waterweg", connecting the port of Rotterdam to the sea. The total capacity of the main pumping stations depends on the tide of the outside water and can vary between 50 and 70 m^3/s. However, in practice this capacity cannot be reached, because of limitations on the transport capacity of the canal system. High pump flows

should be avoided where possible to avoid problems with high flow velocities and steep water level gradients.

Importance of anticipation on extreme events

As can be seen from the quantities mentioned above, inflows to the boezem can exceed maximum outflow capacity by 50 m^3/s. This can last several hours and in extreme cases lead to space averaged water level rises of 20 cm. Because of the limited channel capacity, local water levels can even rise up to 30 cm in a few hours. To accommodate this water level rise, both the lower and the upper margin need to be used. This means that pumps must lower the water level to -0.60 m before the start of the event, so the maximum level of -0.30 m will not be exceeded. To be able to anticipate on these events, it is necessary to use inflow predictions up to several hours ahead. Altogether, management of the Delfland water system is a challenging task. The increase in greenhouse area and the apparent regional climate trend towards more frequent extreme events have led to considerable damage over the last 15 years. The water board is currently executing a large program of structural measures to increase storage and discharge capacity in the system. Apart from the structural measures, a new decision support system is being tested to improve operational management

2.2.2 The Delfland decision support system

Traditionally, all pumping stations were managed by operators of the water board, based on visual observation of water level gauges. Over the years, this process has become increasingly automated. Nowadays, most of the polder-pumping stations switch on and off automatically, based on local water level measurements in the polders. The main pumping stations of the boezem are operated centrally, based on water level measurements, information about the situation in the polders, and expected meteorological conditions. Until recently, this was done purely relying on judgment of the operators. Now, a new Decision Support System (DSS), built by the engineering consultant "Nelen & Schuurmans", aids the operators in controlling the water levels in the boezem by advising flows for the main pumps. When results are satisfactory, it is also possible to switch the system to fully automated mode, in which the main pumping stations are directly controlled by the DSS. While testing the system, the operators give feedback on the decisions the DSS proposes, which helps to elicit extra operational constraints and objectives that the operators take into account through their experience.

2.2.3 Objective of control

Damage occurs if the water system fails during extreme events. Making optimal use of the possibilities to control the boezem system will reduce both failure frequency and impact. Avoiding system failure is not the only objective. In fact, failure in a boezem system should not be viewed as a single Boolean event, but rather as a continuous damage function of

the water level deviations from a target level. Low water levels are associated with long-term effects such as acceleration of land subsidence, decay of foundations and instability of embankments. High water levels can cause flooding, risk of dike breach and the necessity to impose a pump restriction for the polders, causing flooding problems there. The challenge for the operators is to secure the evacuation of water, discharged from the surrounding land, while balancing the short- and long-term costs associated with high and low water levels.

Usually a target water level has been set by the water board with democratically elected representatives of the various interest sectors. These include owners of agricultural land, house owners and inhabitants. Because zero deviation of the target level throughout the system is never attainable, the objective should also quantify the "cost" associated with the deviations, as a function of their direction, magnitude, location and duration. This makes it possible to balance the deviations in a way that minimizes costs or damage. Apart from the water-level related variables, the variables associated with control actions, like the pump flow, could also be part of the objective function. In this way, the operation costs of the pumps can be incorporated into the objective. In most cases, these are relatively low compared to the water level related costs, making minimization of operating costs a secondary objective that becomes relevant for the decisions only in non-critical situations.

High pump flows also cause high water level gradients in the canals, which leads to relatively high water levels in the center of the system (see Fig. 2.2). Therefore, a penalty on high pump flows indirectly also reduces water levels in the center of the system, because it encourages a pumping policy that is spread out more over time. The objective function that was established in collaboration with the operators is quadratic for the deviation from target level and the control flow and linear (by summation over the time steps) for the duration. The spatial variability has not been included, but this is possible by introducing extra state variables for water levels in the different areas. This objective function can either be used as a performance indicator to evaluate control rules, or directly in a control policy, based on real time optimization. In the latter case, optimization should take place over a certain time horizon, to balance current and future costs and to allow anticipation of future events if necessary. This type of optimization is known as Model Predictive Control (MPC), which was introduced in subsection 2.1.3 and will be further explained in section 2.4. The objective is minimizing the cost function J over this time horizon:

$$\min_{u} J = \sum_{i=0}^{n} \left\{ \mathbf{x}^T(k+i|k) Q \mathbf{x}(k+i|k) + \mathbf{u}^T(k+i|k) R \mathbf{u}(k+i|k) \right\} \tag{2.9}$$

in which $\mathbf{x} = e = h - h_{opt}$ is the state vector (deviation of the water level h from target level h_{opt}), k is the current time step, n is the number of time steps within the prediction horizon, i is the counter for these time steps, $\mathbf{u}$ is the control action vector (pump flow), Q and R are the weight matrices for the penalty on states and control actions, respectively. In the example case, these vectors and matrices are scalars, because the system is modeled as one reservoir controlled by one pumping station. Before going into more detail about

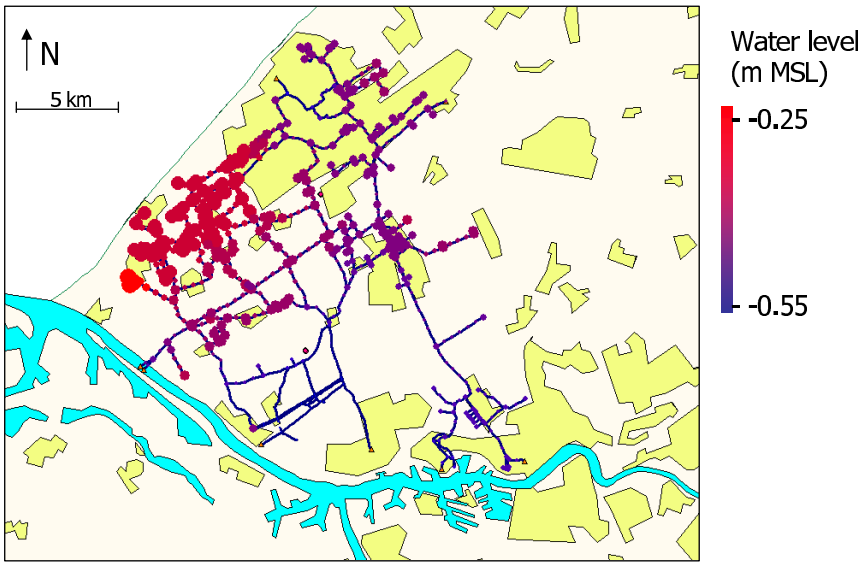


Figure 2.2: The water levels in the Delfland system can have some variability during extreme events. The map shows a snapshot of instantaneous water levels resulting from a simulation of an extreme event in 1998 using the hydrodynamical model Sobek. The large circles correspond to high water levels.

the formulation of the MPC formulation for the Delfland system, the alternative strategy of solving a control problem off-line is analyzed first.

2.3 Off-line optimization of operation

Operation of water systems requires reacting on the current situation and anticipating future events that can influence the water system. In the off-line approach, this reaction and anticipation is formulated as a precalculated rule that specifies how to act upon each specific situation, as summarized by the input variables of the rule. The rules thus map each situation to a control action, where a situation can be described by the state $\mathbf{x}$ and additional information $\mathbf{I}$.

$$\mathbf{u} = f(\mathbf{x}, \mathbf{I}) \tag{2.10}$$

The simplest off-line rules take into account only one local variable. For example, a pump may switch on and off based on the water level of the connected canal. This is called feedback control, because the control action depends on the variable that is controlled. The levels at which the pump switches can then be optimized off-line. This can for example be done by running simulations of a model of the water system including the rule for the pump and changing the rule until it best satisfies the objectives. Finding the optimal rule can be done by trial and error by an optimization algorithm, but can also be done analytically

in some cases. Examples of well-established feedback controllers are proportional (P) or proportional integral (PI) feedback controllers and linear quadratic regulators (LQR) (Kwakernaak and Sivan, 1972).

An improvement to optimal feedback rules can be made when the rules depend on more information. For example, the rule may be varied depending on the season. This is typically the case in many polder systems, where the target water levels are higher in summer, when the demand for water is higher and there is an infiltrating flow from the canals into the adjacent agricultural lands. The different objectives in different seasons can also be interpreted as emerging from the underlying, implicit objective to control the water availability in the root zone of the agricultural fields.

In feedforward control, information about the actual disturbance is used as an input to the rule. The advantage of including this information is that actions can be taken before the disturbance starts to affect the states that have to be controlled. For example, the pumps in the Delfland System can be started at the moment inflow into the system is observed, proportional to its magnitude. This would lead to smaller deviation than postponing action until the inflow causes a measurable water level rise. To effectively use feedforward control, a model of how the disturbance and the control action influence the controlled state is necessary. If the measured disturbance is rainfall, a rainfall runoff model is needed, in combination with a simple balance model that relates inflow, controlled outflow and water level rises. Feedforward control can also be based on experience, in which case the models are implicit.

Control rules can also include anticipation on forecast rainfall. For example, the water level in the canal system may be lowered to an emergency level that is predefined by the water board, when exceedence of a certain rainfall threshold is expected. The value of the threshold can be optimized off-line. When uncertainty in the forecasts is taken into account, the rules can also include a probability threshold, e.g. 'the water level will be lowered with 10 cm if there is a probability of more than 25% that a rainfall of 20 mm in the next 12 hours is exceeded'. Thresholds for the amount, time horizon, and probability can be optimized off-line to meet certain requirements on the water system (van Andel et al., 2008; van Andel, 2009). This is often implicitly the case in Dutch polder and storage canal systems, although the target levels for water systems and rules for anticipation are determined in a negotiation process rather than in a mathematical optimization with predefined objectives. After an optimal rule has been found by off-line optimization, the only computation that needs to be executed in the real time operation of the water system is the implementation of that predefined rule.

2.3.1 Example: the "regelton"

Suppose a water butt will be equipped with a proportional feedback controller (regelton < Dutch: regenton = water butt, regel = control). The water butt serves both for reduction of peak flows and for watering plants. Therefore, there will be a penalty for both spills and shortages. The outflow of the reservoir can be regulated proportionally to the water

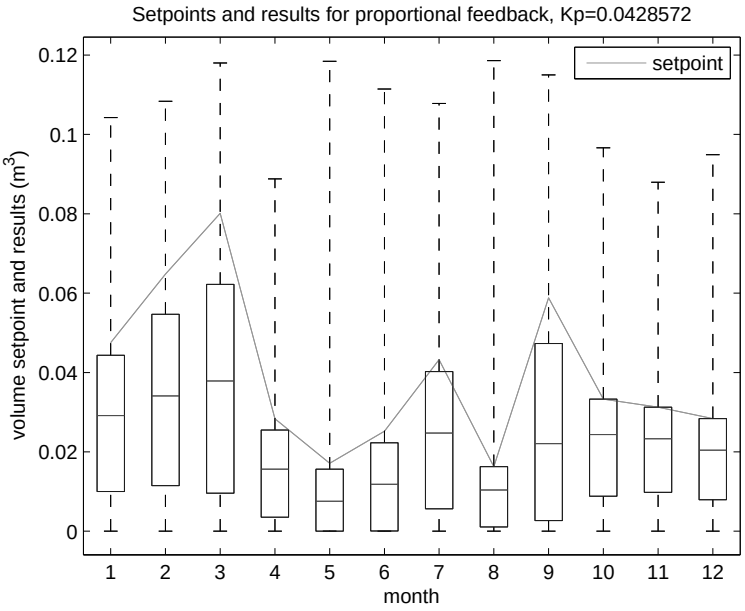


Figure 2.3: The optimal month setpoints and feedback gain for the water butt. The box and whisker plots give the median, quartiles and extremes for the volumes in the butt for each month, based on a simulation of the rule for a 100 year rainfall time series.

level deviation from a certain target level, but is limited to a certain maximum outflow. The proportional feedback controller can have a different target level for each month and has one k_p parameter, that determines how much the outlet is opened when a positive deviation from the target level occurs. The optimization problem that has to be solved is

$$\min_{k_p, h^*_{1..12}} J = k_1 \text{shortage}^2 + k_2 \text{spill}^2 \tag{2.11}$$

subject to the mass balance constraint, the inflow and the maximum outflow. To optimize the 13 parameters for the controller, a global optimization algorithm was used that does not need to make any assumptions on the structure of the problem. During the optimization, many simulations of the controlled water butt are performed, with a 100 year time series of daily measured rainfall as an input. In each simulation, the controller has a different 13 parameter vector. The objective function value over the 100 years is evaluated and used to generate new proposed parameter sets. The optimization was performed using a genetic algorithm developed by Storn and Price (1997), called Differential Evolution (DE). This algorithm efficiently searches the 13-dimensional parameter space by maintaining a good balance between exploration of the whole space and refining results in promising regions. The resulting monthly setpoints and some statistics of the water levels achieved within the 100 year period are shown in figure 2.3.

The problem of finding the optimal setpoints for the water butt is of a similar nature as finding an optimal rule curve for a hydropower reservoir (see chapter 7). The approach that was taken here is empirical and finds the best setpoints and proportional feedback gain for a given time series of disturbances. In this case the amount of information in the data was sufficient to determine parameters that will also perform reasonably well for

future situations. When the control rules that need to be found become more complex or less data are available, it may become more difficult to find good control rules empirically.

2.3.2 Disadvantages compared to online optimization

A problem with finding optimal control actions empirically is that every possible situation must be considered beforehand. Especially if the best action depends on a combination of many variables, or on the distribution in space and time, many different combinations and patterns must be considered to find the best rule. If only a limited amount of data is available, these patterns have to be generated using stochastic models that accurately capture the spatial and temporal dependencies of the inputs. It is important to note, however, that such models usually do not increase the amount of information on which the decision rules are based. More complex control actions, e.g. the use of multiple pumps, need more data to reliably establish the control rules. Projection of the situation to lower dimensions may make it easier to identify the optimal rule, but often results in a loss of information.

If the system to be controlled is well defined and the way the optimal control action depends on the inputs is complex, a better approach is to execute the optimization online, especially if forecasts of reasonable quality are available. This avoids the need to represent the patterns that are the input for the rule in low-dimensional form. Instead, any information that influences the decision through the model is automatically taken into account. Instead of a rule, the result directly consists of the optimal action, suited to that particular present situation. Consequently, the online optimization has to be repeated every timestep.

2.4 Online optimization of operation by model predictive control

MPC solves a constrained optimization problem over a receding horizon. This means that the optimization finds a sequence of actions that optimizes the value of the objective function over the whole prediction horizon, while satisfying the constraints on water levels and control actions. In this optimization, MPC makes use of an internal model of the controlled system, to calculate the future states as a result of the projected actions (see Fig. 2.4). In the case of the Delfland system the model represents the system with one single reservoir, which is controlled by one pumping station representing the total pump flow out of the system. The optimization horizon is 24 hours ahead and the internal model has a calculation timestep of 15 minutes. The pump flow can vary between the constraints of -6 m^3/s to 70 m^3/s. The negative pump flow corresponds to letting fresh high-quality water into the system from a nearby lake. The maximum pump flow constraint varies in time with the outside water level. For predicting the outside water level, the system uses the astronomical tide. Other, second order effects are neglected. The Delfland decision support system uses a post-processing scheme to convert the total projected pump flow

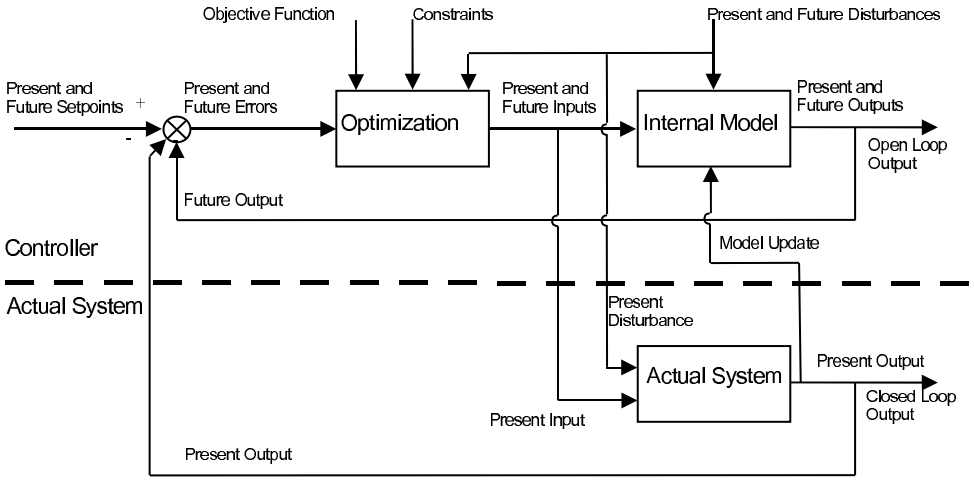


Figure 2.4: Schematic representation of model predictive control on a real system (modified from van Overloop, 2006).

into instructions for the individual pumping stations. For the calculations in the rest of this chapter, a simplified representation of the MPC controller is used, with a fixed maximum pump flow.

2.5 Uncertainties affecting the Delfland system

In this section, a number of sources of uncertainty relevant for the Delfland control problem are reviewed. The estimation of the inflow to the boezem canals is treated in more detail, including a description of the rainfall runoff model that is used in subsequent analyses.

2.5.1 Uncertainties

The optimal pump flow is computed every 15 minutes on the basis of the latest information available. Apart from the actual and forecast inflow, this information includes water levels for 8 different locations in the boezem canal system. These point measurements are used in a weighted average to calculate the representative water level (RWL). This RWL is used in all mass-balance equations and serves as an initial condition for the state of the single reservoir model. A difference of 1 cm between RWL and target level, takes 20 minutes at full pumping capacity to compensate. Especially during high pump flows, the limited conveyance capacity of the canals can cause water level differences up to 10 cm within the system. The limited spatial resolution of the measurements in combination with the spatial variability of water levels is the main source of uncertainty for the actual state of the system. Apart from the actual water level and volume in the boezem system, the actual inflow is important information for the controller maintaining the water level. Part of this inflow is the outflow of polder pumping stations, the rest is originating from higher lying areas and flows into the boezem canals through weirs and other uncontrolled flow processes. Only the polder pump flows are partly known, but the telemetry system is

not yet advanced enough to have this information available centrally in near real time. Therefore, the inflow in the boezem canals can not be measured directly at this moment.

One possibility to estimate the inflow is using a rainfall runoff model of the surrounding land, including the polders. This means that the local controllers switching the polder pumping stations have to be modelled as one of the rainfall runoff processes. This results in the difficulty of modeling the joint emergent behavior of many small systems with hysteresis, which is sometimes also encountered in natural systems; see e.g. the application of "hysterons" by O'Kane and Flynn (2007).

Another possibility for inflow prediction is using the mass-balance of the boezem canal system. Because of the persistence, the past inflow can give quite an accurate estimation of the current inflow. By using the known outflow through the main pumping stations and the changes in volume calculated from the water level measurements, the past inflow can be estimated. Because of the large high frequency errors (see noise in lower graph of Fig. 2.5) in the representation of the volume by the level measurements, this balance calculation needs a considerable balance period to be accurate. A longer period acts as a moving average filter, removing errors due to temporary local drawdown effects and waves. The downside of using a long balance period is that the existent temporal variability of the inflow is also filtered out and the estimation lags behind approximately half of the balance period.

The DSS uses a combination of the two approaches. A fast, lumped rainfall runoff model makes a first estimation of the inflow for the past 12 hours, using rainfall measurements. When this inflow is different from the inflow derived from the water balance over the boezem canals, a correction term is added to the inflow predicted by the rainfall runoff model. In this way, persistent errors in times of slowly varying inflow are compensated by the flow calculated from the balance, while the fast response to precipitation events is secured by the rainfall runoff model. The rainfall runoff model is fed by actual precipitation measurements at 8 locations. Weighted averages are used for different areas. For the inflow prediction, the model is fed by the updated weather forecast, but because of integrator elements in the model, the short term inflow prediction is also very much determined by past rainfall. Uncertainties in the inflow forecast are caused by model uncertainty, predicted rainfall uncertainty and measurement errors in the rainfall.

2.5.2 Data driven inflow model

In the DSS, a conceptual rainfall-runoff model is used, that was developed in Weijs (2004). For the analyses in this chapter, an even more parsimonious model was formulated, making use of the data collected by the DSS since its installation. A rainfall runoff model was identified from the data using linear system identification (Ljung, 1987). First, the measured water levels and pump flows for a two year period were used to estimate the inflows to the canal system, using the mass balance over the canal system. Second, linear system identification was applied to link the inflow signal to the measured rainfall signal.

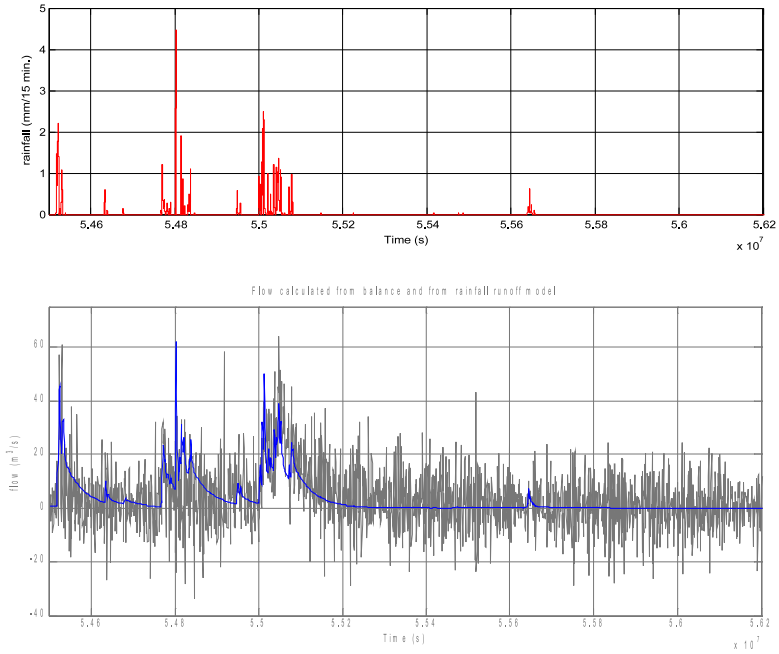


Figure 2.5: The response of the identified rainfall-runoff model to a rainfall event, compared to the inflow signal that was derived from the water balance over the canal system. The noisy signal for the balance-derived inflow results from small fluctuations of the area-averaged water level, which is not necessarily representative for the volume in the system.

Several model structures were tested and their parameters estimated. The model with the best performance was used. The transfer function after the Laplace transform reads

$$G(s) = K\frac{1 + T_Z s}{s\,(1 + T_{P1} s)} \tag{2.12}$$

where s is the laplace transform of the input signal and the constants are (gain, zero, pole) $K = 2.69 * 10^{-7}$, $T_Z = 7.66 * 10^7$ and $T_{P1} = 4968.3$. As can be seen from figure 2.5, this model behaves almost like a linear reservoir, where outflow is linearly dependent on storage, with some additional delay. The model captures the essential dynamics of the transformation of rainfall to inflow into the canal system sufficiently for the purpose of evaluating the control performance under uncertainties in the prediction of precipitation.

2.6 Relevant time horizons for uncertainty and optimization

Because uncertainties in predictions usually increase with the lead time, one would expect that the most problematic uncertainties exist close to the end of the time horizon. However, depending on the problem and formulation, the influence of the prediction on the current decision usually also decreases with lead time. This also means that the optimality of control will be less sensitive to this information and thus to the uncertainties therein. To get insight into the uncertainties in the forecast, use of probabilistic forecasts (e.g.

ensemble forecasts) is proposed. In the light of Bayes' rule, such a forecast can be seen as the prior climatic distribution, conditioned on actual information, to become a posterior distribution (Krzysztofowicz, 1999, 2001; Murphy and Winkler, 1987). The forecast is thus based on the multi-year average and spread of rainfall in a certain season, combined with information about current weather patterns. With increasing lead time, this conditional, posterior distribution approaches the climatic distribution, in which no information about the actual state is contained (see Fig. 2.6 for an illustration).

2.6.1 Time horizons relevant for prediction and control

If we use a probabilistic forecast in real time control, we can define two horizons relevant to the problem:

(a) The information-prediction horizon (TIp)
(b) The information-control horizon (TIc)

Horizon (a) can be defined as the time span from the actual moment until the moment where the conditional distribution of future events, conditional to all actual information, becomes the same as the marginal (climatic) distribution of these events. The length of this horizon depends on the forecast system and the statistical properties of the input forecast. This may also depend on the season. Horizon (b) is defined as the time span from the actual moment until the moment from which information does not influence the control actions anymore. This can be the case because the control action is at one of its constraints, or when optimality requires postponing actions.

Note that in control theory there is also a definition of the control (Tc) and prediction (Tp) horizons. These horizons define, respectively, the number of control moves to be calculated and number of prediction time steps over which the future state is calculated. These are parameters of the controller configuration. Rules to choose these parameters are usually based on characteristic time-constants of the system (Camacho and Bordons, 1999). In contrast to these definitions, here the horizons TIp and TIc are defined, which depend on predictability of inputs and sensitivity of the control action, respectively. Regarding the relation between these four horizons, the following observations can be made:

1. If the controlled system has delays in its behavior, Tp should be much larger than the delay-time.
2. Tp should also be much larger than Tc + delay time. This is necessary to evaluate the full effects of each calculated control action.
3. If TIc > TIp, extending the TIp by more advanced predictions based on more information helps to improve control. Tp and Tc should be chosen larger than TIp. At the end of Tp, the cost-to-go function can be used that is based on a steady state optimization (Faber and Stedinger, 2001; Kelman et al., 1990; Loucks and van Beek, 2005; Negenborn et al., 2005).
4. If TIp > TIc, we can know something about the future after TIc, but it has no influence on the decision now. The information about cost-to-go functions after TIc is not relevant, because the possible control sequences that lead to minimum total cost do not diverge yet. Tp does not need to be longer than TIc.

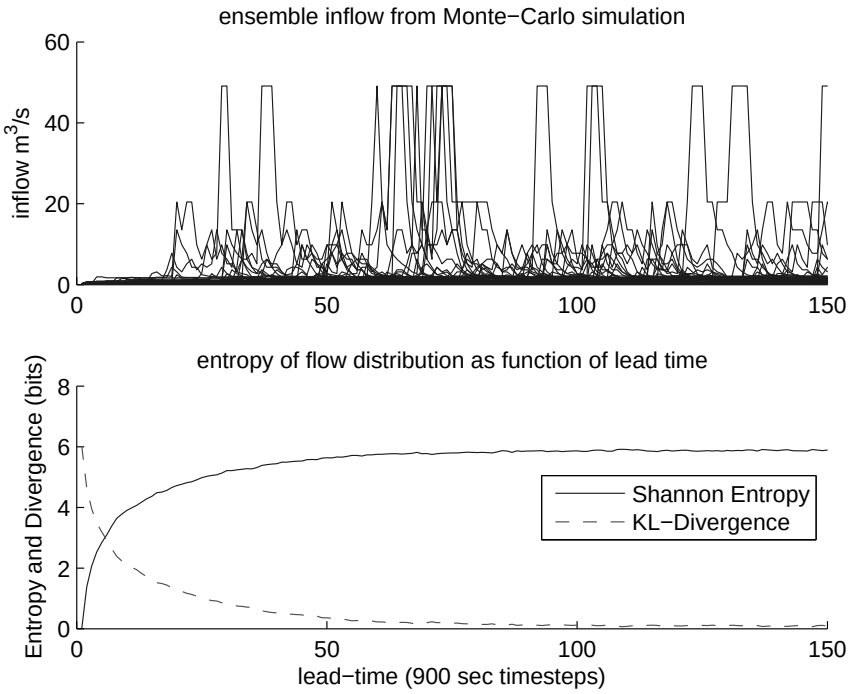


Figure 2.6: For increasing lead time, the uncertainty in the inflow increases and approaches the climatic uncertainty. The picture shows a Monte-Carlo simulation ensemble of a second order Markov Chain, with transition probabilities empirically derived from the modelled inflow series by Eq. 2.12. The Shannon-Entropy is a measure for uncertainty and the KL-Divergence a measure for information relative to the climate. These measures will be introduced in the next chapters.

5. In general, it can be stated that extending Tc and Tp beyond TIc is not necessary, but has no negative influence on the control, except for the computational cost.
6. Extending Tp and Tc beyond TIp is not necessary either, and can have a negative influence, if the forecast system is biased. In that case, the climatic distribution would provide a better estimate than the forecast

2.6.2 The time horizons for the Delfland system

For the model predictive controller used in the DSS for Delfland, the sensitivity of the control action for uncertainties was tested by providing different forecast errors and measuring performance by evaluating the objective function over a closed loop simulation period. The sensitivity can be determined in two steps:

1. Test whether the control action is sensitive to information after a certain time step.
2. If so, test what the resulting reduction in optimality is, by evaluating the closed loop value of the objective function.

Because positive and negative deviations from the target level are punished equally strong in the objective function, anticipatory pumping is only used if pump constraints are likely

to be violated and positive deviations are expected somewhere in the future. In this case, the pumping necessary to counter the positive deviations is postponed as much as possible to avoid long periods with low water levels. Under normal conditions, in which the pump constraints are not relevant, the actual water level and the actual flow, therefore, are the only variables influencing control actions. In this case, the MPC controller behaves similarly to a feedforward controller. In cases where constraints become relevant (when the expected inflow exceeds the maximum outflow), anticipatory pumping becomes necessary. In these cases, there will be a period of time in the future in which the pumps are planned to operate at their full capacity. If the inflow peak is close enough to the actual moment, the pump flow is already at full capacity and will not be sensitive to an increase of the forecast flow.

If the event is farther away or smaller, anticipation will be postponed, so in that case the action will be insensitive to changes in the future. In between these two, there is a relatively small range of situations in which the optimal action consists of anticipation of future constraints, but not at full pump capacity. In that case, the control action is sensitive to any change in forecast, as long as it occurs within the period of projected full pump capacity use. This period can be bounded by physical or operational constraints on the water level, but in the theoretical case that the expected inflow is very close to the maximum pump flow, can be unbounded. In this case, TIc goes to infinity. Summarizing, the controller can be in three situations:

(a) no anticipation necessary
(b) sensitive to prediction
(c) pump at full capacity

Situation (b) is always a transition from situation (a) to (c). In a deterministic optimization with a perfect prediction, the time period of this transition is limited and very much determined by the objective function and the dimensions of the system. In practice, however, predictions are uncertain to a considerable extent. Apart from the small errors in prediction, that may or may not influence the actions depending on the situation, larger uncertainties in the prediction may also exist. These errors may also influence the situation (no anticipation, sensitive, full capacity) of the controller. It is only possible if all future scenarios lead either to situation (a), or all lead to situation (c), that the controller is insensitive to the prediction. If this is not the case, the consequences and probabilities of all inflow scenarios have some impact on the decision.

Forecast accuracy: the information prediction horizon

For the Delfland system, 2.5 years of hourly rainfall forecasts for up to 24 hours ahead are available. The forecasts concern the hourly average rainfall for the Delfland area. When comparing the hourly forecasts from the Delfland dataset with the measured rainfall, different choices can be made about the period over which the sums are taken and the lead time for which to compare the forecasts. Also, there are two different methods to compare: ‘horizontal” and “vertical” methods, which are illustrated in figure 2.7. The

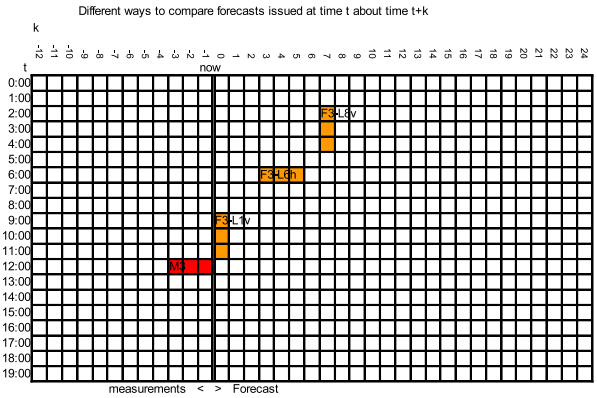
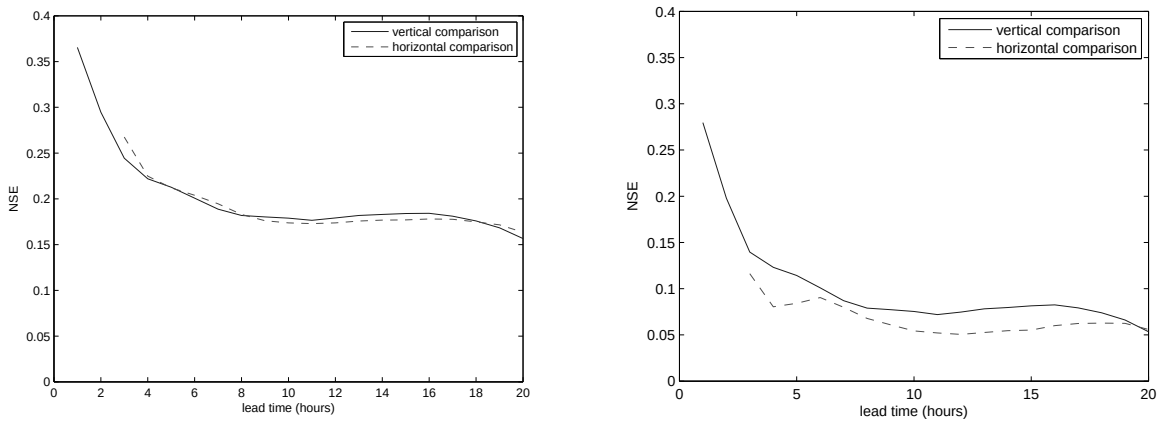


Figure 2.7: The various options to compare forecasts and measured rainfall. M3 is the measured rainfall over a 3 hour period, while F3-L1v, F3-L8v and F3-L6h are the forecasts for lead times of 1, 8 and 6 hours respectively. The first two use the vertical method, while the last uses the horizontal.



(a) All forecasts

(b) Excluding correctly forecast dry days

Figure 2.8: The accuracy of the rainfall forecasts as a function of lead time. The Nash-Sutcliffe efficiency (NSE) of the predicted rainfall amount within a 3 hour period. The measured rainfall is compared with the amount falling at a fixed lead time. (F3-LXv) and within one prediction (F3-LXh).

measured rainfall was obtained from a weighted average of 8 tipping bucket raingauges in the area. This is the same measured rainfall information that is available for the DSS.

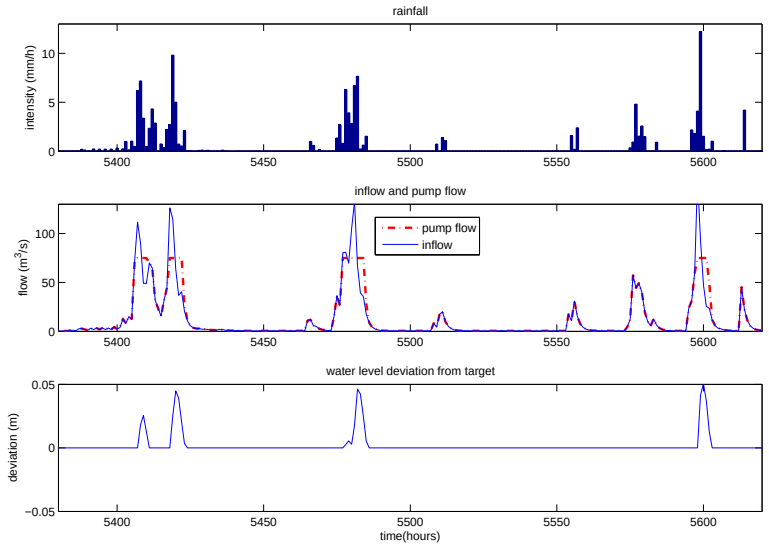
What comparison is relevant depends on the sensitivity of the control decision for those different characteristics of the forecast. For a slow reacting water system for example, it makes sense to compare multi-hour sums instead of single hours. An example of how the forecast accuracy depends on lead time for both the vertical and the horizontal method is given for the 3 hour sum in figure 2.8. The right figure shows the results for the data set where all correct forecasts of no rain have been filtered out. This gives more insight in whether the amounts during events are forecast correctly. The Nash-Sutcliffe efficiency (NSE, Nash and Sutcliffe (1970)), which was used in this analysis, has a number

of drawbacks. The fact that the forecasts have significant skill up to long lead times may be the result of seasonality, which reflects knowledge of the seasonal cycle rather than ability to make good short term forecasts (see Schaefli and Gupta (2007)). Apart from that, the time horizons are formulated in terms of information, which is described by quadratic performance measures like NSE only under certain circumstances. In chapter 5, a more coherent framework for forecast evaluation using information theory is given, which could be applied to determine the time horizons if the forecasts are probabilistic.

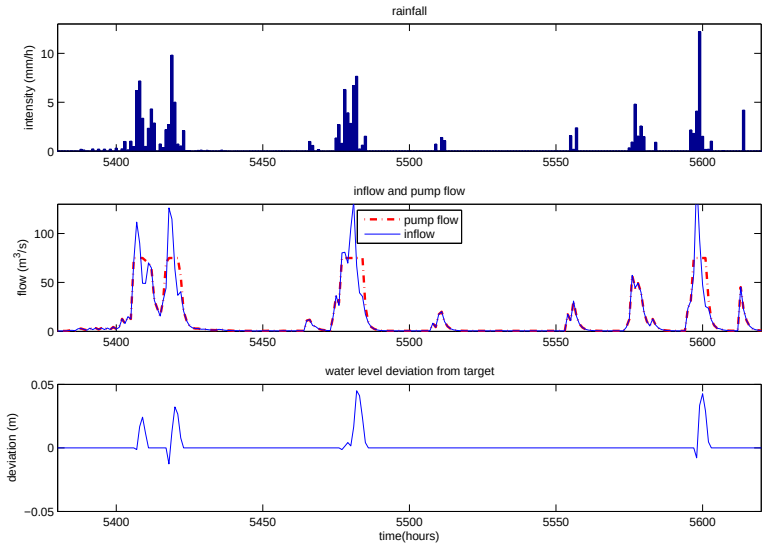
Sensitivity to predictions: the information control horizon

To determine to which extent the controlled system is sensitive to rainfall forecast information, several simulations were made. Because this study focused on the influence of information and uncertainties in rainfall forecasts, uncertainties in water level and rainfall measurements were not considered and the models for the rainfall-runoff process and the storage canal system are assumed to be perfect. In the simulations, both the storage canal system and the MPC controller are simulated. For the MPC controller, different horizons were used for the optimization and the resulting performance in controlling the water levels is compared. The simulations were done using a one hour time step. To be able to compare results with optimal feedback control, the penalty on control actions was set to zero. This led to small differences in behavior compared to the real controller, but does not significantly affect the results on time horizons. First, to determine the Information Control Horizon, simulations were made in which the predictions that the MPC controller uses are taken from measured rainfall data, that is also used to simulate the storage canal system. This corresponds to perfect foresight, which allows the controller to optimally anticipate extreme events. This is only true in case the optimization horizon is larger than the Information Control Horizon. If it is shorter, the controller does not see events in time to take the necessary anticipatory actions. In other words, the Information Control Horizon is the point where extending the optimization horizon does not improve performance anymore. The results for four different optimization horizons with perfect foresight are shown in figure 2.9. It is visible that for shorter horizons, the controller can not anticipate the high inflow in time to avoid relatively high positive deviations. For longer optimization horizons, the controller lowers the water level beforehand, leading to some negative and some positive deviations, which have a smaller total quadratic penalty. The results for the penalty as a function of the optimization horizon are shown in figure 2.11. As can be seen from the dashed line in that figure, for the Delfland storage canals, the Information Control Horizon is approximately 6 to 7 hours, because extending the optimization horizon beyond that point does not improve control performance anymore.

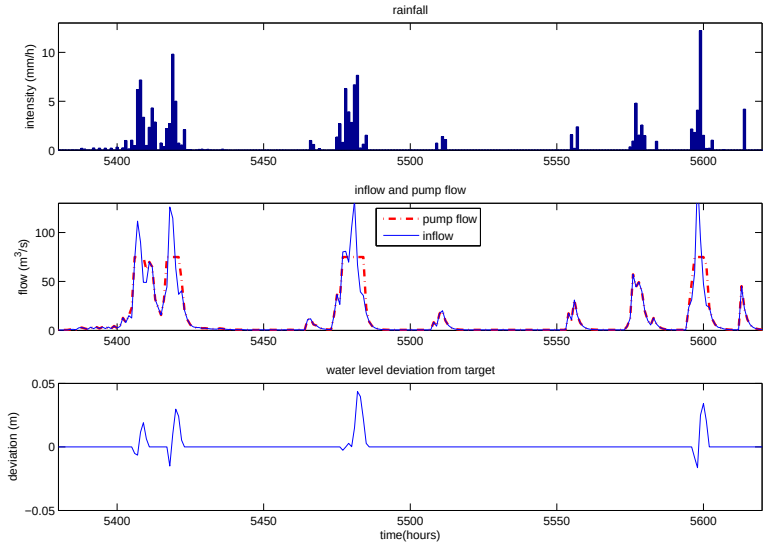
The results for a optimization horizon of 1 hour correspond to feed forward control. This means the controller does not anticipate future inflows, but exactly knows the current inflow. This already assumes perfect rainfall measurements and knowledge of the rainfall-runoff process. Feed forward control is an improvement compared to feedback control, where control actions are based on measured water levels in the system, which leads to actions that always lag behind. As long as the inflow does not exceed the outflow, a feed forward controller can perfectly maintain the water level at target level, but when the



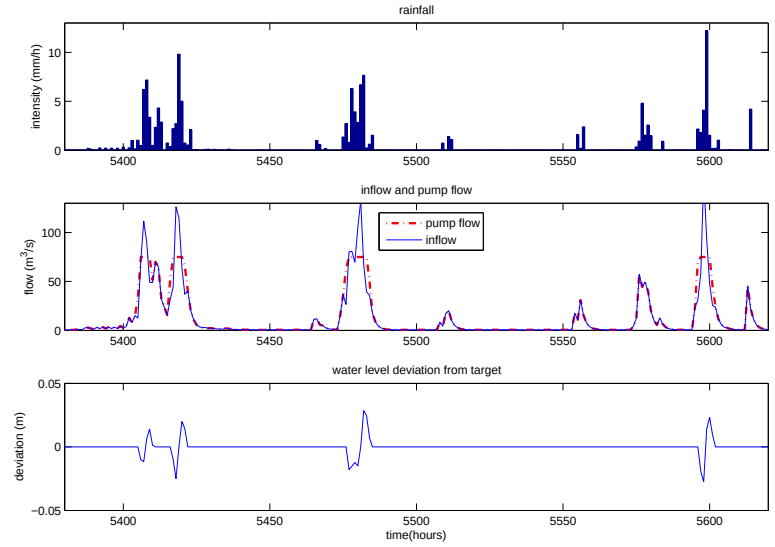
a: 1 hour perfect foresight



b: 2 hours perfect foresight



c: 3 hours perfect foresight



d: 12 hours perfect foresight

Figure 2.9: Influence of the prediction horizons on the control flows, when the predictions are perfect. Figures a to d give the results for the control flows (which slightly differ in timing) and resulting water levels for various time horizons.

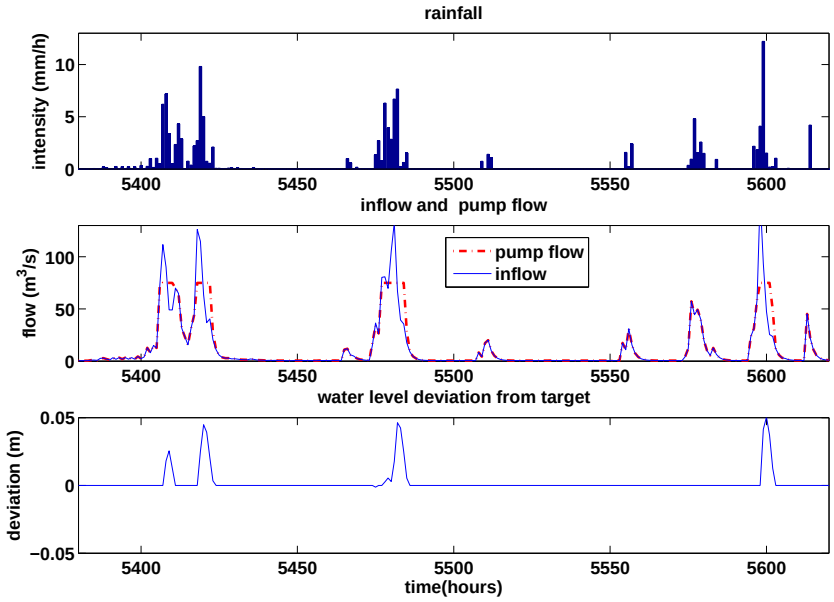


Figure 2.10: The behavior of the controller when fed with the forecasts that were actually available, which are not perfect. It becomes clear that none of the events were anticipated in time to lower the water level, resulting in larger mean squared deviations from target level compared to the perfect predictions. The optimization horizon was set to 12 hours, which is long enough for perfect anticipation if the forecast were perfect (see dashed line in Fig. 2.11).

outflow is exceeded, water levels start to deviate. From the performance, which is proportional to resulting water level squared deviations, it can be seen that anticipation based on perfect forecasts would improve water level control by a factor of about 3 compared to feed forward control, while feed forward control already outperforms an optimally tuned (linear-quadratic regulator) feedback controller (not shown) by a factor 4.5. Secondly, simulations were done in which the controlled system was fed with the measured rainfall, while the controller was fed with the forecasts that were actually available at the time, instead of perfect forecasts (see Fig. 2.10). From the results of this second experiment, it becomes clear that the uncertainties in the forecasts actually reduce most of the value of anticipating. This also shows the huge gain that can be made by improving forecasts, especially for the first 6 hours. For these simulations, the performance as function of lead time is shown as the solid line in figure 2.11.

2.6.3 Results

Analysis of the RMSE of forecast compared to measured rainfall has shown that rainfall forecasts have some predictive power for lead times up to at least 20 hours. The Information Prediction Horizon for this method of forecasting rainfall is more than 20 hours. Analysis of controller performance under perfect foresight as a function of optimization horizon shows that the current control action is only sensitive to events within the first 6 hours. Therefore, extending the optimization horizon beyond 6 hours does not improve control performance. This defines the Information control horizon. However, the value of

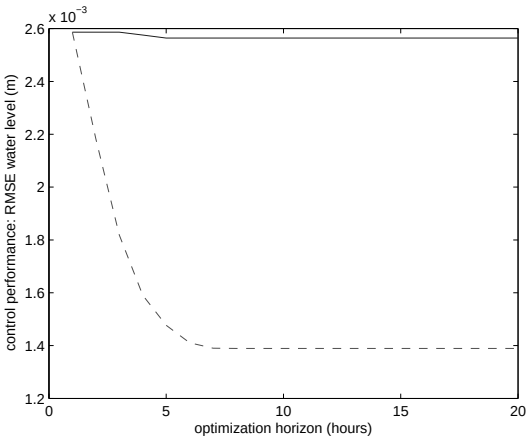


Figure 2.11: The performance of the controller at maintaining the water level as a function of the length of the optimization horizon. The performance is measured in terms of root mean squared error (RMSE), where a lower value indicates better performance. The dashed line represents performance with perfect forecasts, while the solid line was obtained using the actual, imperfect forecasts that were available.

the predictive power in the forecasts for controlling the water levels disappears already at lead times of 5 hours due to prediction errors. Although predictive power is there and the controller is sensitive to events between 5 and 7 hours ahead, no gain is made using these forecasts. A probable cause for this is that performance is mainly dependent on a limited number of events for the cases in which anticipation is necessary or appears to be necessary. Even if forecasts have some predictive power over the whole range of events, they might not have this power for correctly forecasting extreme events that need to be anticipated more than 5 hours beforehand. The way the "information" in the forecasts was measured apparently does not correspond one to one with the useful information for controlling the water system (related discussions can be found in chapter 5 and section 6.6). Another conclusion from simulations with real forecasts is that the value of anticipating largely vanishes as a result of forecast inaccuracies. Comparison between performance with perfect and with real forecasts shows a potential gain in performance by a factor 3 that could be made by improving forecasts. Even without anticipation, simulated feed forward control using perfect knowledge of the current inflow showed performance to be improved by a factor 4.5 compared to LQR-derived optimal feedback control, which only uses measured water levels as inputs. This shows the importance of good rainfall measurements, accurate rainfall-runoff models and real time availability of flow data from the polder pumping stations.

2.7 Certainty equivalence

If uncertainties in the forecast information influence the control decisions, the challenge is to find the best decision, given the probability distribution of the inputs. If a water system is certainty-equivalent, the optimal operation can easily be found by optimizing

the control actions, given the best deterministic forecast, i.e. the expected value of the input (Eq. 2.13). However, the necessary conditions for certainty equivalence require that (Philbrick and Kitanidis, 1999):

1. The objective function is quadratic
2. System dynamics are linear
3. There are no inequality constraints
4. Uncertain inputs are normally distributed and independent

For the Delfland case, the last two conditions are not fulfilled (maximum pump flow, errors are correlated), making it a non-certainty equivalent problem. This means that:

$$u^*_{\text{deterministic}} = \arg\min_u J\left(x, u, E\left\{Q_d\right\}\right) \tag{2.13}$$

$$u^*_{\text{stochastic}} = \arg\min_u E_{Q_d}\left\{J\left(x, u, Q_d\right)\right\} \tag{2.14}$$

$$u^*_{\text{deterministic}} \neq u^*_{\text{stochastic}} \tag{2.15}$$

in which x is the state (in this case e), u is the control vector (in this case Q_c), Q_d is the uncertain inflow, J is the objective function (see equation 2.8) and E is the expectation operator. In this case, optimal operation, given the uncertainties and the available information, is only possible by using a stochastic approach (Eq. 2.14) to the optimization problem, routing the uncertainties to the objective function, which will then express risk instead of costs (Weijs et al., 2006; van Overloop et al., 2008). By minimizing this objective function, expected damage is minimized. This can be approximated by using different inflow scenarios as a discrete representation of the uncertain inflow in a multiple model configuration of MPC (van Overloop, 2006; van Overloop et al., 2008).

2.8 Multiple model predictive control (MMPC)

Model predictive control calculates the optimal actions, given a certain predicted sequence of disturbances. In the past sections it has been shown how uncertainties in the predicted disturbance exist and can negatively affect the performance of the MPC controller. Because the Delfland control problem is not certainty equivalent, the best control action is not equal to the control action that would be best for the most likely disturbance. Instead, the control problem is stochastic, meaning that the action sought should minimize the *expected* value of the cost function over the optimization horizon. A possible way to achieve this in a standard model predictive control setup is to extend the internal model in MPC to contain multiple copies of the system. Each copy can than be fed with a different scenario for the future disturbance to the system, where the spread in scenarios reflects the uncertainty. This method combines well with ensemble forecasts, which have become quite common in meteorology and flood forecasting. This approach is referred as Multiple Model Predictive Control (MMPC). MMPC and its application to the Delfland system is described in more detail in Weijs (2004); van Overloop (2006); van Overloop et al. (2008). This section introduces the basic concept through which risk based water system operation can be implemented in the objective function. Section 2.9 analyzes the method in terms of information flows about future decisions.

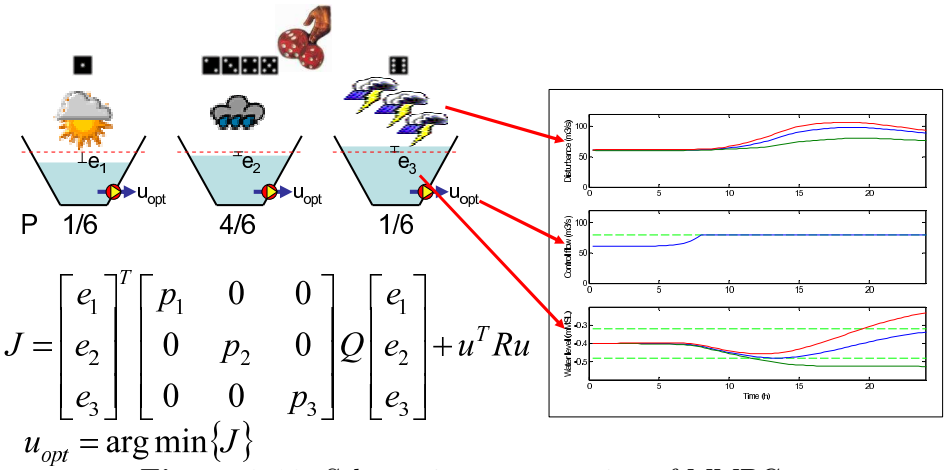


Figure 2.12: Schematic representation of MMPC.

In MMPC formulation, the number of states is multiplied by the number of different inflow scenarios. In the optimization, a sequence of control actions is sought that leads to reasonable results for each copy of the model (see Fig. 2.12). For each possible scenario, the states of the model develop differently over the optimization horizon. If each scenario j represents a certain probability measure p_j in the probability space of future events, the expected value of the cost function (risk) can be calculated by

$$J = \sum_{i=0}^{h} \left[\sum_{j=1}^{n} \left\{ \mathbf{x}_j^T(t+i|t) p_j Q \mathbf{x}_j(t+i|t) \right\} + \mathbf{u}^T(t+i|t) R \mathbf{u}(t+i|t) \right] \tag{2.16}$$

where the symbols are equal to those used in eq. 2.8, and j is the counter for the n different models or scenarios. Minimizing this objective function by a certain control sequence minimizes risk in the open loop behavior. Subsequently, the first action of the sequence can be executed and the optimization is repeated for the next control time step, using the latest information from measurements and updated predictions. In this research, a stochastic inflow model generated 50 inflow scenarios. These where distilled into 3 representative scenarios with unequal weights, in order to reduce the computational cost for the MMPC algorithm. More details can be found in van Overloop et al. (2008). New research focuses on more sophisticated tree based scenario reduction techniques (Raso et al., 2010).

2.9 The problem of dependence on future decisions

The optimization problem that has to be solved in one control time step in the MMPC approach can be viewed from the Dynamic Programming perspective. In fact, the current decision is sought, that minimizes the sum of current and future costs. Because the uncertain future is described by multiple scenarios, the future cost is an expected value or risk. This is calculated by weighting the consequences of each scenario with their probability. However, these future costs not only depend on the external disturbance scenario, but is

also determined by how we react to it in the future. The quality of the decisions in the future influences the decision now. The quality of the decisions is partly influenced by the information that these future decisions are based on.

In the MMPC formulation, the tacit assumption in the calculation of the benefits / penalty of the scenarios is that the whole control sequence is decided to be optimal for the whole range of scenarios, but not changed at the next control steps. In other words, it is assumed that no new information is taken into account during the optimization horizon. The sequence is thus optimal for the case where the operator has a day off tomorrow and programs the pump flows for the whole next day and returns to change the settings only after 24 hours.

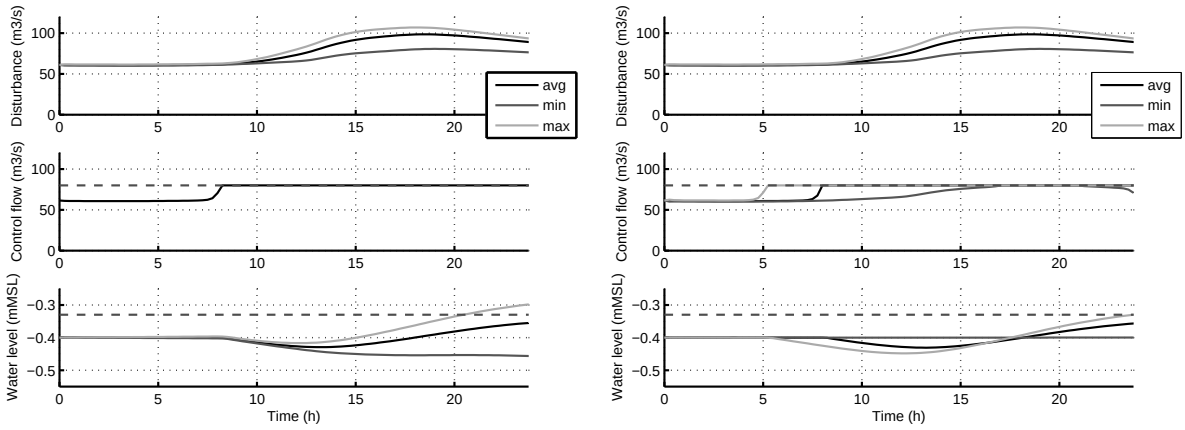
The true closed loop behavior of the control system, however, is that the optimization is executed every control timestep. The true sequence of control actions will therefore be more adapted to the scenario that becomes reality (see figure 2.13b). The true control action on every future timestep will take into account the extra information about the state of the water system at that moment and the information in the updated inflow forecasts. The feedback on the actual water levels that occur for each scenario will result in smaller deviations than assumed in the MMPC formulation. The value of future decision is thus under-estimated in this formulation.

An alternative formulation for the MMPC controller requires only the first control action to be equal for all scenarios. All control actions that follow the first are different for each scenario. The tacit assumption here is that after the first control action, which takes into account uncertainty, the future control actions are based on perfect forecasts. These perfect forecasts are the opposite end of the spectrum compared to the absence of new information that was assumed in the previous formulation. In this new formulation, the value of future decisions is thus over-estimated. This approach, although coming from MPC rather than SDP, is equivalent to a special case of Sampling Stochastic Dynamic Programming (SSDP), as proposed by Faber and Stedinger (2001). In section 4 of that paper, it is also argued that assuming a case where no transitions between scenarios occur, leads to an overestimation of future benefits. The authors propose Bayesian updating of the transition probabilities as a remedy.

The correct estimation of the optimal control action thus has to take into account how much information will be available for the future control actions to estimate their value. This amount of information lies somewhere between no information and perfect information, as assumed in the previous two formulations. In chapter 7, the problem of accounting for future availability of information is further addressed in the context of SDP.

2.10 Summary and Conclusions

Analysis of the control of the Delfland storage canal system revealed that anticipation on heavy rainstorm events is necessary. The sequential decision process of finding optimal



(a) no new info (b) perfect new info

Figure 2.13: The difference between MMPC formulations that assume no new information and perfect new information.

sequences of pump flows can be solved off-line, by finding an optimal decision rule, or online, by finding a sequence of optimal decisions for a finite horizon, based on predictions of inflow into the system. Model Predictive Control (MPC) can solve this optimization problem, making use of an internal model that predicts behavior of the system over the control horizon. Uncertainties in measurements and predictions affect performance of the MPC setup negatively. The largest uncertainties are present in forecasts of events far into the future. At the same time, sensitivity of the current action to errors in prediction diminishes with lead time. Two conceptual time horizons regarding these effects were defined and rules for choosing the optimization horizon of a controller were formulated in terms of these horizons. For the Delfland system, the first eight hours of the prediction are most important for performance, which can be improved significantly by more accurate inflow predictions. The rainfall predictions contain information up to at least 20 hours into the future, according to the Nash-Sutcliffe Efficiency. In chapter 5, a more rigorous approach to evaluating the amount of information in forecasts will be described, using a mathematical theory of information, which is introduced in the next chapter.

Due to constraints on pump flow, the control problem is not certainty equivalent. This demands a stochastic approach, which minimizes risk associated with decisions. This can be implemented in MPC by an extension to multiple parallel models (MMPC). Analysis of the MMPC formulation revealed an intricate interplay between future decisions, current decisions, information and risk. This issue will be revisited in chapter 7, where stochastic dynamic programming is applied to analyze the value (utility) of water as function of forecast information. Information can be seen as a key ingredient to good decisions and appropriate operation of water systems. Taking this viewpoint, the main part of this thesis deals with the application of formal theories about information and uncertainty. The next chapters will first introduce information theory and then show its applications and insights regarding forecasting and modeling for decisions.

Chapter 3

Uncertainty or missing information defined as entropy

> *"My greatest concern was what to call it. I thought of calling it 'information', but the word was overly used, so I decided to call it 'uncertainty'. When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, 'You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one really knows what entropy really is, so in a debate you will always have the advantage.' "*
> - Claude Shannon (Scientific American 1971, volume 225, page 180)

The word "uncertainty" in daily conversation is a qualitative notion meaning lack of certainty. In many uses of the word, the notion is also something quantitative: "There is too much uncertainty to take a decision"; "The uncertainty of forecasts tends to grow with lead-time". Both these expressions reflect the idea that there is such a thing as "the amount of uncertainty". To be useful, the ordinary language concept of uncertainty needs to be explicated. A highly convincing explication of uncertainty was provided by Claude Shannon in his seminal paper "A Mathematical Theory of Communication" (1948), in which he introduced the concept of entropy as a measure for uncertainty. Furthermore he introduced the concepts of conditional entropy, entropy rate and relative entropy (although the term was later used for another concept), and stated and proved many of the main theorems relating to these quantities and their application to the transmission of information. In this thesis, information theory is extensively used as a framework of reference and a basis for methodological development. This chapter therefore introduces the basic measures and theorems, along with some additional interpretations.

3.1 Uncertainty and probability

In the English language, certainty and uncertainty are not necessarily opposites on two ends of a scale. One can be very certain that one particular outcome of an uncertain event will happen or one can be highly uncertain about the outcome of an uncertain event. Both

cases refer to certainty and uncertainty as a quantitative properties. See for example the statements:

1. I'm quite certain that it will rain tomorrow
2. I'm very uncertain about whether it will rain tomorrow or not

Although both expressions seem to refer to the same uncertainty-certainty scale, they do not. When relating the expression of uncertainty to statements in terms of probabilities, the difference becomes clearer. The first statement could imply that my subjective probability for rain tomorrow is 90%, but can not mean that it is 10%. Certainty in this statement means high probability attached to the outcome following the word "that". So "I'm quite certain" is synonymous to "I find it highly probable". In the second statement, however, uncertainty does not refer to the opposite, i.e. low probability of rain, but to close to equal probabilities on both outcomes, e.g. 50% chance of rain. If we would obtain new information that makes us revise our probability of rain to either 10% or 90%, in both cases we would say that we are more certain about whether it will rain tomorrow than in the 50%-50% case. In this thesis, the word uncertainty refers to the meaning in the second statement. The first concept would be more accurately conveyed by the statements in terms of the improbable-probable scale.

3.2 The uncertainty of dice: Entropy

A quantitative definition of uncertainty is related to choice and possibilities. The more possibilities there are, the more uncertainty is associated with a choice. The outcome of throwing a fair die is thus more uncertain than the outcome of a fair coin toss. The uncertainty is related to a lack of information about which of the possibilities is the truth. Therefore, intuitively, uncertainty can be equated with missing information or the amount of information that needs to be gained to obtain certainty. One could express this in terms of the expected number of binary questions that need to be asked. For the coin toss, one binary (yes-no) question is enough to determine the outcome ("is the outcome heads?"). For the die, one question could be enough to know the answer ("is the outcome six?"), but with a probability of $\frac{5}{6}$ the answer will not be enough to determine the outcome and additional questions are necessary. One could continue asking questions like "is the outcome four?" to obtain the answer in five or less questions (see left of Fig. 3.1), with an expected number of questions of

$$E\{\text{number of questions}\} = \frac{1}{6} \times 1 + \frac{1}{6} \times 2 + \frac{1}{6} \times 3 + \frac{1}{6} \times 4 + \frac{2}{6} \times 5 = \frac{10}{3}$$

This is, however, not the optimal way of asking the questions. A scheme where each question divides the possibilities approximately equally leads to certainty in less than three questions, with an expected number of $\frac{8}{3}$ questions (see right of Fig. 3.1). Intuitively, the answer to one yes-no question could serve as a unit of information. Shannon called this unit one "bit" of information, following a suggestion by J.W. Tukey (Shannon, 1948). One bit corresponds to one binary digit (0 or 1) and its derived units, from byte (8 bits) to terabyte, have now become familiar to most people who have ever touched a computer.

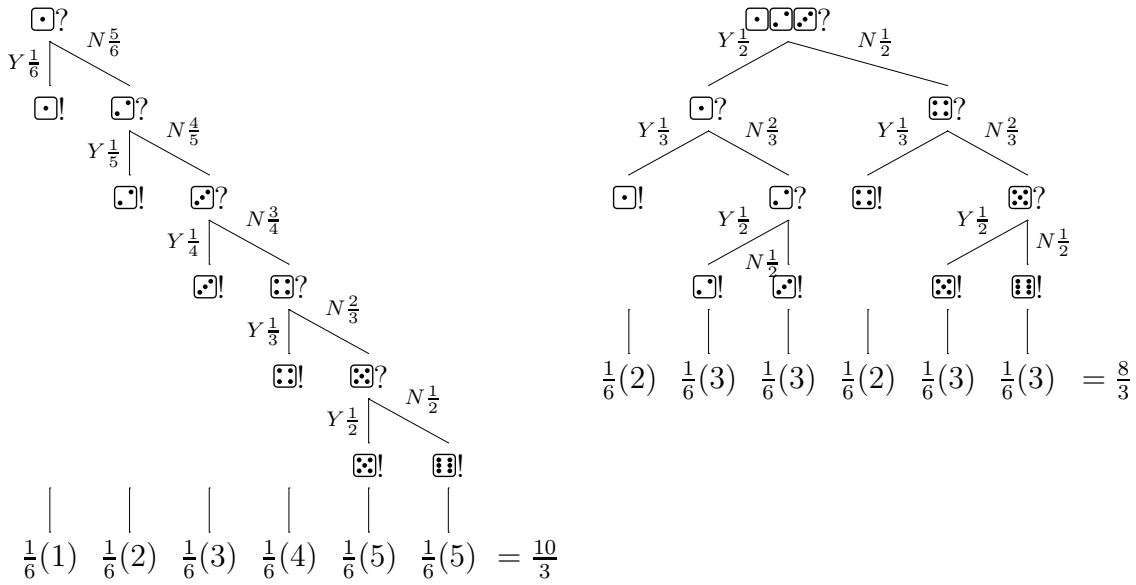


Figure 3.1: The expected number of questions to determine the outcome of a fair die for a risky questioning scheme (left) and an optimal questioning scheme (right).

The expected number of questions does not yet define uncertainty. The missing information still depends on how clever the person asking those questions is (this is the basis of the game "Guess Who?"). The minimum expected number of questions, however, is a function of the probability distribution alone. Unfortunately, this measure of missing information still leaves us with two problems. The first problem is why we would restrict outselves to asking binary questions. The second problem is the counter-intuitive result that a coin toss would always present one bit uncertainty, regardless whether it is a fair coin, or an extremely biased coin where we are 99% sure it will land on heads. In our intuition, the heavily biased coin would present less uncertainty.

Shannon (1948), instead of presenting this interpretation in terms of questions, took a more rigorous approach. He started from three reasonable properties that would be required for a measure of uncertainty, if it would exist. If all that is known about the outcome of an event are the probabilities attached to the different possible outcomes, the measure for uncertainty should be a measure of those probabilities: $H(p_1, p_2...p_n)$. In the case of the fair die, $p_1...p_6$ would all be $\frac{1}{6}$. The requirements for H are (quoting Shannon, 1948) :

1. H should be continuous in the p_i.
2. If all the p_i are equal, $p_i = \frac{1}{n}$, then H should be a monotonic increasing function of n. With equally likely events there is more choice, or uncertainty, when there are more possible events.
3. If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H.

The first of these requirements corresponds to the intuitive notion that adding an infinitely small bias to the die should not lead to a large jump towards more certainty.

The second requirement is the most self-evident of the three and needs no further explanation. The third is requirement becomes clearest when seen in light of the expected number of questions-interpretation. It was explained before that this interpretation was unsatisfactory because the information gained in one answer does not depend on the prior probabilities. However, when the information gained in one question is equated with the uncertainty it resolves (to be defined by the measure that is sought), it will depend on the probabilities. Furthermore, uncertainty can then be defined in a way that does not depend on which questions are asked (this is the interpretation of requirement 3). Instead, the expected information gained in each question depends on the question asked (and the actual information gained depends on luck, cf. "is the outcome six?"). Shannon proved that the only measures simultaneously satisfying these three requirements are of the form

$$H = -K \sum_{i=1}^{n} p_i \log p_i \tag{3.1}$$

Where K is a positive constant that determines, together with the base of the logarithm, the unit in which uncertainty is measured. For $K = 1$ and a base 2 logarithm, the uncertainty is measured in bits. In the rest of this thesis, logarithms have base 2 unless specified otherwise. Shannon named his measure "entropy", because the expression is similar to the concept of entropy in statistical thermodynamics as interpreted by Boltzmann and Gibbs. Some further explanation of this connection will be given in section 3.8. Just after deriving the entropy measure for uncertainty Shannon writes:

> *"This theorem, and the assumptions required for its proof, are in no way necessary for the present theory. It is given chiefly to lend a certain plausibility to some of our later definitions. The real justification of these definitions, however, will reside in their implications."*

Indeed, the implications were many, ranging from data compression in computer science to portfolio theory for the stock market. The fact that fundamental theorems in all these fields can be stated in terms of the measures introduced by Shannon, exposes a surprising unity of underlying principles. It allows an intuitive understanding of connections between methods in many fields, including some of the methods presented in this thesis.

For uniform distributions, H simplifies to $\log n$. The entropy of a fair die is therefore approximately 2.585 bits. The entropy is a lower bound for the minimum expected number of binary questions. This lower bound is achieved if the answer to every binary question in the scheme resolves one bit of uncertainty. This is only the case if the prior information about the answer of each question is as uncertain as a fair coin toss, i.e. 50-50. No questioning scheme about a fair die can meet this property (see figure 3.1). However, as proven in a theorem related to data compression, it is always possible to find a questioning scheme that can achieve an expected number of questions within 1 bit from the bound. The questioning scheme in the right of figure 3.1 is an example of such a scheme for the fair die, achieving an expected number of questions of 2.667, where the entropy bound is 2.585.

3.3 Side information on a die: Conditional Entropy

3.3.1 Conditioning, on average, reduces entropy

The entropy measure introduced in the previous section represents the uncertainty a gambler has about the outcome of a fair die. The uncertainty is a function of the probabilities attached to each outcome and those probabilities follow from the beliefs of the gambler. This means that even after the die has been thrown and the outcome is fixed, the uncertainty for the gambler is not eliminated until he sees the outcome. Conversely, this also means that if the die is thrown with a very precisely controlled and known velocity and spin, the probability distribution changes according to this state of knowledge and uncertainty can be reduced significantly, before the die has been thrown (see e.g. Diaconis et al. (2007) where this is shown for a coin toss). This makes the propensity interpretation of probability, where it is seen as a property of the die (Popper, 1959), untenable. Even when the propensity refers to the whole experimental setting, the interpretation needs many ad hoc complications. A far more simple interpretation of probability is that it is just in our heads, and reflects our incomplete information about the outcome. The probability thus depends on how much we know about the initial conditions of the throw or situation after the throw. An extensive treatment of this view on probability, which is also adopted in this thesis, can be found in Jaynes (2003).

Given this probability interpretation, we can now imagine the following situation, in which some side information on the die is available. After the die has been thrown, the gambler is allowed to observe the face of the die facing him. This observation, although not directly revealing the outcome, gives some information on which side faces upwards. This information reduces the uncertainty of the observing gambler, because it rules out the observed face as an outcome. In absence of other information, the gambler should assign equal probabilities to the remaining possible outcomes. For example, the conditional probability $P(X|y = ⚀) = \left(\frac{1}{5}, 0, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}\right)$, where the random variable X denotes the outcome of the throw, given the observed side-face $y = ⚀$. The entropy of the conditional distribution $H\left(X|x \neq ⚀\right)$[1] is now the same as if the die has only five faces (2.32 bits).

A clever gambler, however, can combine the observation with his prior knowledge on the die to reduce his uncertainty even further: the sum of two opposing faces is always seven. Using this model of the die, ⚅ can be ruled out as well, yielding $P(X|y = ⚀, \text{model}) = \left(\frac{1}{4}, 0, \frac{1}{4}, \frac{1}{4}, 0, \frac{1}{4}\right)$ and a remaining uncertainty of only 2 bits. In this case the gambler conditioned his probability estimate both on the observation and on his model. One of the theorems of information states that conditioning, on average[2], always reduces entropy, sometimes also stated as *"Information can't hurt"* (Cover and Thomas, 2006), although

1. Note that, for ease of notation, a shorthand notation is introduced here, in which the information-theoretical measure is written with a random variable as argument, but in fact is a function of its probability distribution.
2. Note that by taking the average, we arrive at the conditional entropy $H\left(X|Y\right)$ as defined in the next section, which is not the same as the entropy of the conditional distribution $H\left(X|y\right)$. The latter entropy could actually be larger than $H\left(X\right)$ for some particular values of y, although this does not happen in the example with the die, because there conditioning entails eliminating possible outcomes.

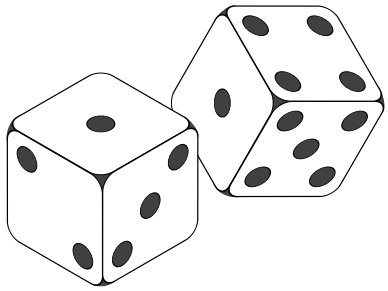


Figure 3.2: Observing the dice can yield a further conditioning model.

this is not always true in some sense, as will be explained in subsection 3.6.2. The model the gambler used for the die, could have been obtained from the various sources familiar to scientists. The model might have been discovered by another gambler who shared his research in a publication; the model could have been derived by observing the die directly and theorizing about the geometry; and the model could also be an empirical one, stating nothing about the seven-eye-sum theorem or even the shape of the die, but just observing that in a large number of throws, ⚄ never comes up when ⚀ is observed and using that for prediction. The gambler's estimate can in that case be seen as conditional not only on the actual observation, but also conditional on all those observations that were used to infer the model.

In a hydrological context, this is analogous with a flood forecast, based on side-information about the weather and the hydrological conditions in the basin. The side information on those conditions enables better flood forecasts. However, without a model the side information is worthless. Observing long time series of streamflow and weather, the models can be improved. Also direct observation of certain properties of the catchment, like slopes and area, may help, but some remain hidden underground (cf. the fact that the die is cubic vs. the fact that it's density is homogeneous). The more specific the observations are, the better the forecast becomes. Also for our gambler, those observations can still be refined...

3.3.2 Conditional entropy

When looking closer at the observations of the die, the gambler might find a further structure in them. Apart from conditioning on the number of eyes observed, the gambler can use the fact that ⚀, ⚁ and ⚂ are only 2-fold rotational symmetric and not 4-fold, like the other 3 faces. In his observation, he can therefore further distinguish between ⚀ and ⚁. Of the four possible outcomes, given two eyes facing him, only two can occur when $y = ⚁$. As can be seen from figure 3.2, $P(X|y = ⚁) = \left(0, 0, \frac{1}{2}, \frac{1}{2}, 0, 0\right)$ leaving only 1 bit of uncertainty.

Besides looking at the remaining uncertainty for one particular case of side information y, the gambler might be interested in his average or expected remaining uncertainty about X, given side information Y, which is now also a random variable. Because for Y the orientation is also of interest, there are more possible outcomes and $H(Y) > H(X)$. The possible

faces for Y are now ⚀ ⚁ ⚂ ⚃ ⚄ ⚅ ⚅ ⚅ ⚅ ⚅ with $P(Y) = \left(\frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{6}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}\right)$, yielding $H(Y) = 3.085$ bits. When observing the outcome y, 3.085 bits of information is thus gained about Y. When having side information Y on the die, while employing the best possible model, the uncertainty about X can with 50% probability be reduced to 2 bits (when observing ⚀, ⚅ or ⚅) and to one bit when one of the other faces is observed, including its orientation. It can then be said that the conditional entropy of X, given Y is 1.5 bits.

$$H(X|Y) = \sum_{y \in Y} \{P(y) H(X|y)\} = \sum_{j=1}^{m} P(y_j) \sum_{i=1}^{n} P(x_i) \log P(x_i|y_j) \tag{3.2}$$

in which y_j are the m possible states the random variable Y can have (the possible observations for the observed side), and x_i the n possible states the outcome of the throw can have, $P(x_i|y_j)$ denotes the probability of the ith outcome, given an observation of the jth outcome for the observed side and $H(X|Y)$ is the conditional entropy of random variable X, given random variable Y. The conditional entropy measures the remaining uncertainty about X, given Y. The information inequality ("conditioning, on average, always reduces entropy") can now be stated as $H(X|Y) \leq H(X)$, with equality only when X and Y are completely independent. In that last case, Y does not reduce uncertainty about X and therefore contains no information about this variable.

3.4 Relations between uncertainty and information gain: Mutual Information and Relative Entropy

Due to the logarithms used in the entropy-definition of uncertainty, the measures introduced have some intuitive additive properties, which are depicted in figure 3.3. In the figure, two new measures appear: joint entropy and mutual information. The joint entropy is simply the combined uncertainty about more than one variable. It is the missing information to obtain certainty about both X and Y and can be calculated with

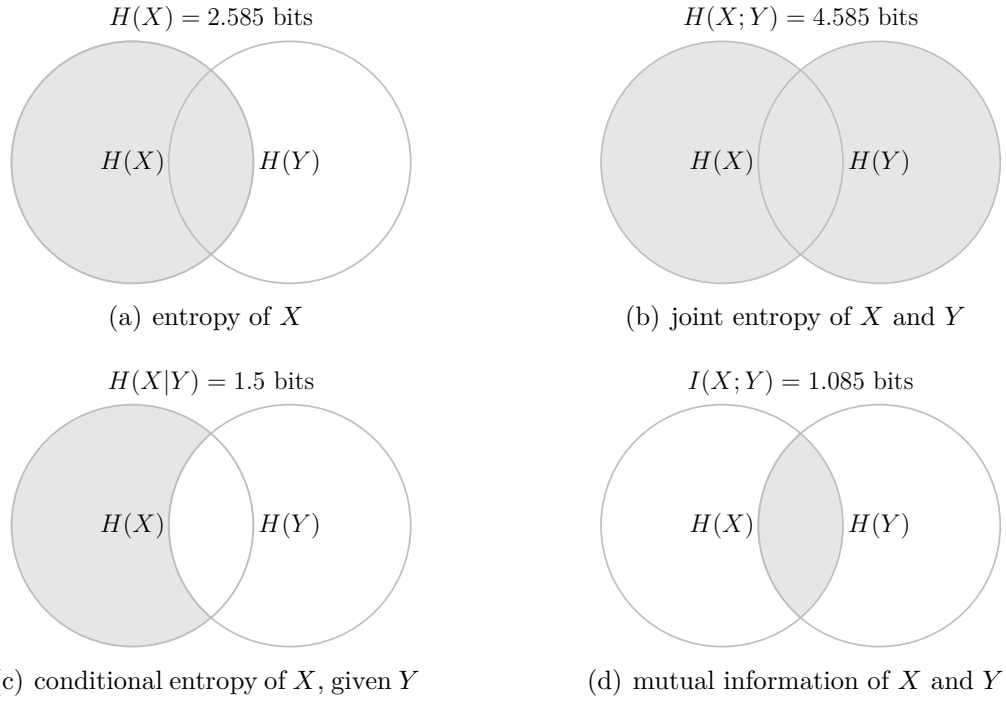
$$H(X;Y) = \sum_{j=1}^{m} \sum_{i=1}^{n} P(x_i, y_j) \log P(x_i, y_j) \tag{3.3}$$

In case the variables are completely independent, the joint entropy is equal to the sum of the entropies of both variables. If the variables are dependent, which is the case shown in the figure, the joint entropy is less than this sum. This is due to the fact that once the questions about Y have been asked and it is known that $Y = y$, the uncertainty about X is not $H(X)$, but $H(X|y)$. On average, the uncertainty about X when also informing about Y is the conditional entropy $H(X|Y)$. This leads to the following equation

$$H(X;Y) = H(X|Y) + H(Y) = H(Y|X) + H(X) \tag{3.4}$$

Furthermore, because this equation shows that X reduces uncertainty about Y exactly by the same amount as Y reduces uncertainty about X, the measure mutual information can be defined as

$$I(X;Y) = H(X) + H(Y) - H(X;Y) \tag{3.5}$$



(a) entropy of X

(b) joint entropy of X and Y

(c) conditional entropy of X, given Y

(d) mutual information of X and Y

Figure 3.3: The relations between the entropy measures of two variables. The numerical values are for the die with side information, using the sophisticated model.

The mutual information defined by this equation can also be calculated directly from the joint probability mass function such as defined in table 3.1.

$$I(X;Y) = \sum_{j=1}^{m} \sum_{i=1}^{n} P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i) P(y_j)} \tag{3.6}$$

It measures the degree of dependency between two random variables. The mutual information $I(X;Y) \geq 0$ with equality only when the variables are completely independent. In the case of the simple model of the die, the mutual information between the number of eyes on the observed face and the outcome is 0.585 bits, while for the sophisticated model using the exact observed pattern of the eyes, it is 1.085 bits. The mutual information in eq. 3.6 can also be interpreted as a measure divergence between the joint distribution of X and Y , $P(x_i, y_j)$, and the hypothetical joint distribution if both variables were independent, which is $P(x_i) P(y_i)$. Equation 3.6 shows that mutual information is the expectation of the logarithm of the ratio between these two probability distributions.

This interpretation of mutual information makes it a special case of a general information-theoretical measure for divergence from one distribution to another. The measure is known as Relative Entropy, Relative Information or Kullback-Leibler divergence and was introduced by Kullback and Leibler (1951). The Kullback-Leibler divergence from distribution $P(X)$ to distribution $Q(X)$ is defined by

$$D_{KL}(P||Q) = \sum_{i=1}^{n} P(x_i) \log \frac{P(x_i)}{Q(x_i)} \tag{3.7}$$

The divergences measures how much more uncertain about X a person is, when having probability estimate $Q(X)$ rather than $P(X)$, while the true probabilities are $P(X)$. The

	1	2	3	4	5	6	$P(Y)$
1	0	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{1}{24}$	0	$\frac{1}{6}$
2	$\frac{1}{24}$	0	$\frac{1}{24}$	$\frac{1}{24}$	0	$\frac{1}{24}$	$\frac{1}{6}$
3	$\frac{1}{24}$	$\frac{1}{24}$	0	0	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{1}{6}$
4	$\frac{1}{24}$	$\frac{1}{24}$	0	0	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{1}{6}$
5	$\frac{1}{24}$	0	$\frac{1}{24}$	$\frac{1}{24}$	0	$\frac{1}{24}$	$\frac{1}{6}$
6	0	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{1}{24}$	0	$\frac{1}{6}$
$P(X)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	

	1	2	3	4	5	6	$P(Y)$
⚀	0	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{1}{24}$	0	$\frac{1}{6}$
⚁	0	0	$\frac{1}{24}$	$\frac{1}{24}$	0	0	$\frac{1}{12}$
⚂	$\frac{1}{24}$	0	0	0	0	$\frac{1}{24}$	$\frac{1}{12}$
⚃	$\frac{1}{24}$	0	0	0	0	$\frac{1}{24}$	$\frac{1}{12}$
⚄	0	$\frac{1}{24}$	0	0	$\frac{1}{24}$	0	$\frac{1}{12}$
⚅	$\frac{1}{24}$	$\frac{1}{24}$	0	0	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{1}{6}$
⚅	$\frac{1}{24}$	0	$\frac{1}{24}$	$\frac{1}{24}$	0	$\frac{1}{24}$	$\frac{1}{6}$
⚅	0	$\frac{1}{24}$	0	0	$\frac{1}{24}$	0	$\frac{1}{12}$
⚅	0	0	$\frac{1}{24}$	$\frac{1}{24}$	0	0	$\frac{1}{12}$
$P(X)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	

Table 3.1: The joint probability mass function of X and Y in case of the die for the simple model (left) and the sophisticated model (right) of the die.

measure is referred to as a divergence and not a distance, because it is not symmetrical and does not satisfy the triangle inequality. In mathematical terms: $D_{KL}(P||Q) \neq D_{KL}(Q||P)$ and $D_{KL}(P||Q) \nleq D_{KL}(P||R) + D_{KL}(R||Q)$. The explanation for the asymmetry is that the divergence depends on which of the distributions is (considered to be) the true one. This is the distribution that is in the first argument and which is used to calculate the expectation. Further interpretation of this measure and the information-theoretical measures is given in the next two sections. In the first it is related to gambling and in section 3.6 it is related to the concept of surprise.

3.5 Rolling dice against a fair and ill-informed bookmaker

The information-theoretical measures introduced in this chapter can be interpreted in terms of gains in a gambling game against a fair bookmaker. These interpretations are presented here along with some essential definitions. The derivation of these results is outside the scope of this thesis and can partly be found in Cover and Thomas (2006). A fair bookmaker is defined as a bookmaker whose odds offered for the various possible outcomes do not lead to any expected gain or loss in an individual bet, if his probability estimate is correct. For example, for a fair die, a fair bookmaker could offer 6 times the stakes that the gambler put every outcome, leading to an expected gain of zero for both the gambler and the bookmaker. When \$1 is bet on the outcome ⚅, and the bet was correct, the bookmaker keeps the \$1 and pays out \$6 to the gambler. If bet was wrong, the bookmaker keeps \$1 and does not pay out anything.

The odds the fair bookmaker offers reflect his subjective probabilities. For every outcome, the probability estimate of the bookmaker is $h_i = 1/r_i$, where r_i are the odds offered for outcome i and h_i is the bookmaker's subjective probability for outcome i. The bookmaker is said to be fair if

$$\sum_{i=1}^{n} \frac{1}{r_i} = 1 \tag{3.8}$$

This indicates that a bookmaker can be fair, regardless of what his probability estimate is, as long as the odds offered reflect some probability distribution where the sum of probabilities on all possible events is one. From his own perspective, the expected gain of

a fair bookmaker is zero in each individual bet, regardless of where the gambler chooses to put his money. However, he can still expect to make money in a long series of bets, if the gambler has worse probability estimates than he has. As will become clear from the remainder of this section, this is related to the fact that a gambler can not bet more money than he possesses.

Suppose a gambler, who starts the game with \$1, wants to maximize his gain in a long sequence of repeated bets against this bookmaker. Because the bets are repeated, the gambler can reinvest all the capital he accumulates in previous bets in each new bet. As a consequence, if he loses all his money, he can not continue gambling and the bookmaker wins. To guard against this situation, he should always put some money on each outcome that is not impossible. Kelly (1956) showed that the best strategy for the gambler is to follow a proportional betting strategy, which means that he should "put his money where his probability is". A gambler betting on a fair die should therefore equally split his capital into six parts and spread his stakes over all possible outcomes. In general, the fraction of the gambler's capital bet on each outcome should be equal to his probability estimate for that outcome. One remarkable consequence of this theorem is that the best strategy for the gambler does not depend on which fair odds are offered, but only on his own beliefs. Some further background on this result and the relations with information-theoretical measures can be found in appendix C, where the gambling interpretation is related to weather forecasts which are the subject of chapter 5. The gambling context is now used to give some additional interpretation to the information-theoretical measures introduced in this chapter, some proofs can be found in Cover and Thomas (2006).

A gambler can make money only if his probability estimates are better than the bookmaker's. If they would be worse than the bookmaker's, he would better not bet. However, he can never knowingly have a worse estimate than the bookmaker, because if he thought the bookmaker had a better estimate than he, he could just adopt the bookmaker's estimate, which can be read from the odds. In that case, the gambler is sure to not win or lose any money.

For the gambler the game becomes interesting when he (thinks he) has side information. If the gambler observes one side and knows the simple seven-eye-sum theorem, he can use this information in his bets against the ignorant, fair, and soon to be poor bookmaker. The mutual information between the side and the outcome, $I(X;Y)$ is 0.585 bits. Because the bookmaker assumed the full uncertainty in setting his odds, the gambler's capital will now on average grow with a factor $2^{I(X;Y)} = 1.5$, per bet. After 30 bets his expected capital will be \$191551. For one single bet, in which the gambler for example observes ⚂ and the bookmaker offers naive fair odds, the gamblers estimate $P_g\left(X|y=⚂\right) = \left(\frac{1}{4}, 0, \frac{1}{4}, \frac{1}{4}, 0, \frac{1}{4}\right)$ and the bookmaker's estimate is $P_b\left(X\right) = \left(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right)$. The gambler then estimates the exponent of his own expected gain at $D_{KL}\left(P_g||P_b\right) = 0.5850$ (a factor 1.5 per bet), while the bookmaker will think the exponent of the gambler's expected gain is $-D_{KL}\left(P_b||P_g\right) = -\infty$, which means he expects the gambler to reduce his capital from \$1 to \$0 in the long run. The asymmetry in Kullback-Leibler divergence can thus be seen as originating from a different view on the "true" probabilities: the bookmaker and the gambler can not both be right. The gambler is more right in this case and will win, because he

has more information than the bookmaker. Also an independent third party can make an estimate of the gambler's winnings. Suppose an experienced second gambler observes the same ⚂, but has knowledge of the sophisticated model of dice. His probability estimate will be $P_{g2}(X|y = ⚂) = \left(0, 0, \frac{1}{2}, \frac{1}{2}, 0, 0\right)$. In his estimate, the first gambler is expected to receive

$$\text{capital bet} \times \left(2^{D_{KL}(P_{g2}||P_b) - D_{KL}(P_{g2}||P_g)}\right) = 1.5 * bet \tag{3.9}$$

meaning that he earns half of his bet as winnings. In this case, the better informed second gambler confirms the own expectation of the first gambler.

3.6 Interpretation in terms of surprise

Apart from the axiomatic derivation, the interpretation in terms of gambling gains, and the interpretation related to the expected minimum number of questions, an intuitive natural language view of information theory is now presented in terms of surprise (Tribus, 1961). Surprise is something we feel when something unexpected happens. The lower the probability we assume something to have, the more surprised we are when observing it. Rain in a desert is surprising, rain in the Netherlands is less surprising and rain on the moon is a miracle yielding almost unbounded surprise. When, following Tribus (1961), the surprise of observing outcome x is defined as $S_x = \log \frac{1}{P(x)}$, surprise can be measured in bits, like information and uncertainty. Observing something that was a certain fact yields no surprise, heads on a fair coin yield one bit of surprise and observing a 1/1000 year flood in some year yields a surprise of $\log 1000 \approx 10$ bits.

Entropy can now be defined as the expected surprise about the true outcome: $H(X) = E_X(S_x) = \sum_{x \in X}\left(P(x)(\log \frac{1}{P(x)})\right)$. Mutual information between X and Y is the expected reduction of surprise about the outcome of X due to knowing the outcome of Y: $I(X;Y) = E_Y\left\{E_X\left\{S_x - S_{x|y}\right\}\right\}$ or, vice versa, $I(X;Y) = E_X\left\{E_Y\left\{S_y - S_{y|x}\right\}\right\}$. The Kullback-Leibler divergence from the bookmaker's estimate to the gambler's estimate $D_{KL}(P_b||P_g)$ is the extra surprise the bookmaker expects the gambler to experience about the true outcome compared to his own surprise, while the reverse divergence $D_{KL}(P_g||P_b)$ is the expected extra surprise the bookmaker will experience, seen from the viewpoint of the gambler (Eq. 3.10).

$$\begin{aligned} D_{KL}(P_g||P_b) &= E_{P_g}\left\{S_{P_b} - S_{P_g}\right\} \\ &= \sum_{i=1}^{n} P_g(x_i)\left\{-\log P_b(x_i) - (-\log P_g(x_i))\right\} \\ &= \sum_{i=1}^{n} P_g(x_i) \log \frac{P_g(x_i)}{P_b(x_i)} \end{aligned} \tag{3.10}$$

In general, uncertainty can now be interpreted as expected surprise about the true outcome. The fact that different expectations can be calculated according to different subjective probability distributions reflects the fact that uncertainty can be both something

objective and subjective. The uncertainty perceived by the person have a subjective probability distribution himself is the entropy of that distribution. Kullback-Leibler divergence can be seen as the additional uncertainty one person estimates the other person to have compared to his own. When the uncertainty of a person is estimated from the viewpoint of a hypothetical all-knowing observer, who knows the true outcome, an objective perfect estimate of the uncertainty (expected surprise about the truth) of that person is obtained. This estimate is $D_{KL}\left(P_{\text{teapot fairy}}||P_{\text{person}}\right)$ and because the estimate of the truth $P_{\text{t}}(X = x_i) = 1$ for the i that corresponds to the true outcome and zero everywhere else, the estimate reduces to $-\log P_g\left(X = x_{\text{true}}\right)$. In chapter 5 we will see that this divergence has an important role as a measure of forecast quality.

3.6.1 Surprise and meaning in a message

When the word *information*[3] is used in daily conversation, it normally comprises two elements; surprise and meaning, as noted by Applebaum (1996), which he demonstrated with the following example of three possible messages:

1. I will eat some food tomorrow.
2. The prime minister and the leader of the opposition will dance naked in the street tomorrow.
3. XQWQ YK VZXPU VVBGXWQ.

Of these messages, message 2 is considered to convey most *information*. Message 1 has meaning but only little surprise, 3 has surprise (when taking English language as a prior), but little or no meaning and only message 2 has both surprise and meaning.

In this thesis, this framework is used to define useful information, which can offer new insights, noting the explicit distinction between meaning and surprise as two components of *information* transfer in the context of decision making. In this context, we can further specify meaning as meaning to a specific user or receiver of *information*. Message 2 in the above example might have meaning to someone interested in politics or dance, but is of little relevance to a farmer deciding whether to irrigate or not. A good forecast in the eyes of this farmer will have both meaning (e.g. exceedence of some rain threshold that will change his decision and has influence on his yield) and surprise (something new to be learned about this meaningful uncertain event).

3.6.2 Can information be wrong?

Is wrong information also information? This question is related to the notion that uncertainty is subjective. Decisions can be improved if uncertainty relative to the truth is reduced. Reduction of uncertainty relative to an arbitrary other (possibly irrational) belief is not guaranteed to help decisions. A message can be true or wrong. It can be noted that surprise, as defined by Applebaum (1996), is a necessary but not a sufficient condition to

3. In this section, the word information is printed in italic font where it refers to information in the ordinary language sense, to distinguish it from the information-theoretical definition.

convey information about the truth. An extension to the framework of Applebaum can therefore be proposed, in which we require the message to be true to truly constitute *information*. If the dancing in message 2 does not take place the next day, the message conveyed the same information (surprise), but wrong information. Since forecasting for decision making is concerned with information about the truth, the information in a forecast should reduce the expected surprise upon hearing the true outcome. Because one can not be surprised about the same information twice, and surprise can not be reduced without information, any reduction of surprise about the truth will involve a surprising message and any message that is true and surprising will reduce the user's uncertainty about the truth. In chapter 5, a mathematical decomposition of Kullback-Leibler divergence is presented that defines the concepts of missing information, true information and wrong information in the context of forecast verification.

3.7 Laws and Applications

Once information is defined as a quantity, some inequalities can be formulated that can be applied to many problems. Once these "laws of information" are allowed to seep into our intuition, they are as useful as concepts such as the conservation laws in thermodynamics and equations of motion in classical mechanics. Just like an apparent perpetuum mobile tells us we are missing some source of energy, results from information theory can for example serve to detect flaws in data manipulation methodologies.

3.7.1 Information cannot be produced from nothing

The data processing inequality states that information can never increase in a Markov chain (Cover and Thomas, 2006). This means that no matter how clever a data manipulation method is, it is not possible to extract more information than there is in the original raw data. Therefore there exists an upper limit to predictability, given the information that is in the predictors; see Westra and Sharma (2010) for an empirical approach to find this limit. An example are the low-frequency components (long term memory) in the climate system, that can be used for seasonal forecasts. The sea-surface temperatures (often summarized in indexes like El Niño Southern Oscillation) at a certain moment in time can contain information about the average weather conditions up to more than a year ahead (see e.g. Namias (1969); Piechota et al. (1998); Barnston et al. (1994)), but the predictability is limited by the mutual information between these temperature patterns and the predictands. Statistical methods can help to optimally extract this information, but can never increase it. The only way to improve these forecasts is to add informative predictors on long timescales, like ice coverage and vegetation. The deep rooting vegetation can for example serve as an indicator for soil moisture in the entire root-zone, which can show considerable long term persistence. New sources of information from remote sensing can provide valuable information for long term predictions.

3.7.2 Information never hurts

On average, information always reduces uncertainty (Cover and Thomas, 2006). As was shown in subsection 3.6.2, this is only true for correct information. This "information inequality" is often also stated as conditioning on average reduces entropy. In this form, the inequality is always valid, because conditioning information is true by definition. Only if Bayes' rule is bypassed if the information consists of a probability estimate that is directly assigned to the outcome, information may be partly wrong. In chapter 5, a framework is presented for distinguishing correct and wrong information. Unfortunately, this can only be done in hindsight. The wrong information is referred to as an unreliable probability estimate.

3.7.3 Given the information, uncertainty should be maximized

The principle of maximum entropy (Jaynes, 1957) states that given certain pieces of information, one should use that information as much as possible, but not use information that is not there. In other words, apart from the information that is there, uncertainty should be maximized. This will lead to probability estimates that reflect both what is known and what is unknown. The principle of maximum entropy (PME, also POME) also defines the concept of maximum entropy distributions, which maximize entropy given certain constraints on for example moments and domain. The constraints constitute the information, and the remaining maximum entropy constitutes the uncertainty that is left. Many frequently used parametric distributions turn out to be maximum entropy distributions for natural constraints. For example, the normal distribution maximizes entropy, given mean and variance. Section 3.9 gives some references for use of PME in hydrology and corresponding maximum entropy distributions. In chapter 4, the principle of maximum entropy is used to add information from low frequency components in the climate system to existing ensembles reflecting climatological uncertainty[4].

3.7.4 Applications of information theory

Although originally developed as a mathematical theory of communication, information theory has found application in a wide range of fields and problems. Initial results dealt with how much information could be sent over a communication line. When extending these results to a noisy communication channel, error correcting codes became important. Without these techniques, every scratch in an audio CD would be audible. The same codes are also important in data recovery. If information is stored in a redundant way, also partly damaged data allows full recovery of the information. This is closely linked to cryptography, where the aim is to make information unrecoverable without the correct key. Data compression is based on removing all redundancy in data, resulting in a new data set where every piece of data is a piece of unique information. If such a piece of data

4. "Climatological uncertainty" here refers to knowledge of historic frequencies, without other conditioning information.

is lost, it can not be recovered by any error correcting code. Therefore, there is a tradeoff between communication speed (compression) and robustness (error correction).

Kelly (1956), presented a new interpretation of information rate, which illustrates the connection between information and gambling. In the end, this connection follows from the strong connection between information and probability on the one hand, and probability and rational decisions on the other. The results for gambling, which were initially for idealized horse races, have been extended to more complicated cases, resulting in optimal portfolio theory for the stock market.

3.8 Relation to thermodynamics

In physics, entropy was used long before Shannon introduced it as a measure of uncertainty in his theory of communication. The term was first coined by Clausius in 1865 for a concept of "transformation content" or the amount of energy that was used or dissipated, referring to Greek "εντροπία" (entropia; from the verb "εντρέπεσθαι" / entrepesthai = to turn into). Clausius' theory of thermodynamics was stated in terms of the macroscopical quantities heat and temperature. He defined the change in entropy associated with the irreversible transformation of a quantity of heat Q from a body of temperature T_1 to a body of temperature T_2 as

$$\Delta S = Q \left(\frac{1}{T_2} - \frac{1}{T_1} \right) \tag{3.11}$$

It was by then understood that heat flowing from warm to cold in a isolated system is an irreversible process. And entropy changes can only be positive.

Boltzmann managed to find an explanation in terms of statistical properties of microscopical behavior, and thereby founded the field of statistical mechanics. The formulation was later refined by Gibbs. In the statistical mechanical perspective, temperature, heat and entropy are related to motion of microscopical particles. The classical theory of thermodynamics emerges from statistical mechanics through statistical relationships between averaged macroscopical variables. These relationships, although emergent from random and complex behavior, are quite accurate, but still only an approximation, given the lack of knowledge of the precise microscopical behavior. This is one of the first times that statistics and incomplete information entered fundamental physics so explicitly. A nice historical account of these developments is given in Lindley (2008).

Shannon, when developing a measure for uncertainty in his theory of communication, found that his measure has the same mathematical form as the statistical mechanics formulation of entropy. Jaynes (1957) investigated the relation between information-theory and statistical mechanics and proposed that the idea of maximum entropy in physics could be seen as a general principle of statistical inference. In this view, maximum entropy is seen not so much as a specific physical law, but more as the best estimate of the distribution by inference from the incomplete information in the macroscopical variables. The entropy of the dice in section 3.3 is reflecting remaining uncertainty for the gambler, given the side

information. In statistical thermodynamics, entropy represents the remaining uncertainty about the microstates (what the molecules are up to), given some conditioning information about the macrostates (pressure, temperature, volume).

For example, the macrostate temperature for an ideal gas corresponds to the average kinetic energy of the molecules. Secondly, we know that these energies must be positive. Given these two constraints, or pieces of conditioning information, the best guess is that the kinetic energies of the individual molecules follow an exponential distribution, which is the maximum entropy distribution under these constraints. The distribution for the velocities then becomes the Maxwell-Boltzmann distribution.

$$f(v) = \sqrt{\frac{2}{\pi}\left(\frac{m}{kT}\right)^3} v^2 \exp\left(\frac{-mv^2}{2kT}\right) \tag{3.12}$$

where $f(v)$ is the probability distribution of the instantaneous velocity of a randomly selected particle, T the temperature of the gas, m the molecular mass of the gas and k is Boltzmann's constant.

The second law of thermodynamics, which states that entropy tends to increase over time, can also be interpreted in terms of information. In that interpretation, it states that we tend to lose track of the microscopical behavior, because energy that is concentrated in ways that can accurately be described in terms of macroscopical constraints tends to spread over microscopical degrees of freedom. When only having access to macroscopical states, this represents a loss of information, which is analogous to a loss of free energy. Since the work of Szilard (1964), who coined Maxwells demon, and Landauer (1961), who connected irreversibility of logical operations to the irreversibility of heat generation, information is increasingly seen as a physical quantity, which is tightly linked with the laws of thermodynamics. Recently, the principle of Maxwell's demon, which can apparently violate the laws thermodynamics by using knowledge, was demonstrated experimentally for the first time. Toyabe et al. (2010) showed that information can be converted into free energy by applying feedback control on the molecular level. However, obtaining and processing that information is impossible without increasing entropy.

The precise connections between thermodynamics and information theory are quite subtle and need a more precise mathematical formulation to be fully appreciated. This is beyond the scope of this thesis, but would be an interesting direction of future research. Especially for hydrology, where inference about complex systems plays a major role, the cross-pollination between information theory and thermodynamics could yield interesting new perspectives.[5]

5. Note that also in fundamental physics, information theory plays a central role through "black hole information paradox" (information seems to be lost in a black hole, while quantum mechanics says that information should be preserved), which led to black hole thermodynamics, the holographic principle and even to new theories for gravity as an emergent entropic force; see Verlinde (2010)

3.9 Applications of information theory in water resources research

Information theory has been applied to several problems in hydrology and water resources research. It is outside the scope of this thesis to give an extensive overview, but a few references are given in this section. For more references, the reader is referred to the review papers by Harmancioglu et al. (1992a,b); Singh and Rajagopal (1987); Singh (1997). One area where information theory has been applied is the derivation, justification and parameter estimation of several statistical distributions that are common in hydrology by the principle of maximum entropy (PME, also POME); see for example (Sonuga, 1972; Singh and Guo, 1997, 1995b,a; Singh et al., 1993; Singh and Singh, 1991, 1985). Also in hydraulics, this principle has been used to derive velocity profiles (Chiu, 1988), and behavior of complex natural systems in terms of their distribution en temporal dependence structure (Koutsoyiannis, 2005a,b, 2011). One of the first applications of information theory in water resources is due to Amorocho and Espildora (1973), who investigated the use of entropy, conditional entropy and transinformation as measures of model performance. Later, information theory has also been applied to investigate predictability and forecast quality, mainly for meteorological applications. For relevant references, see chapter 5, where an information-theoretical framework for forecast verification is presented.

Other applications of information theory concerned the morphological analysis of river basin networks and landscapes (Fiorentino et al., 1993; Rodriguez-Iturbe and Rinaldo, 2001) and the design and analysis of monitoring networks for rainfall and water quality (Krstanovic and Singh, 1992a,b; Alfonso et al., 2010; Alfonso Segura, 2010). Because using information theory was not a predefined objective, this thesis builds mostly on literature about information theory itself, rather than previous applications to water research. The results mainly followed from the idea that information plays a central role in risk based water system operation and that a a rigorous theory of information existed. In this thesis, information theory is used

- to characterize the information that is contained in forecasts (chapter 5),
- to investigate the amount of information in data that is available for inference of models (chapter 6),
- to clarify the distinction between information and value for decisions (chapter 5),
- to give guidelines for the objective function for training models (chapters 5 and 6),
- to point out the philosophical objections to deterministic forecasts (chapter 5),
- to relate future information availability to the value of water (chapter 7)
- and, in the next chapter (4), to ensure that the correct amount of information in a forecast is reflected in a weighted ensemble.

Chapter 4

Adding seasonal forecast information by weighting ensemble forecasts

"Probability is relative, in part to [our] ignorance, in part to our knowledge."
- Pierre-Simon Laplace, 1820

Abstract - This chapter[1] presents an information-theoretical method for weighting ensemble forecasts with new information. Weighted ensemble forecasts can be used to adjust the distribution that an existing ensemble of time series represents, without modifying the values in the ensemble itself. The weighting can for example add new seasonal forecast information in an existing ensemble of historically measured time series that represents climatic uncertainty. A recent article compared several methods to determine the weights for the ensemble members and introduced the pdf-ratio method. In this article, an information-theoretical view on these weighting methods is presented. A new method, the minimum relative entropy update (MRE-update), is presented. Based on the principle of minimum discrimination information, the method ensures that no more information is added to the ensemble than is present in the forecast. This is achieved by minimizing relative entropy, with the forecast information imposed as constraints. The MRE-update is compared with the existing methods and the parallels with the pdf-ratio method are analyzed. The method is illustrated with an example application the a data-set from the Columbia river basin in the USA.

4.1 Introduction

This chapter presents an information-theoretical view on methods to produce weighted ensemble forecasts, as recently addressed by Stedinger and Kim (2010). It is argued that the updating of ensemble weights constitutes information and is therefore amenable to the information-theoretical principle of maximum entropy. Using the information-theoretical concepts, it is shown that the existing parametric update of Croley (2003) is a second order

1. Based on:
   - Weijs, S.V., van de Giesen, N., An information-theoretical perspective on weighted ensemble forecasts, to be submitted for publication.

approximation of the information-theoretical approach. It is also shown that, for normal distributions, the pdf-ratio method of Stedinger and Kim (2010) has a solution of the same shape but with some deviations in parameters. When the pdf-ratio method is forced to exactly match the prescribed conditional mean and variance, the results are identical. Firstly, this is an information-theoretical justification for this version of the pdf-ratio method. Secondly, it indicates the pdf-ratio method as a fast way to solve the MRE-update in case the forecast information consists of a conditional mean and variance. We only give a short introduction to weighted ensemble forecasts here. For more background and references, the reader is referred to Stedinger and Kim (2010).

4.1.1 Use of ensembles in water resources

Decision making about water resources systems often requires uncertainty to be taken into account. For the operation of a system of reservoirs, for example, forecasts at different timescales can improve decisions. Because these decision problems are not certainty equivalent (Philbrick and Kitanidis, 1999), optimal decisions can only be found by explicitly taking uncertainties into account.

Ensembles are a common method to describe uncertainty in forecasts, such as future inflows to a reservoir system. An ensemble consists of several scenarios (also called members or traces), which represent the possible future development of the variables of interest. An ensemble of past measured streamflows can for example be used as a stochastic description of the inputs to a system of reservoirs (Kelman et al., 1990). Ideally, such a historical ensemble represents the climatic uncertainty about the interannual variability, and at the same time contains a realistic stochastic description of spatial and temporal variability and correlations at smaller timescales. Using ensembles directly has the advantage that no statistical models have to be assumed. Ensemble members can be multivariate, e.g. yearly sets of daily time series of various hydrological variables on various locations (Faber and Stedinger, 2001). Other examples of ensembles are a multi-model ensemble of different climate models predicting the global average temperature over the next century, the ensemble weather forecasts of European Centre for Medium range Weather Forecasts (ECMWF), and the ensemble streamflow predictions (ESP) that are used throughout the USA as an input for reservoir optimization models or decision making about flood protection.

These ESP forecasts reduce uncertainty by conditioning on actual basin conditions. The ESP forecast is produced by feeding a distributed hydrological model, which has an initial state consistent with actual basin conditions, with past observed weather patterns (Day, 1985; Wood and Lettenmaier, 2008). The result is an ensemble with one trace for each historical weather pattern. The ensemble reflects the climatic uncertainty, but also the information that is in the actual basin conditions. The most important information is in the states that represent the largest seasonal storage changes, such as soil moisture and snow pack. This information reduces the climatic uncertainty in the flows; see the bottom left of Fig. 4.1.

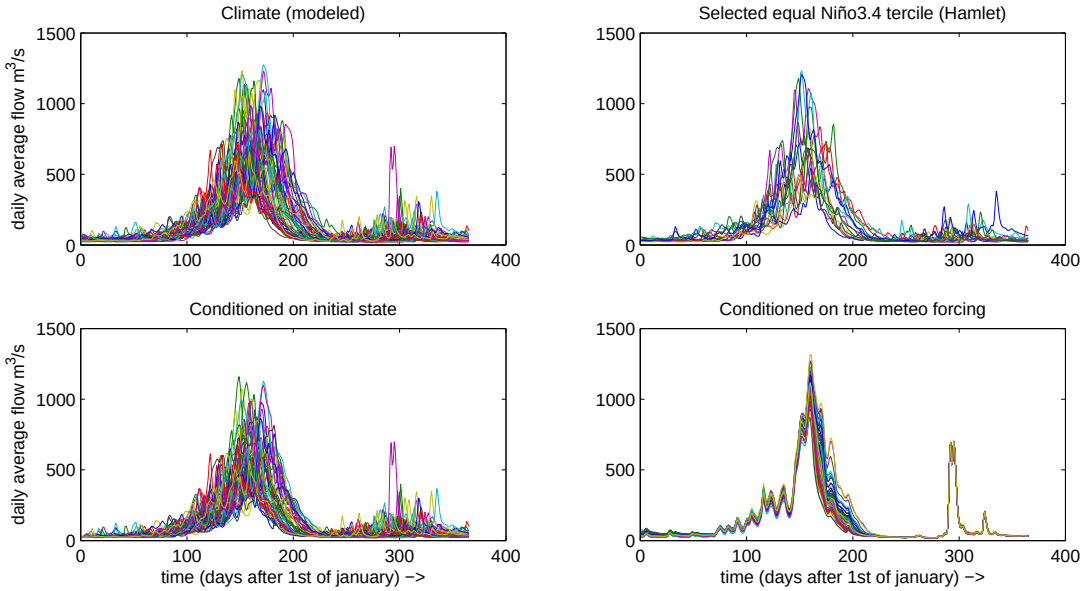


Figure 4.1: The ESP forecasts are generated by forcing a hydrological model with historical weather patterns. The top left figure represents the modeled climatic uncertainty, in which the modeled flow traces obtained by forcing the model with the meteorological data while the initial conditions nor the meteorological data are conditioned on the actual year. The bottom left figure is a typical ESP forecast, in which all traces start from the initial conditions in one particular year (2003), but each trace corresponds to a different historical weather pattern. The top right shows the climatic ensemble, where only traces matching the actual tercile of the ENSO index are selected (also 2003, above normal ENSO). The lower right figure shows traces produced by forcing a model from different initial conditions all with the same weather pattern from 2003. Remarkably, the effect of the uncertainty in the initial conditions only becomes important after 5 months, in the melting season, when the different initial snowpack conditions are translated into different streamflows.

Apart from the initial basin conditions, information about the streamflows might be present in climatic indexes that characterize long term persistence in the atmospheric and oceanic circulation (Piechota et al., 1998). The indexes are usually based on sea surface temperatures and pressure heights. For example, the phase of the El Niño Southern Oscillation (ENSO) and the Pacific Decadal Oscillation (PDO) gives information about the precipitation in the Pacific Northwest of the USA (Hamlet and Lettenmaier, 1999). Hamlet et al. (2002) proposed a method to select only the one third of ESP traces that match the ENSO conditions of the actual year. They calculated that this information, in combination with a more flexible reservoir operation strategy, could lead to "an increase of nonfirm energy production from the major Columbia River hydropower dams, ... resulting in an average increase in annual revenue of approximately $153 million per year in comparison with the status quo." This chapter presents a method to include this type of information into ensemble forecasts by weighting rather than selecting ensemble traces.

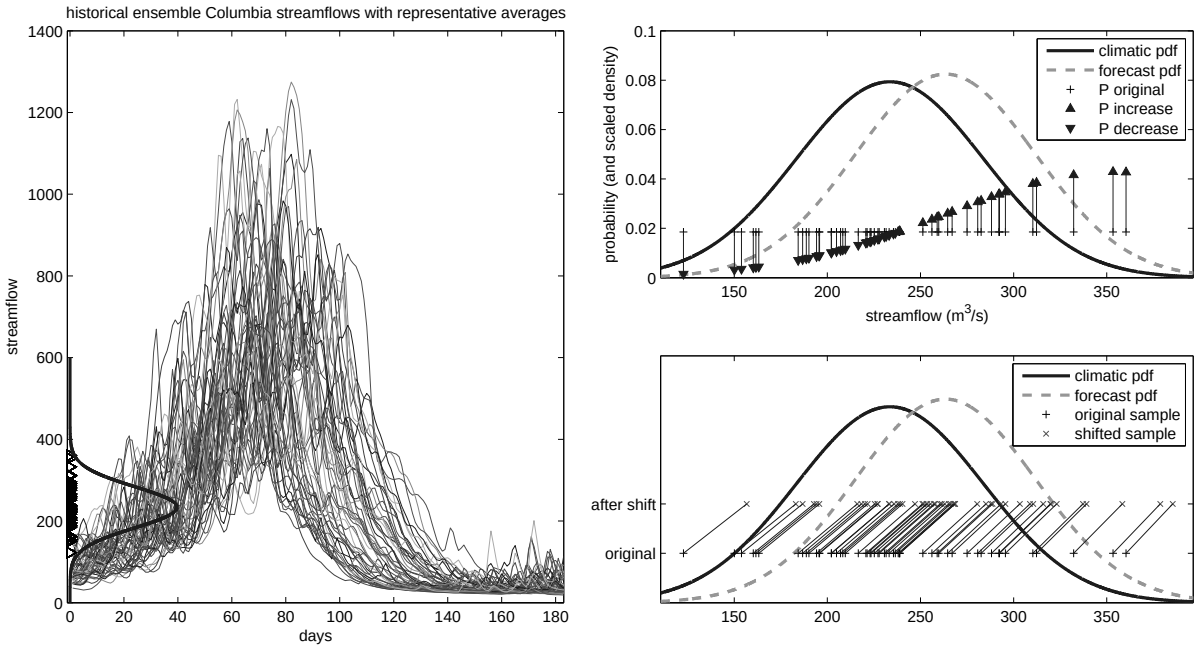


Figure 4.2: Equally weighted ensemble members can represent a nonuniform density. This density can be changed by shifting or by weighting the ensemble traces.

4.1.2 Weighted ensembles

Ensemble traces are often stated to be "equally likely" (see e.g. Cloke and Pappenberger (2009)). This should not be taken too literally. For example, streamflows close to the interannual mean are more likely than streamflows of the most extreme ensemble members. Ideally ensemble forecasts are produced in such a way that all ensemble members can be considered to represent equally probable ranges of the possible outcomes. This is reflected by the fact that scenarios usually lie closer to each other around the mean value. Each scenario represents the same probability, but a different region of the space of the outcome and therefore a different probability density; see Fig. 4.2. In that way, the ensemble is a discrete representation of the underlying probability density function and can be used in risk analysis and decision making (see e.g. (Georgakakos and Krzysztofowicz, 2001) and the MMPC approach in van Overloop et al. (2008); see chapter 2.

Often, long-term forecasts based on for example ENSO do not contain information at a high spatial and temporal detail level, but rather contain information about averages in time and space, e.g. the total flow at the outlet of a river basin, averaged over several months. Yet, risk analysis may depend on events and sequences at shorter timescales and smaller spatial scales. One could attempt to shift the time series in the ensemble (bottom right Fig 4.2), but this could destroy the realism of the traces if the shifting or scaling procedure is not sophisticated enough.

A reasonable alternative to combine detailed information in the ensemble of historical time series with forecast information is to update the probabilities of individual ensemble members, while leaving the time series they contain intact (Stedinger and Kim, 2010). This has the advantage of preserving high-resolution stochastic structure within the ensembles.

The update of the weights is thus based on averages in space and time, derived from the time series, compared to information on these same quantities in the seasonal forecast; see top right of Fig. 4.2.

Such an update of probability weights is consistent with Laplace's principle of insufficient reason, which states that all outcomes of an experiment should be considered equally likely, unless one has information that indicates otherwise. In this case, the information from e.g. ENSO indicates otherwise. As is shown later, Laplace's principle also corresponds to the more general information-theoretical principle of maximum entropy. When updating ensemble probabilities to deviate from equal probabilities, additional information is added to the ensemble. That information is measured by the relative entropy between the original and the new probabilities. To prevent adding more information to the ensemble than is justified by the forecast, the relative entropy is minimized, constrained by the information in the forecast. In this chapter, this method is referred to as the minimum relative entropy update (MRE-update).

Apart from the example in this chapter, concerning climatic ensembles and additional forecast information, the MRE-update method that is introduced is generally applicable whenever information is added to an ensemble by adjusting probabilities. Another possible application could be a bias correction or variance adjustment for ensembles generated by Monte Carlo simulations with models. In finance, the concept of minimum relative entropy has been used to include price information in a Weighted Monte Carlo simulation, which is mathematically equivalent to the proposed bias correction application of the MRE-update (Avellaneda et al., 2001).

4.1.3 Previous work on adding forecast information to climatic ensembles by weighting

The focus of this chapter is how to find a proper adjustment of probabilities in an existing ensemble to reflect forecast information that is given in the form of moments or conditional tercile probabilities, also referred to as probability triplets, which are often used to present seasonal forecasts. Finding such adjustments was recently discussed in Stedinger and Kim (2010). We now shortly review some existing weighting methods, to which the MRE-update in this chapter will be compared. After that, this chapter also addresses the question whether the chosen form of the forecast is an accurate representation of the information that the seasonal forecasts are supposed to convey.

Croley (1996) presents a method for updating ensemble member probabilities, assuming forecast information is given by a third party in the form of conditional tercile probabilities. These are the probabilities of below normal, normal or above normal conditions, which have equal probabilities of 1/3 in the climatic distribution. Croley presents a non-parametric probability adjustment procedure based on minimization of the sum of squared deviations of the probabilities from the uniform distribution. The result is a block adjustment of probabilities, in which all ensemble members within one tercile get assigned the same weight. This is in line with the literal interpretation of the probability triplets as

considered by Wilks (2000, 2002). The method can also deal with multiple forecasts, including deterministic "most probable event" forecasts (Croley, 1997). A procedure that one by one eliminates constraints according to user priorities helps to reach a feasible solution for the probability weights.

Croley (2003) presents an alternative parametric approach, in which sample moments of the forecast distribution can be imposed as equality constraints on corresponding moments of the weighted ensemble. A problem with this method is that often many of the probabilities become zero, so only part of the original ensemble is used. The cause for this partly lies in the objective function that Croley proposed. Although it seems reasonable to minimize the adjustment to the probabilities, there is no clear rationale for using minimum squared deviations as objective function; see page 345 of Jaynes (2003). Among the two methods, the parametric one leads to more reasonable, smoother adjustments than the block adjustment, avoiding sudden jumps in probability between adjacent ensemble members (Stedinger and Kim, 2010).

Stedinger and Kim (2002, 2007, 2010) introduce the pdf-ratio method, also focusing on obtaining a weighted ensemble, but now assuming a forecast is given by the third party as a target conditional distribution. They also argue that forecast information in the form of probability triplets should not be taken literally, but as a representation of a smooth underlying target distribution. They propose that probability triplets should be converted to a likely target distribution that can subsequently be used in the pdf-ratio method. The pdf-ratio method adjusts the probability of each ensemble member with the ratio between marginal (climatic) and conditional (forecast) probability density functions (pdf) at each sample point. The pdf-ratio method then normalizes the probabilities to make them sum to one. Although the method does not seem to be very sensitive to distribution type, one still has the problem of assuming a distribution from only two tercile probabilities or moments. Another problem is that for relatively large deviations from the climatic distribution, significant deviations of the resulting moments from the target moments occur.

In this chapter we analyze the problem of updating ensemble probabilities with forecast information from an information-theoretical viewpoint. We present a new method to include forecast information in a historical ensemble, based on minimum relative entropy. In a comparison between our method and the existing methods, we explicitly show the assumptions in the existing methods and the differences between various ways of presenting forecast information. Before introducing the MRE-update, a short review of relevant information-theoretic concepts and principles is given.

4.2 Information, Assumptions and Entropy

A reduction of entropy implies that information is added about the uncertain event the distribution describes. Information can be added in the form of data or knowledge, but can also enter implicitly by unwarranted assumptions, all reducing the entropy of the

distribution. When new information in, for example, a forecast motivates a revision of the probability distribution from $P(X)$ to $Q(X)$, the relative entropy $D_{KL}(Q||P)$ is an exact measure of the amount of information in that specific forecast. The expectation of $D_{KL}(Q||P)$ over all possible forecasts is equal to the mutual information between the forecasts and the random variable X; see chapter 3. A good overview of information-theoretic concepts reviewed in this section can be found in Cover and Thomas (2006).

4.2.1 The principle of maximum entropy

Among all discrete probability distributions, the uniform distribution, in which all outcomes are equally likely, maximizes entropy. The uniform distribution has maximum missing information amongst all distributions on a finite support set. So without any information available except for the support set, it is rational to assume a uniform distribution. Assuming any other distribution leads to less uncertainty without having the information to justify that reduction. This idea was already formulated by Laplace

Jaynes (1957) first formulated the principle of maximum entropy, which is in fact a generalization of Laplace's principle. It states that when making inferences based on incomplete information, one should choose the probability distribution that maximizes uncertainty (entropy), subject to the constraints provided by the available information. Applying this principle leads to a distribution with maximum uncertainty, but bounded by what is known. This automatically implies that no false certainty is created and only truly existing information is added. The principle of maximum entropy (PME or POME) has been widely applied for derivation of prior distributions and parameter estimation; see e.g. Singh and Singh (1985); Singh and Rajagopal (1987); Singh (1997) for relevant references.

Along the same lines of reasoning, the principle of minimum relative entropy or principle of minimum discrimination information (Kullback, 1997) states that given new facts, a new distribution should be chosen that is consistent with those facts, but apart from that minimizes the information gain with respect to the original distribution. This principle ensures that not more new information is included than is justified by the new facts. The principle leads to results identical to those of PME, but generalizes to non-uniform prior distributions.

4.3 The Minimum Relative Entropy Update

4.3.1 Rationale of the method

We propose to apply the principle of minimum relative entropy to adjust the probabilities of a climatic ensemble to reflect new forecast information. This method is referred to as the minimum relative entropy update (MRE-update). The MRE-update is a constrained minimization of relative entropy to optimally combine new information in the form of constraints with an existing ensemble. In this example, the method is used for updating a climatic ensemble, whose members may contain high resolution spatial and temporal

patterns of several variables. The new information that is added concerns some averaged quantities that characterize the traces in the climatic ensemble. This new information is for example expressed in the form of conditional moments of those averaged quantities. The information is added by adjusting the weights of the ensemble members in such a way that the weighted moments match the forecast. The amount of new information added by the forecast is the relative entropy between the original uniform distribution of probabilities and the updated probabilities assigned to the ensemble. Minimizing this relative entropy, constrained by the information contained in the forecast, will exactly use all information in the forecast, without adding information that is not in the forecast. Consequently, the new ensemble is consistent with the forecast, but does not deviate more than necessary from the observed climatic distribution. The minimum relative entropy update thus optimally combines forecast information with climatic information in an ensemble.

4.3.2 Formulation of the method

In the minimum relative entropy update (MRE-update), we try to find updated probabilities q_i by minimizing relative entropy between the original uniform distribution of probabilities p_i and updated probabilities q_i assigned to the n samples x_i, given the general constraints of probabilities and the constraints posed by the forecast information. This results in a nonlinear optimization problem with objective function:

$$\min_{q_i \ldots q_n} \{\sum_{i=1}^{n} q_i \log(\frac{q_i}{p_i})\} \tag{4.1}$$

Because in this case we start from an uniform distribution of equiprobable ensemble members (p_i is constant), which has maximum entropy, minimizing relative entropy is equivalent to maximizing the entropy of the distribution of q_i:

$$\max_{q_i \ldots q_n} \{-\sum_{i=1}^{n} q_i \log(q_i)\} \tag{4.2}$$

Subject to the constraint that the probabilities sum to one

$$\sum_{i=1}^{n} q_i = 1 \tag{4.3}$$

And that all probabilities are nonnegative:

$$q_i \geq 0 \tag{4.4}$$

This last constraint is never binding in the MRE-update, because the objective function (Eq. 4.1) already ensures positive weights. Without any forecast information, no extra constraints are added. Objective function (4.1) minimizes the divergence from the original

uniform distribution. Because constraints (4.3) and (4.4) are already satisfied by the original distribution, no adjustment is made.

When forecast information is available, it can be introduced by additional constraints to the minimization problem. In case the forecast information is given as probability triplets of below (p_b) and above (p_a) normal conditions, the following constraints are added:

$$\sum_{i \in S_b} q_i = p_b \tag{4.5}$$

$$\sum_{i \in S_a} q_i = p_a \tag{4.6}$$

In which S_b and S_a are the sets of i for which $x_i \leq x_b$ and $x_i \geq x_a$ respectively. With x_b and x_a being the lower and upper terciles of the climatic distribution for X.

In case the forecast information is given as the conditional mean μ_1 and standard deviation σ_1, the following constraints are imposed:

$$\sum_{i=1}^{n} q_i x_i = \mu_1 \tag{4.7}$$

$$\sum_{i=1}^{n} q_i (x_i - \mu_1)^2 = \sigma_1^2 \tag{4.8}$$

The resulting constrained convex optimization problem is subsequently solved using a standard gradient search.

4.4 Theoretical test case on a smooth sample and comparison to existing methods

In this section, the results of the minimum relative entropy update are compared with the results of the Croley nonparametric adjustment (Croley, 1996), the Croley parametric adjustment (Croley, 2003) and the pdf-ratio method Stedinger and Kim (2010). The same example as the univariate case in Stedinger and Kim (2010) was used. In this example an artificially generated smooth climatic sample of $n = 50$ scalar values x_i is updated with forecast information of the previously mentioned forms. In a real-world application, the sample would represent an ensemble of, for example, time series. The sample is created by evaluating the inverse cumulative distribution function of the prescribed original distribution at the Hazen plotting positions $((i - 0.5)/n)$; see Stedinger and Kim (2010). The sample is drawn from a normal distribution with mean 3 and standard deviation 1. In absence of extra information, all 50 samples are considered equiprobable, with probability 1/50.

The challenge is now to update the probabilities of the sample values in such a way that the ensemble reflects the forecast distribution, given the climate information, conditioned on forecast information. Comparing the methods, all using their own required type of input information, we automatically compare forecast information given as three different types:

(TC) The conditional tercile probabilities p_b, p_n and p_a
(N) An assumed forecast normal distribution with given parameters.
(M) The mean μ_1 and standard deviation σ_1 of the forecast distribution

In table 4.1 an overview of the methods and forecast types is given, indicating which combinations are compared and which abbreviations are used for the results. Three of the methods have also been compared by Stedinger and Kim (2010). For the pdf-ratio method, they considered normal, lognormal and gamma type distributions. This section focuses on their results using the assumption of a normal distribution for both climatic and forecast distribution.

Adjustment method	Forecast used		
	Tercile constraints (TC) p_b, p_n, p_a	Conditional distribution (N) $N(\mu_1, \sigma_1)$	Conditional mean and variance (M) μ_1, σ_1^2
pdf-ratio method		(pdf-N)	
Croley non-parametric adjustment	(TC)		
Croley parametric adjustment			(CP-M)
Minimum relative entropy update	(TC)		(MRE-M)

Table 4.1: An overview of the methods and types of forecasts that are compared in this chapter.

In Stedinger and Kim (2010), the forecast information of type TC is converted to type N, using the assumption of a normal distribution. The rationale behind this is that the forecast of type TC is likely to represent a smooth underlying distribution. Results of the pdf-ratio method using forecast N are then compared with the Croley nonparametric adjustment using forecast TC and to CP-M.

We compare results of the MRE-update with both tercile probability constraints (TC) and constraints on mean and standard deviation (MRE-M) to the results of the Croley nonparametric adjustment, the Croley parametric adjustment, and the pdf-ratio method. For the Croley methods, we use the same constraints as for the MRE-update for both the forecast TC (results: TC) and the forecast M (results: CP-M). For CP-M, the MRE-objective (Eq. 4.1) is replaced by minimum squared adjustment (equation 4.9; see Croley (2003)).

$$\min_{q_i}\{\sum_{i=1}^{n}(q_i - p_i)^2\} \tag{4.9}$$

4.4.1 Results in a theoretical test case

From the results it becomes clear that there is a large difference between forecasts in the form of probability triplets and forecasts in the form of moments. When the deviations from the original moments are small, the results for the methods using moments and the pdf-ratio method are similar. When deviations become larger, the Croley parametric method shows clearly different behavior, while the MRE-update and the pdf-ratio method show very similar results in many cases. The results are presented as graphs showing the weights of the individual traces as a function of the value they represent and the cumulative weights, which form an empirical cumulative distribution function (CDF). Next to the graphs, a number of tables shows the resulting tercile probabilities, moments and relative entropies. The relative entropy of each case can be interpreted as the reduction in uncertainty or the information added by the forecast.

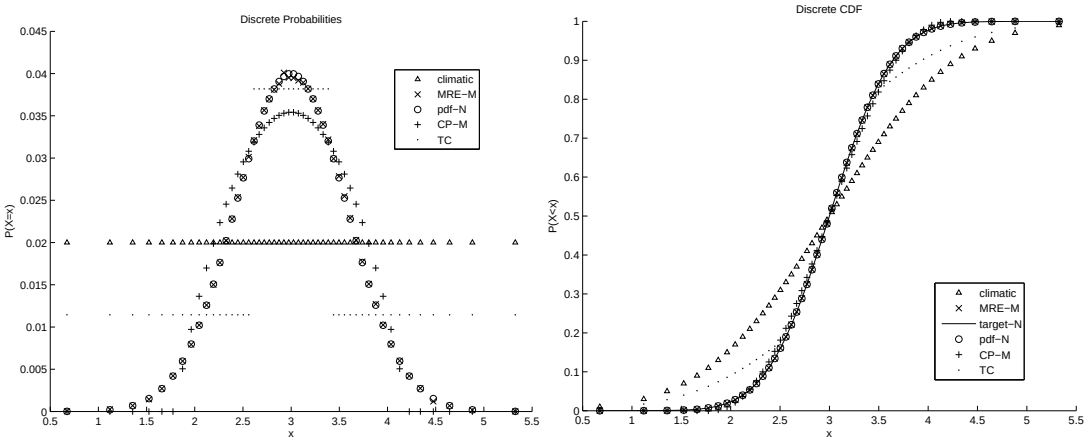


Figure 4.3: Resulting ensemble member probabilities for μ_1=3, σ_1=0.5 and resulting empirical cumulative distribution.

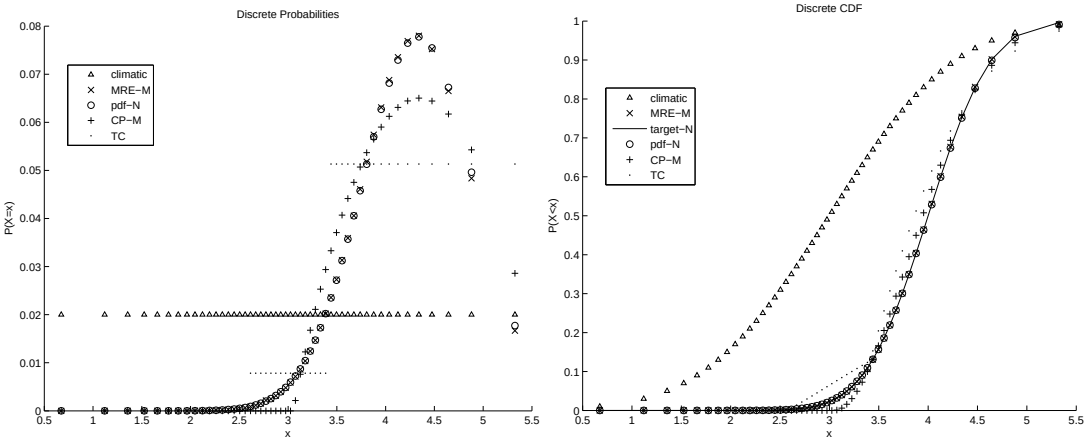


Figure 4.4: Resulting ensemble member probabilities for μ_1=4, σ_1=0.5 and resulting empirical cumulative distribution.

The codes in tables 4.2-4.4 and the legends of figures 4.3-4.6 correspond to the different methods given in table 4.1. For the case of tercile probability constraints (forecast TC), the MRE-update (with objective eq. 4.1) always results in exactly the same block

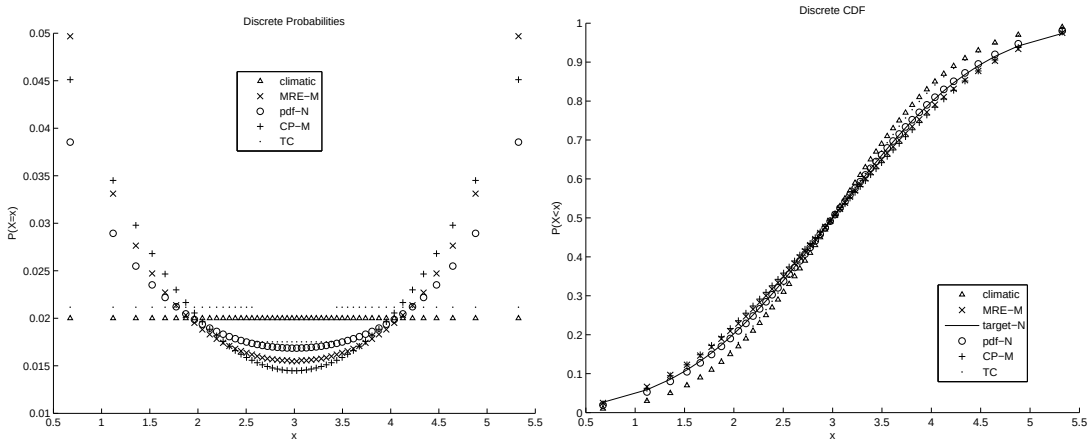


Figure 4.5: Resulting ensemble member probabilities for μ_1=3, σ_1=1.2 and resulting empirical cumulative distribution.

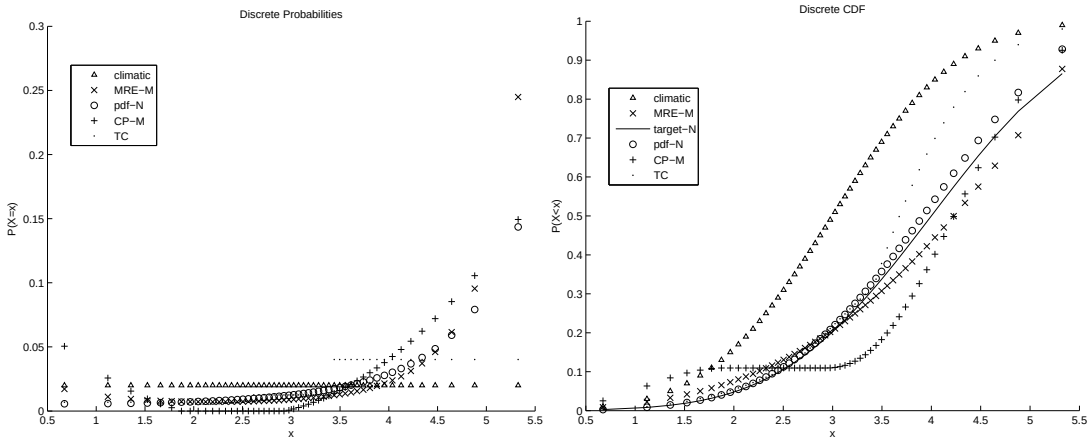


Figure 4.6: Resulting ensemble member probabilities for μ_1=4, σ_1=1.2 and resulting empirical cumulative distribution.

True moments		Target probabilities		Estimated probabilities							
				TC		pdf-N		MRE-M		CP-M	
μ_1	σ_1	p_b	p_a	p_b	p_a	p_b	p_a	p_b	p_a	p_b	p_a
3.00	0.25	0.043	0.043	0.043	0.043	0.049	0.049	0.050	0.050	0.047	0.047
2.00	0.50	0.873	0.002	0.873	0.002	0.880	0.002	0.881	0.002	0.886	0.000
3.00	0.50	0.195	0.195	0.195	0.195	0.205	0.205	0.205	0.205	0.227	0.227
4.00	0.50	0.002	0.873	0.002	0.873	0.002	0.880	0.002	0.881	0.000	0.885
4.50	0.50	0.000	0.984	0.000	0.984	0.000	0.986	n.a.	n.a.	n.a.	n.a.
5.00	0.50	0.000	0.999	0.000	0.999	0.000	0.999	n.a.	n.a.	n.a.	n.a.
3.00	1.00	0.333	0.333	0.333	0.333	0.340	0.340	0.342	0.342	0.342	0.342
3.00	1.20	0.360	0.360	0.360	0.360	0.364	0.364	0.375	0.374	0.382	0.382
4.00	1.20	0.117	0.682	0.117	0.682	0.127	0.669	0.142	0.712	0.110	0.840
4.50	1.20	0.054	0.814	0.054	0.814	0.064	0.791	0.097	0.843	0.081	0.919
5.00	1.20	0.021	0.905	0.021	0.905	0.029	0.877	0.071	0.931	0.071	0.931

Table 4.2: Resulting tercile probabilities for the four methods compared

Target assuming normal				Mean				Standard deviation			
p_b	p_a	μ_1	σ_1	TC	pdf-N	CP-M	MRE-M	TC	pdf-N	CP-M	MRE-M
0.043	0.043	3.00	0.25	3.00	3.00	3.00	3.00	0.41	0.25	0.25	0.25
0.873	0.002	2.00	0.50	2.07	2.00	2.00	2.00	0.61	0.50	0.50	0.50
0.195	0.195	3.00	0.50	3.00	3.00	3.00	3.00	0.76	0.50	0.50	0.50
0.002	0.873	4.00	0.50	3.93	4.00	4.00	4.00	0.61	0.50	0.50	0.50
0.000	0.984	4.50	0.50	4.05	4.52	n.a.	n.a.	0.52	0.50	n.a.	n.a.
0.000	0.999	5.00	0.50	4.07	4.96	n.a.	n.a.	0.51	0.41	n.a.	n.a.
0.333	0.333	3.00	1.00	3.00	3.00	3.00	3.00	0.98	0.99	1.00	1.00
0.360	0.360	3.00	1.20	3.00	3.00	3.00	3.00	1.01	1.13	1.20	1.20
0.117	0.682	4.00	1.20	3.61	3.84	4.00	4.00	0.88	1.04	1.20	1.20
0.054	0.814	4.50	1.20	3.81	4.19	4.50	4.50	0.75	0.95	1.20	1.20
0.021	0.905	5.00	1.20	3.95	4.47	5.00	5.00	0.64	0.84	1.20	1.20

Table 4.3: Resulting mean and standard deviation for the various methods

Target Assuming normal distribution				Resulting divergence (relative entropy) from original distribution (bits)			
μ_1	σ_1	p_b	p_a	TC	pdf-N	CP-M	MRE-M
3.00	0.25	0.043	0.043	1.132	1.324	1.379	1.324
2.00	0.50	0.873	0.002	1.002	1.174	1.251	1.181
3.00	0.50	0.195	0.195	0.257	0.459	0.500	0.459
4.00	0.50	0.002	0.873	1.002	1.174	1.251	1.178
4.50	0.50	0.000	0.984	1.438	2.081	n.a.	n.a.
5.00	0.50	0.000	0.999	1.547	3.356	n.a.	n.a.
3.00	1.00	0.333	0.333	0.001	0.000	0.000	0.000
3.00	1.20	0.360	0.360	0.005	0.036	0.081	0.078
4.00	1.20	0.117	0.682	0.371	0.561	1.333	0.949
4.50	1.20	0.054	0.814	0.712	1.105	2.870	2.283
5.00	1.20	0.021	0.905	1.035	1.709	5.274	5.274

Table 4.4: Resulting relative entropy for the different methods

adjustment as for the Croley nonparametric adjustment (objective eq. 4.9). For tercile constraints, the minima of the objective functions thus coincide. The identical results for these two methods are indicated with TC. Pdf-N indicates pdf-ratio method, using normal climatic and normal forecast distributions (forecast N). The original smooth sample with uniform weights, drawn from the climatic normal distribution is also shown (climatic). In the cumulative distribution plots, also the cumulative distribution function of the target distribution, used in the pdf-ratio method is plotted.

The comparison is made for a hypothetical case, with the various combinations of μ_1 and σ_1 as described in Stedinger and Kim (2010). For the case of tercile probability constraints (TC), target tercile probabilities were derived from μ_1 and σ_1, assuming a normal distribution. Figures 4.3-4.6 show the assigned probabilities for individual ensemble members against their x-values. The right graphs in these figures show the corresponding discrete approximations for the cumulative distribution functions (CDF), using Hazen plotting

positions, following Stedinger and Kim (2010). Table 4.2 shows the target and resulting tercile probabilities for below and above normal conditions (p_b and p_a) for the methods, while table 4.3 shows resulting means and standard deviations. Table 4.4 shows the resulting relative entropy for the set of ensemble member probabilities, relative to the original uniform distribution. The "n.a." entries correspond to combinations of μ_1 and σ_1 constraints for which the optimization based methods (CP-M) and (MRE-M) were not able to find a solution. This means that those μ_1 and σ_1 combinations are not achievable with the given sample.

Results of the methods compared

Small rounding errors are caused due to the limited number of ensemble members and the way the original sample is drawn. The effects become apparent for the case μ_1=μ_0=3 and σ_1=σ_0=1 (no new information). Firstly, the discrete approximation of outer tercile probabilities with 17 of the 50 members, results in probabilities of 0.34 rather then $\frac{1}{3}$. Secondly, the standard deviation of the original sample is not exactly one, but 0.987.

When ignoring small differences due to these numerical effects, table 4.2 shows that for all methods p_b and p_a match the assumed target reasonably well for cases with $\sigma_1 \leq 1$. For the cases with increased variance, the methods using moment constraints show somewhat larger deviation from target probabilities p_a and p_b.

Results for the mean-variance forecast (type M) show a difference between MRE-update (MRE-M) and Croley parametric adjustment (CP-M). The latter tends to result in more ensemble member probabilities set to zero; see Fig. 4.3-4.6. Although results are different, both satisfy the constraints given by the forecast. The difference in results is purely due to the difference in objective function. Naturally, because the moments are imposed as constraints, both methods will exactly match the target mean and standard deviation (table 4.3), like the results for methods using tercile constraints (TC) exactly match p_b and p_a (table 4.2).

Information contained in forecasts

The relative entropies of the resulting discrete distributions, as compared to the original uniform distribution, are shown in table 4.4. These relative entropies are a direct measure of the amount of information added to the ensemble. In all cases, the result for TC has the lowest relative entropy, respectively followed by pdf-N, MRE-M and CP-M. The relative entropy resulting from the MRE-update is the uncertainty reduction by the information in the forecast. Hence we can see that the forecast of type TC is less informative than type M (compare TC and MRE-M). The entropies for (MRE-M) in table 4.4 also show that larger shifts in mean result in larger relative entropy. This corresponds to the intuition that forecasts add more information when they deviate more from climatology.

Because the pdf-N and CP-M methods have no information-theoretical founding, the relative entropy resulting from those methods does not say much about the amount of

information in the forecast, but does indicate the uncertainty reduction in the ensemble. Because CP-M has higher relative entropy than MRE-M, we can say that this first method introduces information (reduction of uncertainty) that is not present in the forecast. This will be further explained in the discussion in section 4.4.2. For cases with reduced variance and not too large a shift in mean, results of pdf-N very closely resemble MRE-M. Apparently, the information contained in the mean and variance constraints is the same information contained in a normal distribution. This is related to the fact that the maximal entropy distribution for a given mean and variance is a normal distribution; see appendix A. In general, there is a duality between sufficient statistics and the constraints of a maximum entropy distribution; see Jaynes (2003), page 520.

If forecast information is given as constraints on conditional tercile probabilities (TC), there are infinitely many adjustments possible to satisfy those constraints. However, the adjustment that minimizes relative entropy is a block adjustment. Generally, when information about a distribution is given in form of constraints on quantile probabilities, the maximum entropy distribution is piecewise uniform. This also holds for the distribution of probabilities over the discrete scenarios. For the case of tercile constraints, the objective of minimum squared deviations of the Croley method has its minimum in the same location, leading to identical results as MRE.

It might seem strange that results for the MRE-update depend on how the forecast information is presented. Forecast TC and M give completely different results. However, considering the tercile probabilities (TC) and moments (M) as different ways to present the same information implies that some underlying information is present, i.e. that we know more about the information than what is presented. In this example forecast M contains the extra knowledge that the forecast distribution is normal. Taking the information-theoretical viewpoint, forecast TC and M contain different information. If we do not know anything but these forecasts, we have to take them literally and thus we get different results for both forecasts. Forecasts of type TC appear to be less informative than forecasts of type M; see table 4.4. Moreover, forecast TC does not seem an appropriate way to represent a smooth forecast distribution. Deriving a mean and standard deviation (forecast M) as given variables from tercile probabilities (forecast TC), as was done in Stedinger and Kim (2010), implicitly introduces the assumption that the forecast distribution is normal and hence results in a smooth update.

MRE compared to Croley

The MRE-update (MRE-M) and the Croley parametric method (CP-M) both use a forecast of type M. Although the same information is used by both methods, they lead to different results. Logically, because it is the objective, the MRE-update results in a smaller divergence (relative entropy) than CP-M (table 4.4). This means that the MRE-update retains more of the climatic uncertainty. However, the Croley parametric method uses the same constraints as the MRE-update, so the amount of forecast information used is the same. Table 4.3 shows that both results are equally consistent with the forecast, because mean and standard deviation are exactly reproduced in both cases.

Consequently, from an information-theoretical point of view, we can say that the Croley method makes an unnecessary extra deviation from climatology, not motivated by the forecast. The higher relative entropy means uncertainty is reduced by artificially introduced information that is not actually there in the forecast. The minimum squared deviation objective therefore results in an over-confident adjustment. This is also demonstrated by the fact that several probabilities are set to zero, without having information that explicitly rules out those scenarios as representing possible future conditions.

MRE compared to pdf-ratio, using equivalent forecasts M and N

Because the maximum entropy distribution for given mean and variance is a normal distribution, forecast M implies a normal forecast distribution. When a forecast used in the pdf-ratio method is a normal distribution (forecast type N) with mean and standard deviation (of forecast type M) as parameters, forecast N and M are equivalent and add the same information. Therefore, the differences in the resulting weights for (pdf-N) and (MRE-M) are purely due to the methods.

MRE-M gives similar results as the pdf-ratio method (pdf-N) for adjustments that are not too large and where the ensemble is sufficiently dense; see fig. 1. In cases where the adjustment is large and the sample values do not cover the entire forecast distribution range, the approximation of the forecast distribution needs probability mass in the range outside the sample values. In the pdf-ratio method, this results in some "missing" probability. This can clearly be seen in the figure 4.6 (right), where the value of the target CDF at the highest sample value is still far from one. The pdf-ratio method (pdf-N) needs a large normalization step in these cases. All ensemble member probabilities are multiplied by the same factor, to make them sum to one. This results in deviations from the target mean and variance, because the missing probability outside the sample range is divided equally over the traces. Although this leads to smooth adjustments, it also results in weighted ensembles that do not conserve the mean and variance of the new information, possibly biasing results of planning and risk analysis.

The (MRE-M) method distributes the missing probability to the sample values in a way to match exactly the target mean and variance, as long as the constraints do not become over-restrictive. Especially when a high adjustment in the mean and a small variance are required, the problem might become infeasible (the "n.a." entries in the tables). An infeasible MRE-update indicates that the new information is conflicting with the historical ensemble and use of a weighted ensemble may be questionable.

When the MRE-update is asked to match the resulting moments of the weighted ensemble resulting from pdf-N, the results of MRE-M and pdf-N are identical. Conversely, when pdf-N is forced to exactly match the target moments from the forecast, it will yield a result identical to MRE-M with the original target moments. This can, for example, be achieved by changing the parameters of the target distribution of pdf-N in an optimization, until the resulting moments after normalization exactly match the targets. In appendix A, it is shown analytically why the methods yield the same results. It is also shown that the Croley parametric method (CP-M) results in a second order approximation of the MRE-result.

4.4.2 Discussion

An information-theoretical view

The previous results showed that the pdf-ratio method does not exactly match the target moments in the case of large shifts. Stedinger and Kim (2010) discussed whether it is desirable to exactly match target moments, arguing that if the moments in the forecast can not be trusted completely, it might be better to not exactly match them. The question can then be asked what justifies that deviation and in what way the resulting moments should deviate from the forecast. We now take a look at this problem from an information-theoretical perspective.

In the information-theoretical framework, the information is the reduction in uncertainty. A requirement of the distribution of updated weights is therefore that the uncertainty is maximum, given a quantity of information that is added. If less information is taken from the forecast because it is not completely trusted, the maximum permissible reduction in uncertainty will also be less, and vice versa. This can be visualized as a tradeoff between forecast information lost and uncertainty lost. Figure 4.7 shows the tradeoff as a Pareto front. Points below the Pareto front are not attainable, because the change in weights to include a given portion of the forecast information inevitably leads to a given minimum loss of uncertainty. Solutions above the Pareto front, however, lose more uncertainty than permitted by the information. In other words, these weighted ensembles incorporate a gain in information that did not come from the forecast.

The Pareto front in figure 4.7 was obtained by formulating a multi-objective optimization problem and solving it by using the fast global optimization algorithm AMALGAM, developed by Vrugt and Robinson (2007). The problem consisted of minimizing two objectives by finding Pareto-optimal vectors Q of 50 weights for the ensembles. The first objective is the maximization of the entropy of the weights, here plotted as the minimization of the difference with the entropy of uniform weights (Eq. 4.10). This objective is plotted on the vertical axis.

$$\min_{Q} \left\{ H_{\text{uniform}} - H\left(Q\right) \right\} \tag{4.10}$$

The second objective is the minimization Kullback-Leibler divergence of the sought distribution from the closest distribution that exactly matches the target moments. This objective, plotted on the horizontal axis, measures the information loss with respect to the exact forecast.

$$\min_{Q} \left\{ D_{KL}\left(P_{\text{target}} || Q\right) \right\} \tag{4.11}$$

where

$$P_{\text{target}} = \arg\min_{P} \left\{ D_{KL}\left(P||Q\right) \right\} \tag{4.12}$$

and P_{target} is subject to the constraints on sum, mean and variance in equations 4.3, 4.7 and 4.8.

From the figure, it can be seen that the climatic distribution, and the MRE-update both lie on the Pareto-front. In contrast, the Croley parametric method is not Pareto optimal

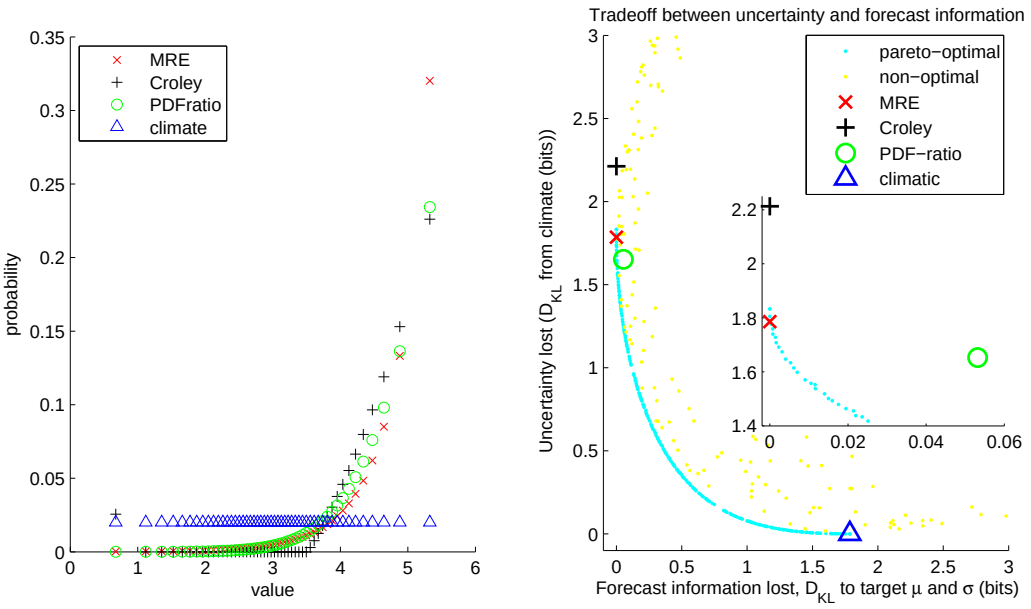


Figure 4.7: Pareto-front showing the trade-off between losing as little uncertainty as possible and losing as little information in the forecast as possible. The different points of the front represent different levels of trust in the forecast. The result is shown here for $\mu_1 = 4.5$ and $\sigma_1 = 0.8$.

according to these criteria. Although it exactly matches the forecast and loses no information from the forecast, it does so with more reduction in uncertainty than strictly needed. Also the pdf-ratio method does not reach the Pareto-front, although is comes close in many cases. However, it is conjectured that when the forecast is seen as two separate pieces of information, one about the mean and one about the variance, the pdf-ratio solution would lie on the Pareto front in a 3 dimensional space, where lost information with respect to mean and variance would be plotted on separate axes.

While all solutions on the pareto front indicate rational sets of weights given varying degrees of trust in the forecast, the MRE-update completely trusts the forecast. The onus is therefore on the forecaster to reflect both the information and the uncertainty about the variable under consideration in that forecast. The MRE-update reflects exactly this information and uncertainty in the weighted ensemble, and does not further increase or decrease the uncertainty. A discussion on why forecasters should communicate carefully chosen summary statistics or preferably their entire probability estimates can be found in Weijs et al. (2010a); see also chapter 5.

About the use of the weighting methods

When forecast TC is received, additional information should be gathered about the moments, support set, or distribution types to assume. If really no other information is available, the MRE-update can be used directly with the forecast, resulting in block adjustment. When information about moments or other appropriately summarized statistics

of the forecast distribution is available, the MRE-update is the most suitable method, as it exactly uses the available information and does not make implicit assumptions.

The use of the optimization based adjustments proposed by Croley (1996; 1997; 2003) are a second order approximation of the MRE-update, but can introduce a reduction of uncertainty that is not supported by the forecast information. The MRE objective function should be preferred over the quadratic objective on information-theoretical grounds. For implementing the MRE-update in places where the Croley method is applied, it suffices to replace the quadratic objective function by relative entropy. This also resolves the problem of many probabilities set to zero.

Because the pdf-ratio method does not need to solve an optimization problem, it is easier to apply and faster then the MRE-update. Another advantage of the pdf-ratio method is that it is relatively easy to include a large amount of information, included in estimated climatic and forecast distributions. In many practical cases, the forecast distribution lies well within the climatic distribution, and normalization is not required in the pdf-ratio method. In those cases, the pdf-ratio method provides a fast and correct adjustment, given that no unfounded assumptions are introduced in the estimation of climatic and forecast distributions. When an extra optimization is done to exactly match the target moments, it can be used as a fast solver for the MRE-update.

The MRE-update uses the full information from the forecast, provided the information contained in the forecast distribution can be converted into mathematical constraints for the optimization problem. In principle the MRE-update offers possibilities to include constraints on for example skew, variance of the log-transformed variable, other quantiles or correlations in a multivariate setting. Many known parametric distributions are in fact maximum entropy distributions for combinations of these types of constraints; see e.g. Singh and Singh (1985); Singh and Guo (1995a). This offers the possibility to reformulate pdf-ratio problems as a MRE-update problem. Conversely, it allows fast parametric solution of the MRE-update by using the pdf-ratio method which is forced to exactly match the constraints.

Making more use of all available information

When we have more information available about the forecast distribution than only mean and variance, like the complete time series of the predictors and responses , it is possible to estimate a joint pdf for them. Bivariate kernel density estimators, as applied by Sharma (2000), would then be a good way to derive continuous climatic and target distributions for the pdf-ratio method. Once one has the joint pdf, the marginal climatic and conditional forecast pdfs can be derived from it and used in the pdf-ratio method. If the conditional distribution from the kernel density estimate can be summarized in a number of constraints, it can also be used in the MRE-update. Figure 4.8 shows for example how an extra constraint on skewness results in a different update.

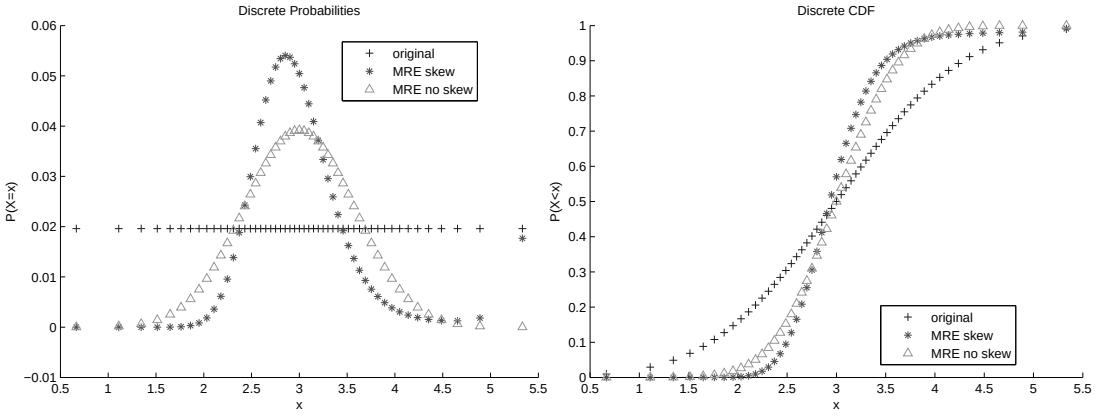


Figure 4.8: Resulting weights (left) and CDF (right) when an extra constraint on skewness is imposed in the MRE-update. The result is shown for $\mu_1 = 3$, $\sigma_1 = 0.5$ and a target skewness of 2.

4.4.3 Conclusions from the theoretical test case

There is an important difference in the information that is conveyed by forecasts in the form of conditional tercile probabilities (TC) and forecasts in the form of a mean and a variance (M). The TC forecast is less informative and taken literally suggests a reweighting of the ensemble in the form of a block adjustment, following Croley (2001). Probability triplets therefore do not seem an appropriate way to convey forecast information. For forecasts in the form of moments, the Croley parametric method (Croley, 2003) makes an adjustment that reduces the uncertainty represented by the ensemble more than is warranted by the information in the forecast. It excludes some of the scenarios in the ensemble by setting their probabilities to zero, without receiving information to justify that.

The pdf-ratio method (Stedinger and Kim, 2010), used with Gaussian distributions, does not use all information in the forecast. It results in an adjustment of the same form as the MRE-update, but with resulting moments that deviate from the moments of the forecast information. The distribution of weights that is found by this version of the pdf-ratio method is not Pareto-optimal in the two dimensional trade-off of lost uncertainty versus lost forecast information. The solution loses more uncertainty than is justified by the partial information that is taken from the forecast. However, the method is Pareto-optimal in a 3D objective space, with objectives minimum lost uncertainty, minimum lost information from the mean and minimum lost information on the variance. This results from the fact that the solution of the pdf-ratio method is identical to the MRE-update solution for the moments that result from it; see appendix A.

An adaptation of the pdf-ratio method is possible, that adjusts the parameters of the target distribution in such a way that the resulting moments exactly match the target moments of the forecast (Jery Stedinger, personal communication). This offers an opportunity to significantly reduce the dimensionality of the optimization problem for the MRE-update in case of a mean-variance forecast. Instead of seeking values for all individual weights, it suffices to optimize the 2 parameters of the target normal distribution and

the normalization factor. Appendix A shows that this amounts to finding the 3 Lagrange multipliers in the analytical solution to the MRE-update.

4.5 Multivariate case

Another important case is updating ensemble probabilities to reflect forecast information on multiple variables (Stedinger and Kim, 2010). For more background on the importance of the multivariate case; see Stedinger and Kim (2007, 2010). For all variables, constraints on mean and variance can be specified separately. Because the size of the ensemble stays the same, the dimensionality of the optimization problem does not increase. The only difference is the addition of more constraints, which results in a slightly higher risk of the optimization problem becoming infeasible. However, tests show that for most practical problems, enough degrees of freedom exist to find a solution. Another important issue is the preservation of cross-correlations (Stedinger and Kim, 2010), especially in cases where risk depends on the joint occurrence of for example high water temperatures and low flows. Preservation of the cross-correlations can be ensured by imposing additional equality constraints on the weighted cross-correlation of the adjusted sample.

Although in this chapter we concentrate on the univariate case, we briefly show some results to demonstrate the potential of the MRE-update also for multivariate updates. We consider the theoretical example from (Stedinger and Kim, 2010) for comparison, using the exact same data. We chose the same bubble plots to show the resulting weights as a function of both variables. The top plot in figure 4.9 shows the results for the MRE-update exactly matching mean and variance of both variables, but without explicitly preserving the initial cross-correlation by including it as a constraint. The middle plot shows the MRE-update result when a constraint enforced the preservation of the initial cross-correlation of 0.8. The bottom plot in figure 4.9 shows the resulting weights when the MRE-update is asked to exactly match the means, variances and cross correlation resulting from the pdf-ratio method with a bivariate normal distribution ($\sigma_{1x} = 1.145$ $\sigma_{1y} = 1.292$ $\rho_1 = 0.751$). Also for the multivariate case, it turned out that the MRE-update using means, variances, and cross-correlation is equivalent to the pdf-ratio method with a bivariate normal distribution, when its moments and cross-correlation would be forced to exactly match the targets.

4.6 Application to ESP forecasts

In this section, the various methods for generating weighted ESP forecasts are applied on a data set with hindcast ESP forecasts for the Columbia river basin in the Pacific Northwest of the USA. The data concerns a “climatic” ensemble. The ensemble traces are modeled flows from the VIC hydrological model (Wood et al., 1992) that were generated using different historical initial conditions and historical weather patterns from 1950 to 2005; see Wood et al. (2005) for a more detailed description of this data set.

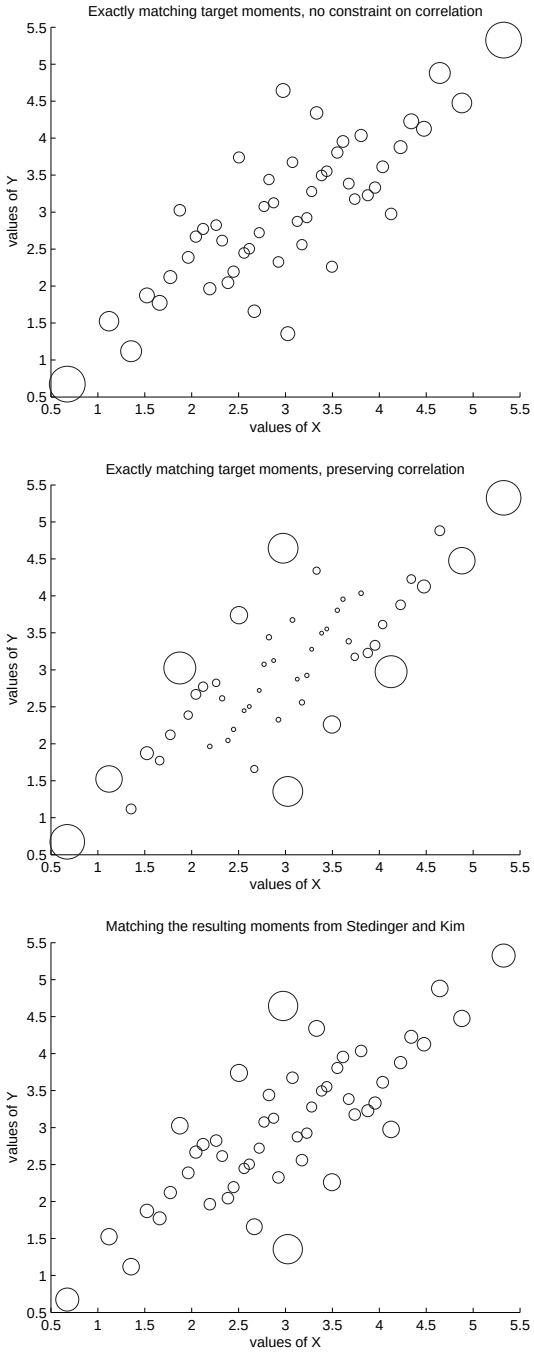


Figure 4.9: Results for the bivariate update of a sample with initial means $\{\mu_{0x}, \mu_{0y}\} = \{3, 3\}$ and initial standard deviations $\{\sigma_{0x}, \sigma_{0y}\} = \{1, 1\}$, the new means and standard deviations are $\{\mu_{1x}, \mu_{1y}\} = \{3, 3\}$ and $\{\sigma_{1x}, \sigma_{1y}\} = \{1.5, 1.5\}$. The area of the circles represent the weights. The above two graphs result from the MRE-update and match these targets exactly. The middle plot maintains the original cross-correlation $\rho_0 = 0.8$, while the upper plot results in a cross-correlation of 0.934. The bottom plot shows the result of the MRE-update for the target moments that are the resulting moments of the pdf-ratio method with bivariate normal target. The resulting weights are identical to the pdf-ratio solution.

4.6.1 Seasonal forecast model

The predictor that was used is the ENSO climate index "Nino3.4" as defined by Trenberth (1997), obtained from the NOAA server[2]. The inclusion of other information, like de phase of the PDO, can probably improve predictions further, but for the purpose of demonstrating the weighting methods, the simple linear regression model with ENSO is considered sufficient. Two different modes, "forecast" and "hindcast", were investigated. In hindcast mode, the whole data set (from 1950 to 2005) about the streamflows and ENSO is used in the regression to derive the linear relation. For each forecast, the current year is excluded from the data set. In forecast mode, which is more representative for a real situation, only data up to the year that is forecast is used. For the forecast for 1970, for example, only the ENSO indexes and flows from 1950 to 1969 are used in the regression model. Furthermore, the ESP forecasts consist of historical weather patterns only from 1950 to 1969. The number of traces in the ESP forecast thus grows each year that a new forecast is made. To avoid problems with small ensemble sizes and little training data for the regression, a "warm up" period of 20 years is used in forecast mode.

The linear model that is found by the regression in hindcast mode is

$$Q_f = 235.33 - 10.96 * \mathrm{ENSO}_{11..2} \quad (R^2 = 0.08) \tag{4.13}$$

where Q_f is the average flow in the months April to September and $\mathrm{ENSO}_{11..2}$ is the average ENSO index for November the year before to February. In forecast mode, the regression coefficients varied from year to year. The deterministic forecasts were subsequently converted to normally distributed probabilistic forecasts using three different methods.

Mean and variance forecast

To obtain a forecast in the form of a mean and a variance of Q_f, the joint distribution of previous forecasts and observed average flows is used. The joint distribution is assumed to be a bivariate normal distribution and the parameters μ_{Q_f}, $\mu_{Q_{\mathrm{obs}}}$, σ_{Q_f}, $\sigma_{Q_{\mathrm{obs}}}$, ρ are estimated using the values of Q_f and Q_{obs} over the years up to the forecast (forecast mode) or for the entire dataset (hindcast mode).

Subsequently, the conditional mean and variance, given the actual forecast $Q_f(t)$ from the regression model, can be calculated using

$$\mu_Q(t) = \mu_{Q_{\mathrm{obs}}} + \rho\sigma_{Q_{\mathrm{obs}}} \frac{Q_f(t) - \mu_{Q_f}}{\sigma_{Q_f}} \tag{4.14}$$

$$\sigma_Q(t) = \sigma_{\mathrm{obs}}\sqrt{1-\rho^2} \tag{4.15}$$

2. `ftp://ftp.cpc.ncep.noaa.gov/wd52dg/data/indices/sstoi.indices`

Kernel Density Estimate (KDE)

Kernel density estimation is a method to estimate an empirical distribution from a sample. It can be interpreted as a smoothed histogram, which is a sum of kernels around the sample values. Sharma (2000) describes a method to use bivariate kernel density estimation of forecasts and responses to derive the joint distribution and the conditional distribution of the response, given a certain value of a predictor (e.g. a climate index or a linear combination of several variables). This method was used to estimate a joint distribution of past ENSO indexes and past streamflow values. For the kernels, a correlated bivariate normal distribution was used, for which the parameters where estimated from the data, using the method described in (Botev, 2006). This method introduces relatively few assumptions and is based on information-theoretic concepts. One still has to make an assumption about the kernel shape, but in most cases, the result is not very sensitive to that (Wand and Jones, 1993). After the joint distribution has been estimated, the conditional distribution at the current predictor value can be found, which is a weighted sum of Gaussian kernels.

The kernel weights and resulting kernel density estimates for the climatic and the conditional distribution are shown in Fig. 4.11. These distributions can then be used in the pdf ratio method, yielding the updated weights for the ensemble traces, which are shown in the same figure.

ESP forecast with information about the hydrological initial state

Next to the weighted climatic ensembles, also ESP forecast were used. The ESP ensembles are obtained by forcing a hydrological model with the observed weather patterns from historical observations, while the initial conditions are the same for each model run. These initial conditions are based on the hydrological state of the catchment in the year of the forecast. As a result of this, the ESP forecasts are based on information that is not available to the weighting methods. A direct comparison of the forecast skill can therefore not lead to conclusions about the weighting methods, but can lead to conclusions about the predictors basin conditions vs. teleconnections; see also Wood et al. (2005); Wood and Lettenmaier (2008).

The weighting can also be combined with the conditional ESP forecasts. The conditional ensemble contains the information about the initial state, in the form of an ensemble that differs from the climatological ensemble. Adjusting the weights of this conditional ensemble also adds the information from the teleconnections.

4.6.2 Results

Figure 4.12 shows how the probabilities are updated by the different methods. For the MRE update on the conditioned ensemble, the weights found for the original ensemble were used. Because the streamflow magnitudes in the different years can reverse order when conditioned on initial basin state, the weights are no longer a smooth function of

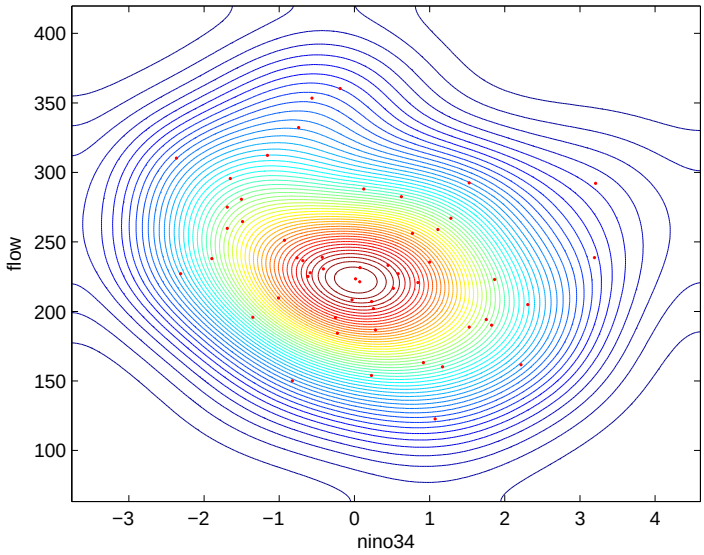


Figure 4.10: Bivariate kernel density estimate of the joint distribution of the ENSO index (November-February) and the average streamflow (April-August).

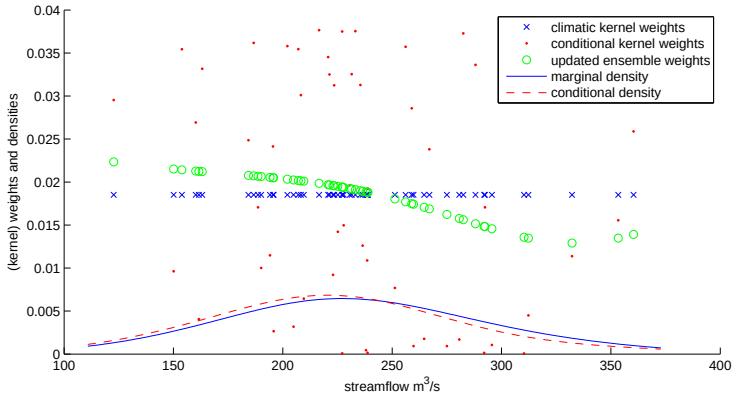


Figure 4.11: An example of the kernel density estimate used in the pdf-ratio method for the year 2003 (hindcast mode).

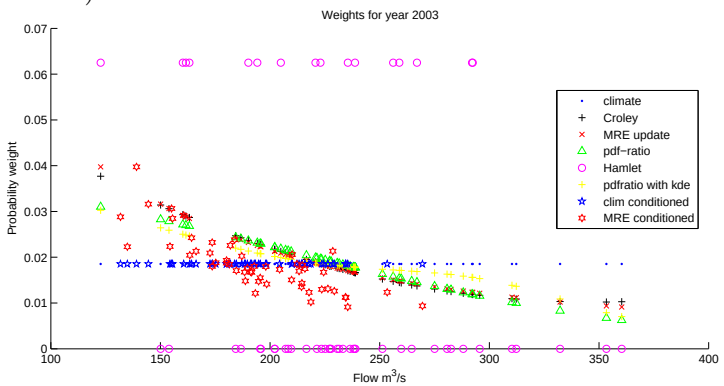


Figure 4.12: Ensemble weights for 2003, plotted against the average streamflow of the each trace from April to September (hindcast mode).

the streamflows. Changes in order occur for example when a relatively dry but warm meteo-year is used as a meteo trace with snowy initial conditions. In the conditional forecast, the increased snowmelt gives a high flow in the melting season, while for the climatic forecast the flow might be one of the lowest.

After the probabilities are updated, the empirical CDFs are shifted. The shifts differ between the various methods, as can be seen from figure 4.13. Weighting methods make vertical shifts to the points in the CDF, while the conditioned ensembles of the ESP forecasts have their CDF points shifted horizontally. The lower graphs make clear that, for this year, the change in CDF due to conditioning on initial basin state is far larger than the change due to weighting based on ENSO information. The observed value for the average flow for the forecast period in 2003 was 188.7 m^3/s. In this case, the shift towards lower values was correct and the forecasts, especially the ones in the lower graphs, were an improvement compared to the climatological forecast.

Over the whole time period, the forecasts were evaluated using the Ranked Probability Skill Score (RPSS) Epstein (1969), which is common for these forecasts. The RPSS is defined as

$$\mathrm{RPSS} = 1 - \frac{\mathrm{RPS}}{\mathrm{RPS}_{\mathrm{clim}}} \tag{4.16}$$

where

$$\mathrm{RPS} = \sum_{i=1}^{n} (P_i - O_i)^2 \tag{4.17}$$

where P_i is the CDF value of the forecast for bin i out of n bins and O_i the CDF value of the observation, which is 0 below the observed value and one above it. $\mathrm{RPS}_{\mathrm{clim}}$ is the RPS score for the reference climatological CDF. An RPSS of 0 therefore indicates no skill over climatology and an RPSS of 1 indicates a perfect forecast. In chapter 5, an alternative skill score based on information theory is presented. In this chapter, the traditional RPSS is used. The resulting scores for the different weighting methods are shown in figures 4.15 (forecast mode) and 4.14 (hindcast mode). It can be observed that the score for the pdfratio forecast using KDE fluctuates less, because the probability distribution is smoothed somewhat by the Gaussian kernels. The methods that depart more from the climatic distribution have a more fluctuating score, most notably the deterministic forecasts. In chapter 5 it is argued that this fluctuation in skill should in fact range between 1 and $-\infty$.

From table 4.5 it becomes clear that in terms of RPSS, all weighted ensemble probabilistic forecasts have some skill over climatology, while the deterministic forecasts have not. The strong variation of forecast quality between the different years make it hard to draw definitive conclusions about the relative performance of the different weighting methods. Longer periods are necessary to demonstrate the practical advantages of one method over the other. In forecast mode, results are most representative of a real situation, but only a short time series is available for evaluation.

The conditioned ensemble yields an RPSS of 0.4528 for uniform weights and 0.4310 for an MRE-updated set of weights (both in forecast mode). In hindcast mode, results are

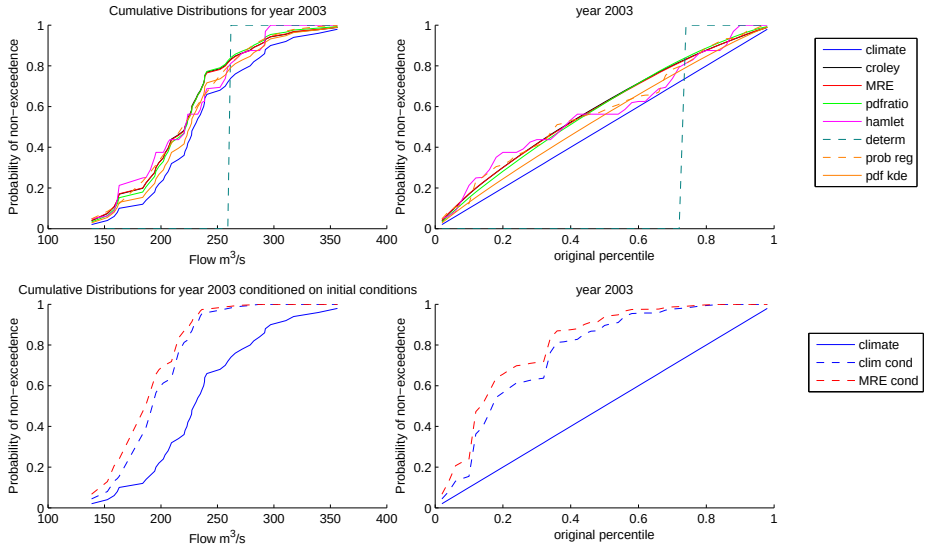


Figure 4.13: The empirical CDFs for the different methods. The right figures are quantile-quantile (QQ) plots, clearly showing the changes by the weighting. The lower two graphs show the CDFs for the ESP forecasts that are conditioned on initial basin conditions.

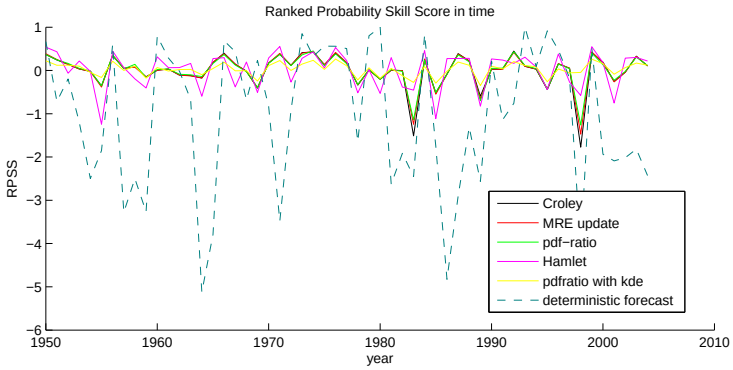


Figure 4.14: The RPSS for the weighted ensemble forecasts by the different weighting methods for the whole forecast period (hindcast mode)..

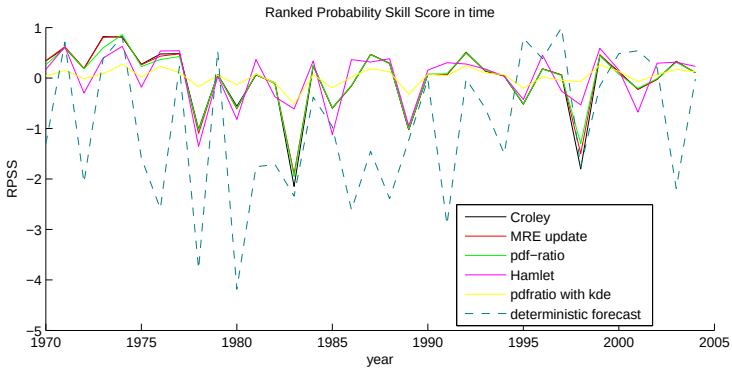


Figure 4.15: The RPSS for the weighted ensemble forecasts by the different weighting methods for the period from 1970 (forecast mode, starting from 20 traces).

Method	RPSS forecast	RPSS hindcast
Croley	0.0365	0.0228
MRE	0.0479	0.0327
pdfratio	0.0323	0.0350
Hamlet	0.0182	0.0214
ESP	0.4528	0.4487
MRE weighted ESP	0.4310	0.4298
Pdfratio KDE	0.0247	0.0356
Deterministic	-0.3970	-0.4638

Table 4.5: The resulting skill score for the different methods.

0.4487 for uniform weights and 0.4298 of the updated weights. First of all, this shows that conditioning on initial basin conditions leads to far more skillful forecasts than weighting on the basis of ENSO. The deterioration when going from the ESP forecast to the weighted ESP forecast could be attributed to the fact the initial basin conditions and the ENSO phase are not independent. By basing the weights on a regression of ENSO with the climatic flows, the total dependence of the flows on ENSO is reflected in the weights. The part of this dependence that is also present in the initial basin conditions is therefore used twice. Using the same information twice results in overconfident forecasts, which leads to a reduction in forecast skill. Using the MRE-update with weights based on regression of ENSO with the conditioned ensemble prevents using the information twice. This leads to an RPSS of 0.0117 with the conditioned ensemble as a reference, showing that the weighting does not add much info over the conditional ESP forecast. Most of the ENSO information, is already present in the basin conditions.

4.6.3 Discussion

The ensemble weighting methods seem to produce an improvement of skill over the method of selecting traces, similar to Hamlet and Lettenmaier (1999). Weighting based on ENSO has less predictive power than generating ensembles conditioned on initial basin conditions. The advantage of the weighting method, however, is that it can also be applied to historically measured flows, without the need for a detailed hydrological model and measurements of initial basin conditions. Combining the weighting methods with the conditional ensemble would be an interesting direction of future research.

The weighting methods can add information to the ESP forecasts based on any feature of the weather pattern that correlates with ENSO or other long timescale predictors. Such features could, for example, also include timing of the precipitation or average temperature in some months. This information can be added to the ensemble by using the multivariate version of the MRE-update. Care should be taken, however, not to include the same information twice.

The extension of the pdf-ratio method using kernel density estimate showed good results in hindcast mode, but in forecast mode, it performed worse than the pdf-ratio with normal distributions. This may indicate the kernel density estimates are over-fitted to the data.

However, we can not draw any general conclusions based on this single case, which just served to demonstrate the different methodologies.

4.6.4 Conclusion

The different weighting methods show a small positive skill compared to climatology. The skill fluctuates from year to year, which makes it difficult to draw significant conclusions from this case about the performance of the different weighting methods. The skill is also mostly determined by the quality of the forecasts and the predictors they are based on and less by the weighting method to combine the forecasts with an ensemble. However, the MRE-update shows promising results, also in a practical case.

4.7 Conclusions and recommendations

In this chapter, we introduced the minimum relative entropy update (MRE-update) as an approach to update ensemble member probabilities. Our method is based on the minimization of relative entropy, with forecast information imposed as constraints. The main advantage of the method is that it optimally combines available climatic and forecast information, without introducing extra assumptions. Results were compared with three existing methods to make probability-adjustments to ensembles, based on different type of forecast information. We considered forecast information given in the form of conditional tercile probabilities, a normal distribution and a given mean and variance. Analysis of the results from an information-theoretical viewpoint explicitly revealed the differences in information contained in these different types of forecasts and the way existing methods interpret them.

The block adjustment that results from the Croley nonparametric method may be undesirable due to the discontinuities in weights at the arbitrarily selected quantiles. However, the result is in line with the literal information-theoretical interpretation of the probability triplets. When interpreting the forecasts in any other way, information is added. It is important to be aware of this and think carefully about what information is added. Conversely, forecasts that require this extra interpretation are in fact incomplete, leaving too much interpretation to the user of the forecasts. Ideally, seasonal forecasts should provide pieces of information that are a good summary of the probability estimate. For smooth distributions, a mean and variance are more appropriate than probability triplets. See also chapter 5 and Weijs et al. (2010a) for more discussion on how forecasts should be presented to be most informative.

The information contained in the mean-variance forecast and in a normal forecast distribution is the same. The MRE-update results in a weighted ensemble that exactly incorporates this information. The pdf-ratio method diverges from the forecast information by not exactly matching the given moments. From an information-theoretical perspective, it is unclear how to justify this divergence. A multi-objective optimization was performed to

find a Pareto-front that represents the tradeoff between lost information from the forecast and lost initial uncertainty. An analysis of the methods in this objective-space revealed that in some cases, the pdf-ratio method reduces uncertainty more than is justified by the partial information taken from the forecast. This results in a solution that is not Pareto-optimal. Also the Croley parametric method lies above the Pareto-front. It uses the full information from the forecast, but reduces uncertainty more than that information permits. One of the symptoms of this false certainty are the ensemble members which get weight zero, although nothing in the forecast rules them out. By definition of the chosen objectives, the MRE-update results in a Pareto-optimal solution in which the complete information from the forecast is used and no other information is added to the weighted ensemble.

The pdf-ratio method has the advantage that it is fast and does not require an optimization search for all individual weights. An adaptation of the pdf-ratio method that includes a search for parameters that result in an exact match of the target moments is possible. This is equivalent to finding the values of the Lagrange multipliers for the analytical solution of the MRE-update (see appendix A) and results in an optimization problem that is much easier to solve. In this chapter, the equivalence for the univariate and bivariate normal distributions was demonstrated, but it is anticipated that similar results can be found for other distributions for which sufficient statistics exist.

In a test case in a practical application, the variations in skill of the seasonal mean-variance forecasts for Columbia river streamflows were too large to enable strong conclusions on the performance of the weighting methods. The methods showed similar performance, which indicated a small skill over climatology, which was considerably less than the skill of ensembles that where conditioned on initial basin conditions. Optimally combining the information from ENSO with the basin state information is an important challenge. The ENSO-based forecasts have the advantage that they can be extended to longer lead times, and can be issued before the snow accumulates in the basin. For reservoir operation, having such forecasts, with a seasonal jump in predictive power, presents an interesting control problem, where it might be necessary to include knowledge of the future growth of information when snow falls into the operation strategy; see also the discussion in chapter 7.

The MRE-update can incorporate information that can be formulated in terms of constraints. This chapter showed also how skew can be included. In addition, a multivariate example was given in which information in both means and variances was matched while also preserving initial cross-correlation. Generation weighted ensembles using information from non-parametric forecast pdfs, such for example obtained from kernel density estimation, remains an open issue.

The Matlab-source code (still in development) for the MRE-update will be available from the website `www.hydroinfotheory.net`.

Chapter 5

Using information theory to measure forecast quality

> *"Ignorance is preferable to error, and he is less remote from the truth who believes nothing than he who believes what is wrong."*
> - Thomas Jefferson, 1781

Abstract - This chapter[1] presents a score that can be used for evaluating probabilistic forecasts of discrete events. The score is a reinterpretation of the logarithmic score or Ignorance score, now formulated as the relative entropy or Kullback-Leibler divergence of the forecast distribution from the observation distribution. Using the information-theoretical concepts of entropy and relative entropy, a decomposition into three components is presented, analogous to the classical decomposition of the Brier score, which is also extended to the case of uncertain observations. The information-theoretical twins of the components uncertainty, resolution and reliability provide diagnostic information about the quality of forecasts. The overall score measures the the uncertainty that remains after the forecast. As was shown recently, information theory provides a sound framework for forecast verification. The new decomposition, which has proven to be very useful for the Brier score and is widely used, can help acceptance of information-theoretical scores in meteorology and hydrology.

5.1 Introduction

Forecasts are intended to provide information to the user. Forecast verification is the assessment of the quality of a single forecast or forecasting scheme (Jolliffe and Stephenson, 2008). Verification should therefore assess the quality of the information provided by the

1. based on:
   - S.V. Weijs, R. van Nooijen, and N. van de Giesen. Kullback–Leibler divergence as a forecast skill score with classic reliability–resolution–uncertainty decomposition. *Monthly Weather Review*, 138, (9): 3387–3399, September 2010
   - S.V. Weijs, G. Schoups, and N. van de Giesen. Why hydrological forecasts should be evaluated using information theory. *Hydrology and Earth System Sciences* ,14 (12), 2545–2558, 2010
   - S.V. Weijs and N. van de Giesen, Accounting for observational uncertainty in forecast verification: an information–theoretical view on forecasts, observations and truth, *Monthly Weather Review*, early online release, 2011.

forecast. It is important here to note the distinction between quality, which depends on the correspondence between forecasts and observations, and value, which depends on the benefits of forecasts to users (Murphy, 1993). In this chapter, it is assumed that the verification is intended to quantitatively measure quality. Several scores and visualization techniques have been developed that measure certain desirable properties of forecasts with the purpose of assessing their quality. One of the most commonly used skill scores (Stephenson et al., 2008) is the Brier score (BS) (Brier, 1950), which is applicable to probabilistic forecasts of binary events. The Brier skill score (BSS) measures the BS relative to some reference forecast, which is usually climatology. Murphy (1973) showed that the BS can be decomposed into three components; uncertainty, resolution and reliability. These components give insight into some different aspects of forecast quality. The first component, uncertainty, measures the inherent uncertainty in the process that is forecast. Resolution measures how much of this uncertainty is explained by the forecast. Reliability measures the bias in the probability estimates of the probabilistic forecasts. A perfect forecast has a resolution that is equal to (fully explains) the uncertainty and a perfect reliability.

Information theory provides a framework for measuring information and uncertainty; see Cover and Thomas, 2006 for a good introduction. As forecast verification should assess the information that the forecaster provides to the user, using information theory for forecast verification appears to be a logical choice. A concept central to information theory is the measure of uncertainty named entropy. However, consulting two standard works about forecast verification, it was noted that the word entropy is mentioned only thrice in (Jolliffe and Stephenson, 2003) and not one single time in (Wilks, 1995). This indicates that the use of information-theoretical measures for forecast verification is not yet widespread, although some important work has been done by Roulston and Smith (2002); Ahrens and Walser (2008); Leung and North (1990); Kleeman (2002).

Leung and North (1990) used information-theoretical measures like entropy and transinformation in relation to predictability. Kleeman (2002) proposed to use the relative entropy between the climatic and the forecast distribution to measure predictability. The applications of information theory in the framework of predictability are mostly concerned with modeled distributions of states and how uncertainty evolves over time. Forecast verification, however, is concerned with comparing observed values with the forecast probability distributions. Roulston and Smith (2002) introduced the Ignorance score, a logarithmic score for forecast verification, reinterpreting the logarithmic score (Good, 1952) from an information-theoretical point of view. They related their score to relative entropy between the forecast distribution and the "true PDF", which they defined as "the PDF of consistent initial conditions evolved forward in time under the dynamics of the real atmosphere". Ahrens and Walser (2008) proposed information-theoretical skill scores to be applied to cumulative probabilities of multi-category forecasts. Very recently, Benedetti (2010) showed that the logarithmic score is a unique measure of forecast goodness. He showed that the logarithmic score is the only score that simultaneously satisfies three basic requirements for such a measure. These requirements are additivity, locality (which he interprets as exclusive dependence on physical observations) and strictly proper behav-

ior. For a discussion on these requirements, see Benedetti (2010). Furthermore, Benedetti (2010) analyzed the Brier score and showed that it is equivalent to a second order approximation of the logarithmic score. He concludes that lasting success of the Brier score can be explained by the fact that it is an approximation of the logarithmic score. Benedetti also mentions the well-known and useful decomposition of the Brier score into uncertainty, resolution and reliability as a possible reason for its popularity.

This chapter, follows a similar route as Benedetti, but from a different direction. From an analogy with the Brier score, it is proposed to use the Kullback-Leibler divergence (or relative entropy) of the observation from the forecast distribution as a measure for forecast verification. The score is named 'divergence score' (DS). When assuming perfect observations, DS is equal to the Ignorance score or logarithmic score, and can be seen as a new reinterpretation of Ignorance as the Kullback-Leibler divergence from the observation to the forecast distribution. By presenting a new decomposition into uncertainty, resolution and reliability, analogous to the well-known decomposition of the Brier score (Murphy, 1973), insight is provided in the way the divergence score measures the information content of probabilistic binary forecasts. The decomposition can help acceptance and wider application of the logarithmic score in meteorology and hydrology.

Section 5.2 of this chapter presents the mathematical formulation of the DS and its components. Section 5.2 also shows the analogy with the Brier score components. Section 5.3 compares the divergence score with existing information-theoretical scores. It is shown that the DS is actually a reinterpretation of the Ignorance score (Roulston and Smith, 2002) and that one of the ranked mutual information scores defined by Ahrens and Walser (2008) is equal to the skill score version of DS, when the reliability component is neglected (perfect calibration assumed). A generalization to multi-category forecasts is presented in Section 5.4. The inherent difficulty found in formulating skill scores for ordinal category forecasts is also analyzed and leads to the idea that this can be explained by explicitly distinguishing between information and useful information for some specific user. This distinction provides some insights in the roles of the forecaster and the user of the forecast. Section 5.5 presents an application to a real data set of precipitation forecasts. Section 5.6.3 introduces the cross entropy score and another decomposition, which become relevant if the observations are not perfect and contain uncertainties. A paradox which arises from the use of deterministic forecasts is analyzed in section 5.7. Section 5.8 summarizes the conclusions and restates the main arguments for adopting the divergence score.

5.2 Definition of the divergence score

5.2.1 Background

By viewing the Brier score as a quadratic distance measure and translating it into the information-theoretical measures for uncertainty and divergence of one distribution from another, in this chapter an information-theoretical twin of the Brier score and its components is formulated. First, some notation is introduced, followed by formulation of the

Brier score. Then the information-theoretical concept of relative entropy is presented as an alternative scoring rule. In the second part of this section, it is shown how the new score can be decomposed into the classical Brier score components; uncertainty, resolution and reliability.

5.2.2 Definitions

Consider a binary event, like a day without rainfall or with rainfall. This can be seen as a stochastic process with two possible outcomes. The outcome of the event can be represented in a probability mass function (PMF). For the case of binary events, the empirical PMF of the event after the outcome has been observed is a two element vector, denoted by $\mathbf{o} = (1-o, o)^T$. Assuming certainty in the observations, $o \in \{0, 1\}$. Therefore, $\mathbf{o} = (0, 1)^T$ if it rained and $(1, 0)^T$ otherwise. Now suppose a probabilistic forecast of the outcome of the binary event is issued in the form of a probability of occurrence f. This can also be written as a forecast PMF $\mathbf{f} = (1-f, f)^T$, with $f \in [0, 1]$. If for example an 80% chance of rainfall is forecast, this is denoted as $\mathbf{f} = (0.2, 0.8)^T$.

The Brier score for a single forecast at time t measures distance between observation and forecast PMFs by the square Euclidean distance

$$\mathrm{BS}_t = 2\left(f_t - o_t\right)^2 = \left(\mathbf{f}_t - \mathbf{o}_t\right)^T \left(\mathbf{f}_t - \mathbf{o}_t\right) \tag{5.1}$$

For a series of forecasts and observations, the Brier score is simply the average of the Brier scores for the individual forecasts.

$$\mathrm{BS} = \frac{1}{N} \sum_{t=1}^{N} \left(\mathbf{f}_t - \mathbf{o}_t\right)^T \left(\mathbf{f}_t - \mathbf{o}_t\right) \tag{5.2}$$

Note that this is the original definition by Brier (1950). Nowadays, the Brier score is almost always defined as half the value of (5.2) (Ahrens and Walser, 2008).

5.2.3 The divergence score

The entropy of a forecast distribution gives an indication of the uncertainty that is represented by that forecast. It measures the expected surprise $S_{\mathbf{f}}$ upon hearing the true outcome, when the state of knowledge is $\mathbf{f}$, with respect to probability distribution $\mathbf{f}$. For example, the uncertainty associated with a binary probabilistic forecast of 70% chance of precipitation, $\mathbf{f} = (0.3, 0.7)^T$, is

$$H(\mathbf{f}) = E_{\mathbf{f}}\left\{S_{\mathbf{f}}\right\} = -\sum_{i=1}^{n} \left[\mathbf{f}\right]_i \log\left[\mathbf{f}\right]_i = 0.88 \text{ bits} \tag{5.3}$$

where $H(\mathbf{f})$ is the entropy of $\mathbf{f}$, calculated in the unit "bits" because the logarithm is taken to the base 2 (throughout this chapter). $E_{\mathbf{f}}$ denotes the expectation operator with

respect to $\mathbf{f}$, and n is the number of categories in which the outcome can fall, in this case 2.

Next to this entropy-measure, introduced by Shannon (1948), there is also a definition of relative entropy, or Kullback-Leibler divergence (Kullback and Leibler, 1951). This is a measure for the expected amount of additional surprise a person is expected to experience, compared to another person having a more accurate and reliable probability estimate. Therefore, it is a relative uncertainty. The divergence score is based on this idea and measures the expected extra surprise that a person having the forecast $\mathbf{f}_t$ for a instance t will experience, compared to a person knowing the observation $\mathbf{o}_t$

$$\mathrm{DS}_t = \mathrm{D}_{KL}(\mathbf{o}_t||\mathbf{f}_t) = E_{\mathbf{o}_t}\left\{S_{\mathbf{f}_t} - S_{\mathbf{o}_t}\right\} = \sum_{i=1}^{n} [\mathbf{o}_t]_i \log\left(\frac{[\mathbf{o}_t]_i}{[\mathbf{f}_t]_i}\right) \tag{5.4}$$

The divergence score (DS), replaces the quadratic distance from the BS with the Kullback-Leibler divergence. For one single forecast, the DS functions as a scoring rule. It is the Kullback-Leibler divergence of the forecast distribution from the observation distribution over the $n = 2$ possible events i.

The DS over a series of forecast-observation pairs measures the average divergence of the forecast distribution from the observation

$$\mathrm{DS} = \frac{1}{N}\sum_{t=1}^{N} D_{KL}(\mathbf{o}_t||\mathbf{f}_t) \tag{5.5}$$

The divergence score can be interpreted as the information gain when one moves from the prior forecast distribution to the observation distribution. When this information gain is zero, the forecast already contained all the information that is in the observation and therefore is perfect. If the information gain from the forecast to the certain observation is equal to the climatological uncertainty, the forecast did not contain more information than the climate, and therefore was useless. Another way to view the divergence score is the remaining uncertainty about the true outcome, after having received the forecast; see Fig. 5.1.

5.2.4 Decomposition

The new score can be decomposed in a similar way as the Brier score. The classical decomposition of the Brier score (BS) into reliability (REL) , resolution (RES) and uncertainty (UNC) is

$$\mathrm{BS} = \mathrm{REL}_{\mathrm{BS}} - \mathrm{RES}_{\mathrm{BS}} + \mathrm{UNC}_{\mathrm{BS}} \tag{5.6}$$

$$\mathrm{BS} = \frac{1}{N}\sum_{k=1}^{K} n_k(\mathbf{f_k} - \bar{\mathbf{o}}_\mathbf{k})^2 - \frac{1}{N}\sum_{k=1}^{K} n_k(\bar{\mathbf{o}}_\mathbf{k} - \bar{\mathbf{o}})^2 + \bar{\mathbf{o}}\,(1 - \bar{\mathbf{o}}) \tag{5.7}$$

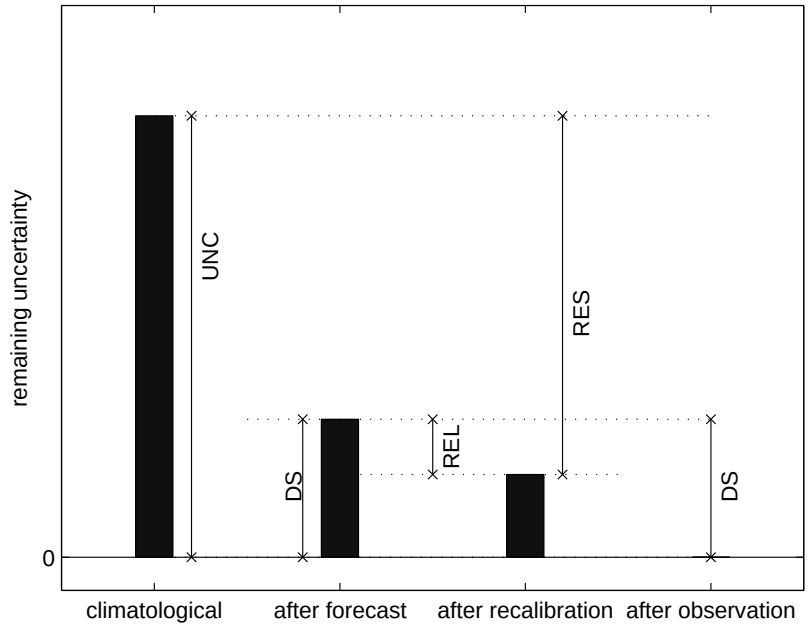


Figure 5.1: The components measure the differences in uncertainty at different moments in forecasting process, judged in hindsight. The divergence score measures the remaining uncertainty after taking the forecast at face-value.

with N being the total number of forecasts issued, K the number of unique forecasts issued, $\bar{\mathbf{o}} = \sum_{t=1}^{N} \mathbf{o}/N$ the observed climatological base rate for the event to occur, n_k the number of forecasts with the same probability category and $\overline{\mathbf{o}}_\mathbf{k}$ the observed frequency, given forecasts of probability $\mathbf{f_k}$. The reliability and resolution terms in (5.7) are summations of some distance measure between two binary probability distributions, while the last term measures the uncertainty in the climatic distribution, using a polynomial of degree two. Now, information-theoretical twins are presented for each of the three quadratic components of the BS, using entropy and relative entropy. It is shown that they add up to the divergence score proposed earlier.

The first component, reliability, measures the conditional bias in the forecast probabilities. In the DS, it is the expected divergence of the observed probability distribution from the forecast probability distribution, both stratified (conditioned) on all issued forecast probabilities. In the ideal case, the observed frequency is equal to the forecast probability for all of the issued forecast probabilities. Only in this case is the reliability 0. This is referred to as a perfectly calibrated forecast. Note that reliability is defined in the opposite direction as the meaning of the word in the English language. A perfectly reliable forecast has a reliability of 0. The reliability can be calculated with

$$\mathrm{REL}_{\mathrm{DS}} = \frac{1}{N} \sum_{k=1}^{K} n_k \mathrm{D}_{KL}(\bar{\mathbf{o}}_k || \mathbf{f}_k) \tag{5.8}$$

with N being the total number and K the number of unique forecasts issued, n_k the number of forecasts with the same probability category, $\overline{\mathbf{o}}_\mathbf{k}$ the observed frequency distribution for forecasts in group k and $\mathbf{f}_k$ the forecast PMF for group k.

The second component, resolution, measures the reduction in climatic uncertainty. It can be seen as the amount of information in the forecast. In the DS, it is defined as the expected divergence of the conditional frequencies from the marginal frequency of occurrence. The minimum resolution is 0, which occurs when the climatological probability is always forecast or the forecasts are completely random. The resolution measures the amount of uncertainty in the observation explained by the forecast. In the ideal case the resolution is equal to the uncertainty, which means all uncertainty is explained. This is only the case for a deterministic forecast that is either always right or always wrong. In the last case, the forecast needs to be recalibrated.

$$\mathrm{RES}_{\mathrm{DS}} = \frac{1}{N}\sum_{k=1}^{K} n_k \mathrm{D}_{KL}(\bar{\mathbf{o}}_k||\bar{\mathbf{o}}) \tag{5.9}$$

From Equation (5.9) it becomes clear that the resolution term is the expectation over all forecast probabilities of the divergence from the conditional probability of occurrence to the marginal probability of occurrence. In information theory this quantity is known as the mutual information (I) between the forecasts and the observation.

$$\begin{aligned} \mathrm{RES}_{\mathrm{DS}} &= \underset{k}{\mathrm{E}}\{\mathrm{D}_{KL}(\bar{\mathbf{o}}_k||\bar{\mathbf{o}})\} && (5.10)\\ &= \underset{k}{\mathrm{E}}\{\mathrm{D}_{KL}((\bar{\mathbf{o}}|\mathbf{f}_k)||\bar{\mathbf{o}})\} = I(\mathbf{f};\mathbf{o}) && (5.11) \end{aligned}$$

The third component, uncertainty, measures the initial uncertainty about the event. This observational uncertainty is measured by the entropy of the climatological distribution ($\mathrm{H}(\bar{\mathbf{o}})$). It is a function of the climatological base rate only and does not depend on the forecast. The uncertainty is maximum if the probability of occurrence is 0.5 and zero if the probability is either 0 or 1.

$$\mathrm{UNC}_{\mathrm{DS}} = \mathrm{H}(\bar{\mathbf{o}}) = -\sum_{i=1}^{n}\{[\bar{\mathbf{o}}]_i \log[\bar{\mathbf{o}}]_i\} \tag{5.12}$$

Like for the BS, for a single forecast-observation pair, uncertainty and resolution are 0 and the total score is equal to the reliability, which acts as a scoring rule. Over a larger number of forecasts uncertainty approaches climatic uncertainty and reliability should go to zero if the forecast is well calibrated. In appendix B it is shown that, just like in the Brier score, the relation DS = REL − RES + UNC holds. The relation between the components and the total score (DS) is

$$\mathrm{DS} = \frac{1}{N}\sum_{t=1}^{N}\mathrm{D}_{KL}(\mathbf{o}_t||\mathbf{f}_t) = \frac{1}{N}\sum_{k=1}^{K} n_k \mathrm{D}_{KL}(\bar{\mathbf{o}}_k||\mathbf{f}_k) - \frac{1}{N}\sum_{k=1}^{K} n_k \mathrm{D}_{KL}(\bar{\mathbf{o}}_k||\bar{\mathbf{o}}) + \mathrm{H}(\bar{\mathbf{o}}) \tag{5.13}$$

Note that this decomposition is valid for forecasts of events with an arbitrary number of categories and is not restricted to the binary case.

5.2.5 Relation to Brier score and its components

For the binary case, the Brier score can be seen as a second order approximation of the divergence score (also noted by Benedetti (2010)). Both scores have their minimum only with a perfect forecast. When the forecast is not perfect, the Brier score is symmetric in the error in probabilities, while the divergence score is not, except for the case where the true forecast probability is 0.5. Therefore the divergence score, like the logarithmic score, is a double valued function of the Brier score (Roulston and Smith, 2002). Consequently, when two forecasting systems are compared, the forecasting system with the higher Brier score may have the lower divergence score.

The uncertainty component in the Brier score is a second order approximation of the uncertainty term in the divergence score (entropy), with the same location of zero uncertainty (100% probability of one of the two events) and maximum uncertainty (equiprobable events with 50% probability; see Fig. 5.2). In the Brier score, the maximum of the uncertainty component is 0.5, while in the divergence score it is 1 (bit). Resolution in the Brier score is the variance of conditional mean probabilities. It is a mean of squared deviations from the climatic probability. Resolution in the divergence score is a mean of divergences. Divergences are asymmetric in probabilities. The resolution in both the Brier and the divergence score can take on values between zero and the uncertainty. In both scores, it can be seen as the amount of uncertainty explained. The resolution in the Brier score is the second order approximation of the resolution of the divergence score, satisfying the condition that the minimum is zero and in the same location (the climatic probability) and that the maximum possible value is equal to the inherent uncertainty of the forecast event; see Fig. 5.3. Reliability in the Brier score is bounded between zero and one, while in the divergence score, the reliability can reach infinity; see Fig. 5.4. This is the case when wrong deterministic forecasts are issued. Generally the reliability in the divergence score is especially sensitive to events with near deterministic wrong forecasts. Overconfident forecasting is therefore sanctioned more heavily than in the Brier score.

5.2.6 Normalization to a skill score

Because the Brier score depends on the climatological probability, which is independent of the forecast quality, it is common practice to normalize it to the Brier skill score (BSS) with the climatology forecast as a reference. A perfect forecast is taken as a second reference. For the DS it is possible to use the same normalization procedure, yielding the divergence skill score (DSS).

$$\mathrm{DSS} = \frac{\mathrm{DS} - \mathrm{DS}_{\mathrm{ref}}}{\mathrm{DS}_{\mathrm{perf}} - \mathrm{DS}_{\mathrm{ref}}} = 1 - \frac{\mathrm{DS}}{\mathrm{DS}_{\mathrm{ref}}} \tag{5.14}$$

The score for a perfect forecast ($\mathrm{DS}_{\mathrm{perf}}$) is zero. In the climatological forecast ($\mathrm{DS}_{\mathrm{ref}}$), both resolution and reliability are 0 (perfect reliability, no resolution). The DSS therefore reduces to:

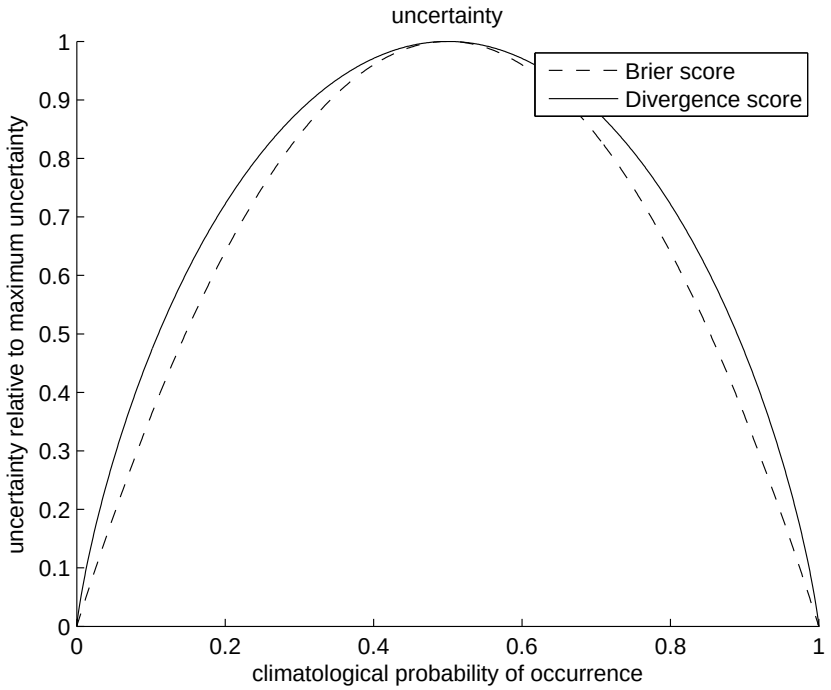


Figure 5.2: The uncertainty component of the Brier score is a second order approximation of the entropy, with coinciding minimum and maximum values. The uncertainty for the Brier score is divided by its maximum value of 0.5, to allow a clear comparison with the uncertainty term of the divergence score, measured in bits.

$$\mathrm{DSS} = 1 - \frac{\mathrm{UNC} - \mathrm{RES} + \mathrm{REL}}{\mathrm{UNC}} = \frac{\mathrm{RES} - \mathrm{REL}}{\mathrm{UNC}} \tag{5.15}$$

This leads to a positively oriented skill score that becomes one for a perfect forecast and zero for a forecast of always the climatological probability. Also a completely random forecast that has a marginal distribution equal to climatology gets a zero skill score. Negative (i.e. "worse than climate") skill scores are possible if the reliability is larger (worse) than the resolution. In case the resolution is significant, calibration of the forecast can yield a positive skill score, meaning that a decision maker using the recalibrated forecast is better off than a decision maker using climatology or a random strategy. This shows the importance of looking at the individual components when diagnosing a forecast system's performance.

Summarizing, the divergence score and its components combine two types of measures to replace the quadratic components in the Brier score decomposition. Firstly, the quadratic distances between probability distributions are replaced by Kullback-Leibler divergences, which are asymmetric. Care should therefore be taken in which direction the divergence is calculated. Secondly, the polynomial uncertainty term is replaced by the entropy of the climatology distribution. The total scores and components are visualized in Fig. 5.1.

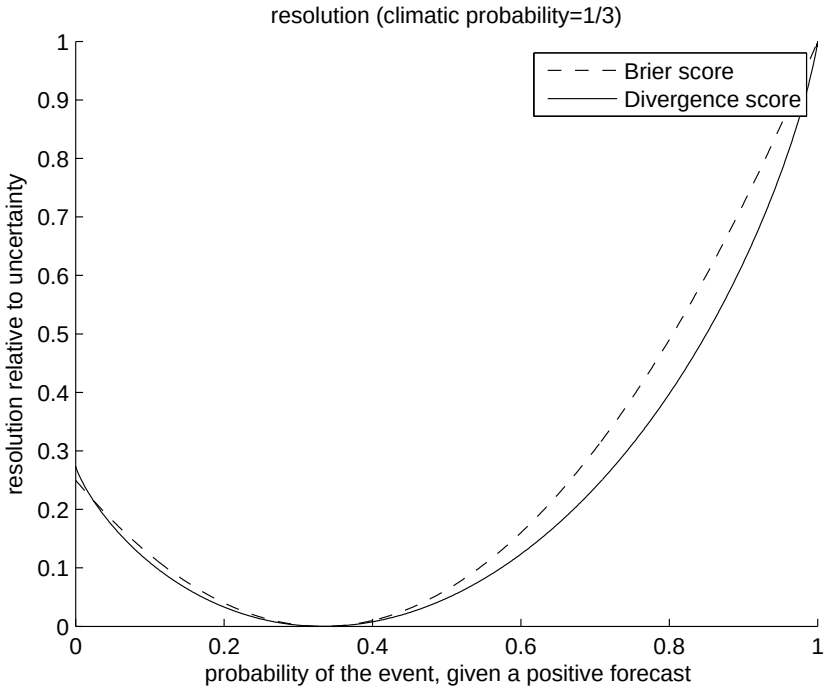


Figure 5.3: The resolution term of the divergence score is asymmetric in probability while the Brier score resolution is not.

5.3 Relation to existing information-theoretical scores

5.3.1 Relation to the ranked mutual information skill scores

Ahrens and Walser (2008) proposed the ranked mutual information skill score (RMIS). The score is intended for use with multi-category forecasts, which will be treated later in this chapter. For the special case of forecasts of binary events, the RMIS_O can be written as the mutual information between forecasts and observations divided by the entropy of the observations

$$\mathrm{RMIS}_O = \frac{I(\mathbf{f}, \mathbf{o})}{H(\bar{\mathbf{o}})} \tag{5.16}$$

When comparing Eq. (5.16) with Eqs. (5.10), (5.12) and (5.15), it becomes clear that

$$\mathrm{RMIS}_O = \frac{\mathrm{RES}_{\mathrm{DSS}}}{\mathrm{UNC}_{\mathrm{DSS}}} \tag{5.17}$$

This means that for the case of a binary forecast, RMIS_O equals the DSS in case the reliability is perfect (zero). In case the forecast is not well calibrated, RMIS_O neglects the reliability component and measures the amount of information that would be available to a user after calibration. The DSS measures the information conveyed to a user taking the forecasts at face-value. The individual components of the DSS also indicate the potentially extractable information as measured by the RMIS_O.

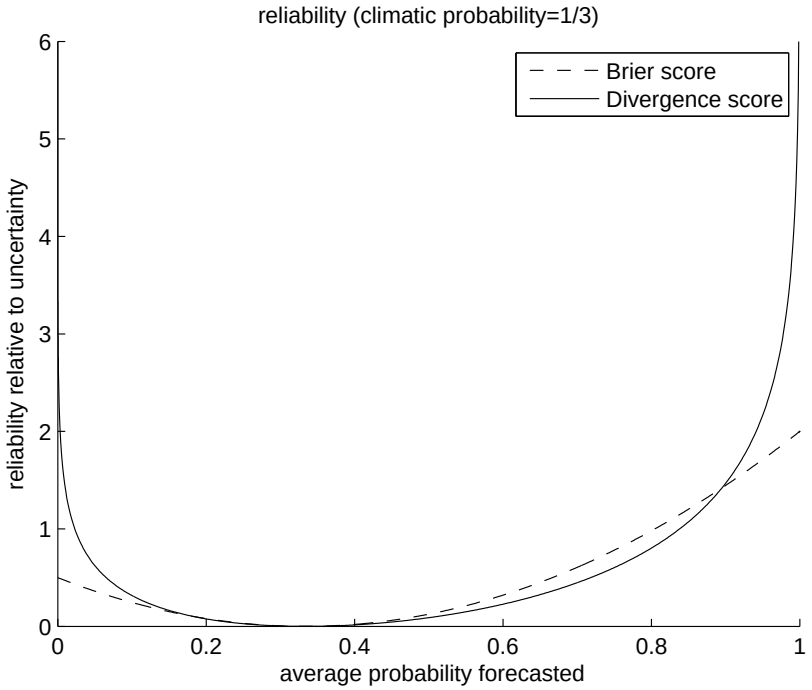


Figure 5.4: The reliability is asymmetric in the DS, while symmetric in the BS. In the BS, it is bounded, while in the DS, it can reach infinity.

5.3.2 Equivalence to the Ignorance score

Roulston and Smith (2002) defined the Ignorance score from the viewpoint of using the forecast probability distribution as basis for a data compression scheme. The scoring rule measures the Ignorance or information deficit of a forecaster, compared to a person knowing the true outcome of the event (j). The Ignorance scoring rule is defined as

$$\mathrm{IGN} = -\log_2 f_j \tag{5.18}$$

In which f_j is the probability that the forecaster had assigned to the event that actually occurred. The Ignorance score is a reinterpretation of the logarithmic score by Good (1952).

By expanding the relative entropy measure that is used as a scoring rule, it becomes clear that divergence from the certain observation PMF ($\mathbf{o}$) to the forecast PMF ($\mathbf{f}$) is actually the same as the Ignorance (IGN) or the logarithmic score.

$$\mathrm{D}_{KL}(\mathbf{o}_t||\mathbf{f}_t) = \sum_{i=1}^{n} o_i \log\left(\frac{o_i}{f_i}\right) = o_{i\neq j}\log\left(\frac{o_{i\neq j}}{f_{i\neq j}}\right) + o_{i=j}\log\left(\frac{o_{i=j}}{f_{i=j}}\right) \tag{5.19}$$

Because $o_{i\neq j} = 0$ and $o_{i=j} = 1$, this reduces to

$$\mathrm{D}_{KL}(\mathbf{o}_t||\mathbf{f}_t) = 0\log\left(\frac{0}{f_{i\neq j}}\right) + 1\log\left(\frac{1}{f_{i=j}}\right) = -\log f_j = \mathrm{IGN} \tag{5.20}$$

This means that the divergence scoring rule presented in this chapter (DS*)* is actually equal to the Ignorance (IGN). The Ignorance is therefore not only "A scoring rule ... closely related to relative entropy" as defined by Roulston and Smith (2002), but actually also is a relative entropy. The difference is in the distributions that are used to calculate the relative entropy. Roulston and Smith (2002) refer to a relation to the divergence between the unknown "true" distribution $\mathbf{p}$ and the forecast distribution $\mathbf{f}$; see Eq. 5.21.

$$D_{KL}(\mathbf{p}||\mathbf{f}) = \underset{\mathbf{p}}{\mathrm{E}}[IGN] - H(\mathbf{p}) \tag{5.21}$$

The divergence that is used in the divergence score is calculated from the PMF after the observation ($\mathbf{o}$) instead of $\mathbf{p}$. That makes the second term of the RHS vanish and IGN equal to the divergence.

Using the decomposition presented in Eq. (5.13), the Ignorance score for a series of forecasts can now also be decomposed into a reliability, a resolution and an uncertainty component. This decomposition, until now only applied to the Brier score, has proven very useful to gain insight into the aspects of forecast quality. Furthermore the new interpretation of the Ignorance score as the average divergence of observation PMFs from forecast PMFs, links to results from information theory more straightforwardly.

5.3.3 Relation to information gain

Peirolo (2010) defines "information gain" as a skill score for probabilistic forecasts. The information gain is defined as

$$\mathrm{IG}_f \quad = \quad \log_2 \frac{f_j}{c_j} = \mathrm{IGN}_c - \mathrm{IGN}_f \tag{5.22}$$

where f_j denotes the probability attached to the event that actually occurred and c_j the climatological probability of that event. For a series of forecasts, the score is simply the average over the different timesteps.

$$\mathrm{IG}_f \quad = \quad \frac{1}{T} \sum_{t=1}^{T} \log_2 \frac{f_{k(t)}}{c_{k(t)}} \tag{5.23}$$

From this definition it becomes clear that the information gain, as defined by Peirolo (2010) as a positively oriented skill score, is equal to the reduction in uncertainty from climatic uncertainty to the remaining uncertainty after the forecast ($\mathrm{IG}_f = \mathrm{UNC} - \mathrm{DS}$). Alternatively, the information gain can be expressed as the correct information minus the wrong information ($\mathrm{IG}_f = \mathrm{RES} - \mathrm{REL}$).

5.4 Generalization to multi-category forecasts

5.4.1 Nominal category forecasts

When extending verification scores from forecasts of binary events to multi-category forecasts, it is important to differentiate between nominal and ordinal forecast categories. In

the case of nominal forecasts, there is basically one question that is relevant for assessing their quality: How well did the forecaster predict the category to which the outcome of the event belongs? In nominal category forecasts, there is no natural ordering in the categories into which the forecast event is classified. For this case of forecast verification, there is no difference between the categories in which the event did not fall. Although the probability attached to those events conveys information at the moment the forecast is received, the only probability relevant for verification, after the observation has been made, is the probability that the forecaster attached to the event that did occur. The quadratic score of Brier (1950) can also be used for multiple category forecasts. In that case, $\mathbf{f}_t$ and $\mathbf{o}_t$ are the PMFs of the event before and after the observation, now having more than two elements. The problem with this score is that it depends on how the forecast probabilities are spread out over the categories that did not occur. For nominal events this dependency is not desirable, as all probability attached to the events that did not occur is equally wrong.

The divergence score (DS) does not suffer from this problem, because it only depends on the forecast probability of the event that did occur; Eq. 5.20. The DS as presented in Eq. 5.4 can directly be applied on nominal category forecasts. A property of the score is that a high number of categories makes it more difficult to obtain a good score. To compare nominal category forecasts with different numbers of categories, the DS should be normalized to a skill score (DSS); see Eq. 5.14.

5.4.2 Ordinal category forecasts

When dealing with forecasts of events in ordinal categories, there is a natural ordering in the categories. This means that the cumulative distribution function (CDF) starts to become meaningful. There are now two possible questions that can be relevant for verification of the probabilistic forecast.

1. How well did the forecaster predict the category to which the outcome of the event belongs?
2. How well did the forecaster predict the exceedence probabilities of the thresholds that define the category boundaries?

The first question is equal to the one that is of interest for nominal forecasts. However, in the ordinal case there is a difference between the categories in which the observed event did not fall. Forecasts of categories close to the one observed are preferred over categories more distant from the one observed. Therefore, skill scores for ordinal category forecasts are often required to be "sensitive to distance" (Epstein, 1969; Laio and Tamea, 2007; Murphy, 1971, 1970). This requirement has led to the introduction of the ranked probability score (RPS) (Epstein, 1969), which is now widely used .The DS is not sensitive to distance in the sense of the RPS, because DS is insensitive to the forecasts for the non-occurring events. However, there still is an apparent sensitivity to distance introduced through the forecast PMF. A forecaster will usually attach more probability to the categories adjacent to the one that is considered most likely, simply because they are also likely. Therefore, missing the exact category of the observation with the most likely forecast still leads to a relatively

low penalty in the score, if the uncertainty estimation of the forecaster was correct and significant probability was forecast for the other likely (often neighboring) categories. Over a series of forecasts, the apparent distance-sensitivity of the penalty given by the DS is therefore defined by the PMF of the forecaster alone, and independent of what the categories represent. In verification literature, the property of only being dependent on the probability assigned to the event that actually occurred is known as locality, which is often seen as a non-desirable property of a score. Whether or not locality is desirable can be questioned (Mason, 2008). In this chapter it is argued that in absence of a context of the users of the forecast there is no justification for using non-local scores, which require some sensible distance measure to be specified apart from the natural distance sensitivity introduced by the forecast PMF. When assessing value or utility of forecasts, as opposed to quality, non-local scores can be used; see also the analogy with a horse race on page 117. However, in that case the distance measure should depend on the users and associated decision processes, as these determine the consequences of missing the true event by a certain number of categories distance. In non-local verification scores, the distance measure is often not explicitly specified in terms of utility, making it unclear what is actually measured. In those cases, the utility function of the users becomes more like an emerging property of the skill score instead of the other way around. Benedetti (2010) also presents locality as a basic requirement for a measure of forecast goodness, interpreting locality as "exclusive dependence on physical observations". He correctly states that it is a violation of scientific logic if two series of forecasts that assign the same probabilities to a series of observed events gain different scores, based on probabilities assigned to events that have never been observed. For a more elaborate treatment of this view on the fundamental discussion about locality, the reader is referred to Benedetti (2010) and Mason (2008).

The second question, regarding the forecast quality of exceedence probabilities, differs from the first, because all the thresholds are considered at once. Therefore, the quality of a single forecast depends on the entire PMF of the forecast and not only on the probability forecast for the event that occurs. Therefore, scores that are formulated for cumulative distributions can never be local. This means that apart from the physical observations, the importance attached to the events influences the score. So some assumption about value is added and the score is not a pure measure of quality alone. The RPS evaluates the sum of squared differences in CDF values between the forecast and the observation of the event (5.24)

$$\mathrm{RPS} = \frac{1}{n-1} \sum_{m=1}^{n-1} \left[\left(\sum_{k=1}^{m} f_k \right) - \left(\sum_{k=1}^{m} o_k \right) \right]^2 \tag{5.24}$$

The RPS can be seen as a summation of the binary Brier scores over all $n-1$ thresholds defined by the category boundaries. The summation implies that the Brier scores for all thresholds are weighted equally. Whether the BS for all thresholds should be considered equally important depends on the users. It has been shown that the RPS is a strictly proper scoring rule in case the cost-loss ratio is uniformly distributed over the users (Murphy, 1970). In that case the RPS is a linear function of the expected utility.

5.4.3 The Ranked divergence score

Now an information-theoretical score is presented for ordinal category forecasts, which are defined in terms of cumulative probabilities. An equivalent to the RPS would be the ranked divergence score (RDS), averaging of the DS over all $n-1$ category thresholds m.

$$\mathrm{RDS} = \frac{1}{n-1} \sum_{m=1}^{n-1} \mathrm{DS}_m, \tag{5.25}$$

with DS_m denoting the divergence score for the forecast of the binary event $j \leq m$. This assumes equally important thresholds. The RDS, just like the DS, is dependent on the climatological uncertainty. To make the score comparable between forecasts, the RDS can be converted into a skill score. Now, two intuitive options exist to do the normalization. The first is to normalize the individual DS_m scores for each threshold m to a skill score for that threshold, like (5.15), using the climatic uncertainty for the binary event defined by that threshold (5.26)

$$\mathrm{DSS}_m = 1 - \frac{\mathrm{DS}_m}{\mathrm{UNC}_m} \tag{5.26}$$

and then averaging the resulting skill score over all thresholds (5.27)

$$\mathrm{RDSS}_1 = \frac{1}{n-1} \sum_{m=1}^{n-1} \mathrm{DSS}_m \tag{5.27}$$

This means that the relative contributions to the reduction of climatic uncertainty about each threshold m are considered equally important. In other words, all skills of forecasts about the exceedence of the $n-1$ thresholds are equally weighted.

The second option for normalization is the first to sum the DS_m and then normalizing with the climatic score for the sum (5.28).

$$\mathrm{RDSS}_2 = 1 - \frac{\sum_{m=1}^{n-1} \mathrm{DS}_m}{\sum_{m=1}^{n-1} \mathrm{UNC}_m} \tag{5.28}$$

The formulation of the RDSS_2 according to Eq.(5.28) does not normalize the scores for the different thresholds individually, but applies the same normalization to every DS_m. This means that the merits of the forecaster for all thresholds are implicitly weighted according to the inherent uncertainties in the climate. In this way, the forecast of extreme (nearly certain) probabilities, which are often associated with extreme events, are hardly contributing to the total score, while they could in fact be most important for the users.

5.4.4 Relation to Ranked Mutual Information

An alternative skill score defined in terms of cumulative probabilities is the RMIS_O (5.16) as defined by Ahrens and Walser (2008). The version of the RMIS_O for multiple category

forecasts can be written as

$$\mathrm{RMIS}_O = 1 - \frac{\sum_{m=1}^{n-1} I\left(\mathbf{f}_m, \bar{\mathbf{o}}_m\right)}{\sum_{m=1}^{n-1} H(\bar{\mathbf{o}}_\mathbf{m})}, \tag{5.29}$$

with $\mathbf{f}_m$ denoting the series forecast probabilities of exceedence of threshold m, $\mathbf{o}_m$ denoting the corresponding series of observations and $\bar{\mathbf{o}}_m$ the average observed occurrence. For a perfectly reliable forecast, the RMIS_O is therefore equal to the RDSS_2 formulated in Eq.(5.28). For forecasts that are not well calibrated, the RMIS_O measures the amount of information that would be available after calibration, while the RDSS measures the information as presented by the forecaster. By using the decomposition presented in Eq. (5.15) it is possible to write

$$\mathrm{RDSS}_1 = \frac{1}{n-1}\sum_{m=1}^{n-1}\frac{\mathrm{RES}_m}{\mathrm{UNC}_m} - \frac{1}{n-1}\sum_{m=1}^{n-1}\frac{\mathrm{REL}_m}{\mathrm{UNC}_m} \tag{5.30}$$

By presenting both terms the resolution and the reliability term of Eq. (5.30) separately, both the potential information and the loss of information due to imperfect calibration are visible.

Apart from including the reliability or not, another question is how to weight the scores for the different thresholds, to come to one aggregated score. As every binary decision by some user with a certain cost-loss ratio can be associated with some threshold, the weighting reflects the importance of the forecast to the various users. No matter what aggregation method is chosen, there will always be an implicit assumption about the user's importance and stakes in a decision making process. This is inherent to summarizing forecast performance in one score. A diagram that plots the two skill score components against the thresholds contains the relevant information characteristics for different users. In this way each user can look up the score on the individual threshold, that is relevant for his decision problem, and compare it with the performance of some other forecasting system on that threshold.

5.4.5 Information and useful information

Forecasting is concerned with the transfer of information about the true outcome of uncertain future events that are important to a given specific user. The information in the forecast should reduce the uncertainty about the true outcome. It is important to note the difference between two estimations of this uncertainty. Firstly, there is the uncertainty a receiver of a forecast has about the truth, estimated in hindsight, knowing the observation. This uncertainty is measured by the divergence score. Secondly, there is the perceived uncertainty about the truth in the eyes of the user after adopting the forecast, which is measured by the entropy of the forecast. The first depends on the observation, while the second does not. Note that information-theoretical concepts measure information objectively, without considering its use. The usefulness of information is different for each specific user. The amount of useful information in a forecast can explicitly be subdivided into two elements:

1. reduction of uncertainty about the truth (the information theory part)
2. the usefulness of this uncertainty reduction (the user part)

The first element is only dependent on the user's probability estimate of the event's outcome before and after the forecast is received and on the true outcome. If the (subjective) probability distribution of the receiver does not change upon receiving the forecast, no information is transferred by it. If the probability distribution changed, but the divergence to the observation increased, the forecast increased the uncertainty about the truth as estimated from the observation, which is in itself an estimation of the unknown truth (although for the decomposition it was assumed to be a perfect estimation). A forecast is less informative to a user already having a good forecast. To make the information-theoretical part of useful information in a forecast independent of the user, remaining uncertainty is estimated instead of its reduction.

The second element of useful information in a forecast, usefulness, is user and problem specific. A forecast is useful if it is about a relevant subject. Communicating the exceedence probability of a certain threshold that is not a threshold for action for a specific user does not help him much. Usefulness also depends on how much importance is attached to events. This can be, for example, the costs associated with a certain outcome-action combination, typically reflected in a so-called payoff matrix. Implicitly, also information-theoretical scores make some assumption on the usefulness of events. The assumption is that the user attaches his own importance to the events by placing repeated proportional bets, each time reinvesting his money. This is referred to as Kelly-betting. For more detailed explanation; see Kelly (1956); Roulston and Smith (2002) and appendix C. In other words, the assumption is that the user maximizes the utility of his information in a fair game by strategically deciding on his importance or stakes.

The explicit consideration of usefulness of information brings up an interesting question about the roles of the forecaster and the user of forecasts. The divergence score measures the remaining uncertainty after adopting the forecast, which is completely independent of the user. This focuses the score on evaluating a main task of the forecaster, which is to give the best possible estimate of probabilities. It might also be argued however, that a forecaster should not just reduce uncertainty, but also deliver utility for user's decisions. To be able to judge forecasts on that criterion, assumptions need to be made about the users and their decision problems. When scores based on these objectives are used to improve forecasting procedures, maximizing these two objectives does not always lead to the same answer. In such cases an improvement of the utility of forecasts may coincide with a reduction in informativeness. Chapter 6 further looks into this issue, which is strongly related to model complexity, overfitting and calibration versus validation.

5.5 An example: rainfall forecasts in the Netherlands

As an illustration of the practical application of the divergence score and its decomposition, it was applied to a series of probabilistic rainfall forecasts and corresponding

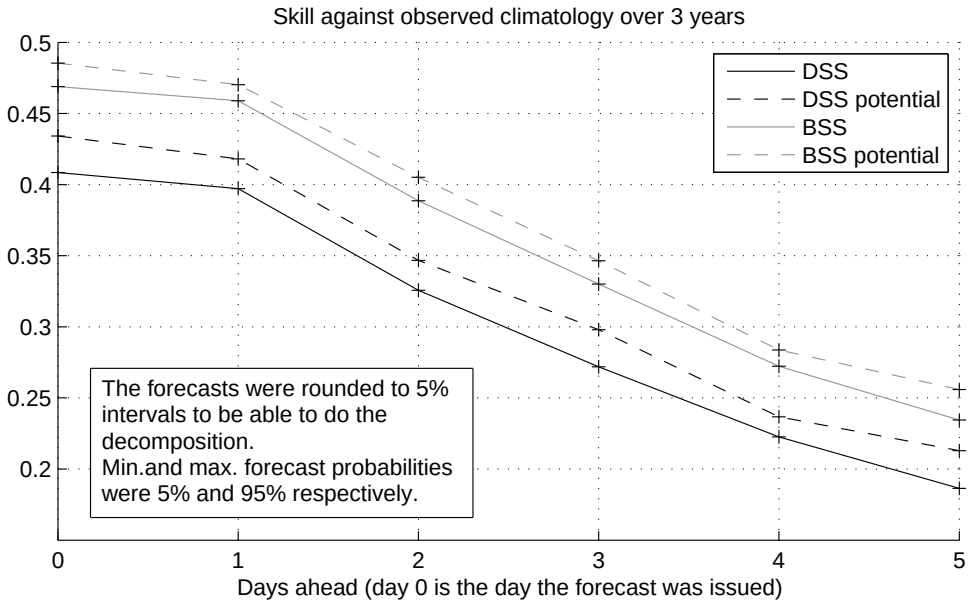


Figure 5.5: The skill according to both scores decreases with lead-time. The potential skills, which should be interpreted with caution, indicate the part of uncertainty that would be explained after recalibration.

observations for the measurement location De Bilt, the Netherlands. The forecast series consist of daily forecasts of the probability of a precipitation amount equal or larger than 0.3 mm, for zero to five days ahead. They are derived using output of both the ECMWF deterministic numerical weather prediction model and the ECMWF ensemble prediction model. Predictors from both models are used in a logistic regression to produce probability forecasts for precipitation. The data covers the period from Dec. 2005 to Nov. 2008, in which the average probability was 0.4613, leading to an initial uncertainty of 0.9957 bits.

Figure 5.5 confirms the expectation that both the Brier skill score and the divergence skill score show a decline with increasing lead time. It also shows that the forecasts possess skill over annual climatology up to a lead time of at least 5 days. The dashed lines show the potential skill that could be attainable after recalibration. The estimation of this potential skill, however, is dependent on the correct decomposition. The decompositions of both the Brier and the divergence score need enough data (large enough n_k) to calculate the conditional distributions $\mathbf{f}_k$ for all unique forecasts $k \in \{1 \dots K\}$. To be able to calculate the contribution to reliability of all the 99% forecasts, for example, at least 100 of such forecasts are necessary to not surely overestimate reliability. Also for a larger number of forecasts, there is a bias towards overestimation of the reliability, that decreases with the amount of data available per conditional to be estimated.

A solution for estimating the components with limited data is rounding the forecast probabilities to a limited set of values. In this way, less conditional distributions $\mathbf{f}_k$ need to be estimated and more data per distribution are available.

Figure 5.6 shows that for these three years of data, using finer grained probabilities as forecasts leads to an increasing overestimation of reliability. The skill scores themselves

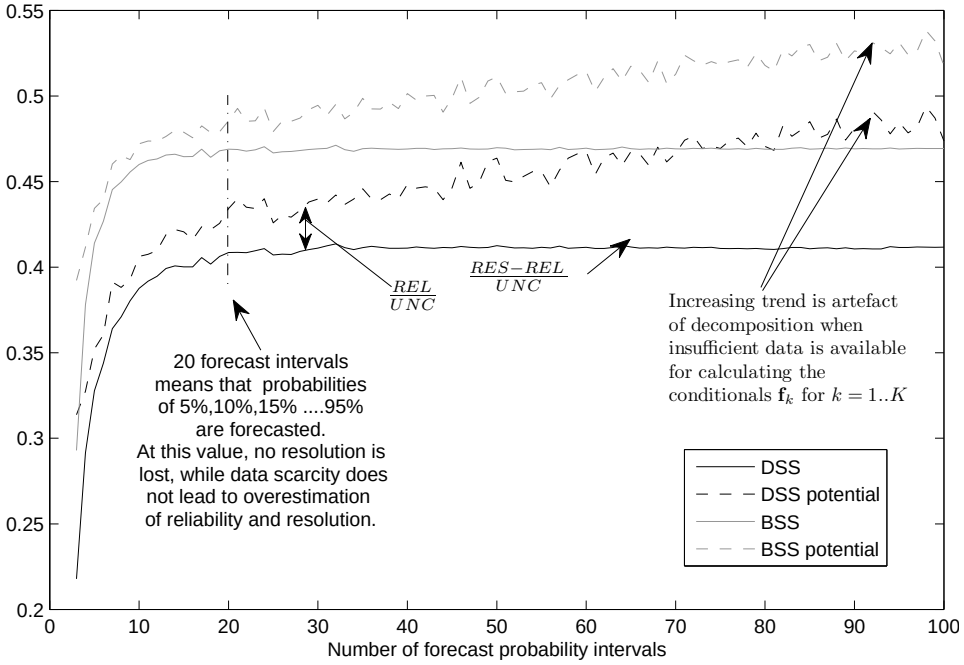


Figure 5.6: The calculation of reliability is sensitive to the rounding of the forecasts. If not enough data are available, rounding is necessary to estimate the conditionals. Too coarse rounding causes an information loss.

are not sensitive to this overestimation, because the lack of data causes a compensating overestimation of resolution. The potential skill, however, should be interpreted with caution, as solving reliability issues by calibration on the basis of biased estimates of reliability does not lead to a real increase in skill. From the figure it can also be noted that too coarse grained probabilities lead to a real loss of skill. In this case, giving the forecasts in 5% intervals seems the minimum needed to convey the full information that is in the raw forecast probabilities.

Fig. 5.7 sheds more light on the relation between the Brier skill score and the divergence skill score, based on 5 day ahead forecasts from a second data set, which covers February 2008 to December 2009. For this set, the forecast probabilities ranged from 1 to 99 %. The black dots indicate the scores that were attained for single forecast observation pairs. The dots show that the BSS and DSS have a monotonic relation as scoring rules. The limits of this relation are at $(1, 1)$ for perfect forecasts and, in this case, $(-\infty, -3.095)$ for certain, but wrong forecasts. The worst forecast was 98%, while no rain fell.

The total scores for different weeks of forecasts are plotted as gray dots. They are averages of sets of 7 black dots. Because the relation of the single forecast scores is not a straight line, a scatter occurs in the relation of the weekly average scores, which is therefore no longer monotonic. The scatter implies that two series of forecasts can be ranked differently by the Brier score and the divergence score. In this example, the scatter is relatively small ($r^2 = 0.9938$) and will probably have no significant implications, but it would be larger if many overconfident forecasts were issued. An interesting example of differently ranked

forecast series are the two weeks indicated by the triangles, where the scores disagree on which of the two weeks was better forecast than climate. The downward pointing triangle marks the score for forecasts in week A, where performance according to the divergence score was worse than the climatological forecast (DSS=-0.0758), but according to the brier score was slightly better than climate (BSS=0.0230). Conversely, the upward pointing triangle marks week B, where the forecasts according to Brier were worse than climate (=-0.0355), but still contained more information than climate according to the DSS (=0.0066).

Given that the scatter in the practical example is small, the Brier score appears to be a reasonable approximation of the divergence score and is useful to get an idea about the quality of forecasts. More practical comparisons are needed to determine if the approximation can lead to significantly different results in practice. These are mostly expected in case extreme probabilities are forecast.

The severe penalty the divergence gives for errors in the extreme probabilities, which is sometimes seen as a weakness[2], should actually be viewed as one of its strengths. As the saying goes: "you have to be cruel to be kind". It is constructive to give an infinite penalty to a forecaster that issues a wrong forecast that was supposed to be certain. This is fair, because the value that a user would be willing to risk when trusting such a forecast is also infinite.

5.6 Generalization to uncertain observations

The previous sections introduced an information-theoretical decomposition of Kullback-Leibler divergence into uncertainty, reliability and resolution. In this section, this decomposition is generalized to the case where the observation is uncertain. Along with a modified decomposition of the divergence score, a second measure is presented, the cross-entropy score, which measures the estimated information loss with respect to the truth instead of relative to the uncertain observations. The difference between the two scores is equal to the average observational uncertainty and vanishes when observations are assumed to be perfect.

5.6.1 Introduction

Section 5.2.4 introduced a decomposition of the divergence score analogous to the decomposition of the Brier score. See also Bröcker (2009) for a general form of this decomposition

2. Note for example that Selten (1998), the 1994 laureate of the Nobel prize in economics, takes it as an axiom that scores should not exhibit infinite penalties, even when zero probability is attached to the true outcome. Selten regards this "hypersensitivity" as an unacceptable value judgment and together with other axioms, he gives an axiomatic justification for using the Brier score, which treats such errors more mildly. It is left to the reader to speculate what the economic consequences may be of wrongly ruling out improbable events.

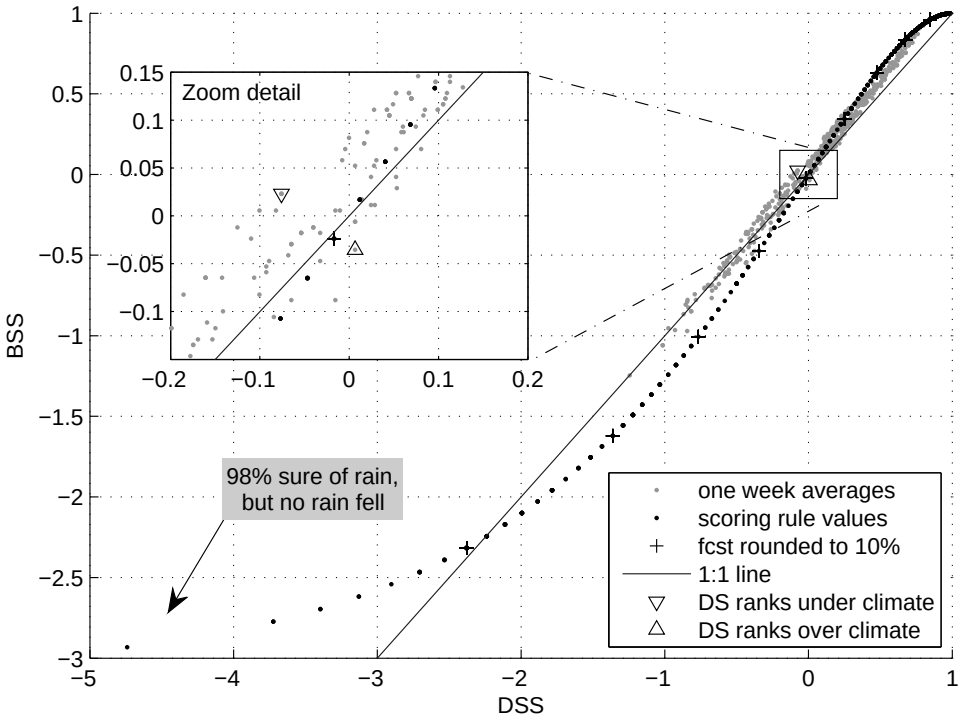


Figure 5.7: Relation between the Brier skill score and the divergence skill score. For single forecasts they have a monotonic relation, but for averages of series of forecasts, a scatter in the relation can cause weekly average forecast skill to be ranked differently by both scores. There is only little difference in the overall number ranked better than climatology. For this example, the Brier score ranked 79.2% and the divergence score ranked 79.4% of weekly average skills better than climate.

for proper scoring rules. A possible interpretation of the new information-theoretical decomposition is "The remaining uncertainty is equal to the missing information minus the correct information plus the wrong information" (see figure 5.8). A forecast that is reliable but has no perfect resolution does not give complete information, but the information it does give is correct. In the decompositions of both the Brier score (BS) and the divergence score (DS), it was assumed that the observations are certain and correspond to the truth.

In reality, no observation can be assumed to correspond to the true outcome with certainty. For example, in the evaluation of binary probabilistic precipitation forecasts of the Dutch Royal Meteorological Institute (KNMI), the observation that corresponds to "no precipitation" is defined as an observed precipitation of less than 0.3 mm on a given day. Given the observational errors in the exact precipitation amount, measured values close to the threshold would be best represented by a probabilistic observation, accounting for the uncertainties; see fig. 5.9. Briggs et al. (2005) noted that uncertainty in the observation must be taken into account to assess the true skill of forecasts. This requires either "gold standard" observations or subjective estimates of the observation errors. Bröcker and Smith (2007) proposed to use a noise model for the observation to transform the forecast, using this to define a generalized score.

In this section, it is proposed to define an "uncertain observation" as the conditional distribution of the true outcome of event or quantity that was forecast, given the reading on one or more measurement instruments. For example, when the spatial scale or location of the measurements and the forecasts differs, the distribution can be based on spatial statistics of various instruments. In another case, the distribution may be derived from an model of the observational noise (e.g. due to wind around a raingauge, noise in the electronics). Note that the correctness of such uncertainty models cannot be verified, because the "true" value cannot be observed directly. Although the term verification suggests a comparison between forecasts and truth (Latin: veritas = truth), both the divergence and the Brier score are actually comparing the forecasts with observations, which are an estimate of the unknown truth. An uncertain observation acknowledges this by representing the uncertainty explicitly by a probabilistic best estimate.

When the uncertainty in the observations is accounted for by representing them with probability distributions with nonzero entropy (i.e. the observation assigns probability to more than one outcome), the decomposition in section 5.2.4 does not hold. The divergence score as a whole, however, is still a useful measure of correspondence between forecasts and observations. It would therefore be interesting to define a meaningful decomposition of the divergence score that is applicable in the case of uncertain observations. A second point is whether the quality of forecasts should be measured with respect to the known probabilistic observations or estimated with respect to the unknown truth.

In this section, a modification to the decomposition in section 5.2.4 and Weijs et al. (2010b) is presented, which generalizes to the case of uncertain observations. The new decomposition is interpreted in terms of uncertainty and information. Furthermore a second, related measure for forecast quality is presented. In information theory, this measure often referred to as cross-entropy, which in this case estimates the uncertainty relative to the truth instead of relative to the observation. A decomposition for this score is also presented. The scores are applied to a real data set for illustration.

5.6.2 Decomposition of the divergence score for uncertain observations

The decomposition of the divergence score (DS) for a series of N forecasts that was presented in Eq. 5.13 on page 93 relies on the assumption that observations are certain, i.e. $\mathbf{o_t} = (0,1)^T$ or $\mathbf{o_t} = (1,0)^T$. In appendix B this assumption is used to rewrite the closing term of the decomposition $\frac{1}{N}\sum_{t=1}^{N} \mathrm{D}_{KL}(\mathbf{o}_t||\bar{\mathbf{o}})$ as the uncertainty component $\mathrm{H}(\bar{\mathbf{o}})$. The uncertain observation $\mathbf{o_t}$ is the probability mass function (PMF) of the true outcome of the uncertain event that is forecast, given the available information after it occurred. Because measurements are usually indirect, the observation can be regarded as a (usually subjective) conditional distribution of the true outcome, given the information from the measurement equipment. When we want the decomposition to be valid for uncertain observations, the last step of the derivation in appendix B can be omitted. The uncertainty component, the last term in Eq. 5.13, is thus replaced by the average Kullback-Leibler

divergence from the uncertain observations to the average observation (i.e. the observed climatic distribution), the last term in Eq. 5.31.

$$\mathrm{DS} = \frac{1}{N}\sum_{k=1}^{K} n_k \mathrm{D}_{KL}(\bar{\mathbf{o}}_k||\mathbf{f}_k) - \frac{1}{N}\sum_{k=1}^{K} n_k \mathrm{D}_{KL}(\bar{\mathbf{o}}_k||\bar{\mathbf{o}}) + \frac{1}{N}\sum_{t=1}^{N} \mathrm{D}_{KL}(\mathbf{o}_t||\bar{\mathbf{o}}) \tag{5.31}$$

The last term in Eq. 5.31 represents the expected climatological uncertainty relative to the observation, which is depicted in figure 5.8 as $\mathrm{UNC}_{\mathrm{DS}}$. By writing the uncertainty term of the divergence score decomposition in this way, it remains valid for uncertain observations. The original uncertainty term, the entropy $\mathrm{H}(\bar{\mathbf{o}})$, can be seen as representing the estimated climatological uncertainty relative to the truth, which from now on will be denoted as $\mathrm{UNC}_{\mathrm{XES}}$, because it is part of a decomposition of XES, which will be introduced in section 5.6.3. Likewise, DS represents the average remaining uncertainty relative to the observations, which in the case of uncertain observations can become different from the estimated remaining uncertainty relative to the truth.

Analogy for the Brier score decomposition

Analogously to the new decomposition of the DS, the Brier score decomposition introduced by Murphy (1973) can be modified in a similar manner to remain valid for uncertain observations. This can be achieved by replacing the uncertainty term in the original decomposition by the average squared Euclidean distance from the observations to the average observation. The modified decomposition is shown in equation 5.32. For perfect observations, equations 5.7 and 5.32 are the same. When observational uncertainty is considered, Eq.5.7 does not hold but Eq. 5.32 does.

$$\mathrm{BS} = \frac{1}{N}\sum_{k=1}^{K} n_k(\mathbf{f_k} - \bar{\mathbf{o}}_\mathbf{k})^2 - \frac{1}{N}\sum_{k=1}^{K} n_k(\bar{\mathbf{o}}_\mathbf{k} - \bar{\mathbf{o}})^2 + \frac{1}{N}\sum_{t=1}^{N} (\mathbf{o}_t - \bar{\mathbf{o}})^2 \tag{5.32}$$

5.6.3 Expected remaining uncertainty about the truth: the cross-entropy score

Now, the cross-entropy score (XES) will be introduced. The expected uncertainty relative to the unknown truth can be expressed by taking the expectation, with respect to the PMF that represents the uncertain observation, of the Kullback-Leibler divergence from the hypothetical truth to the forecast distribution.

$$\mathrm{XES} = \frac{1}{N}\sum_{t=1}^{N} E_{\mathbf{o}_t}\mathrm{D}_{KL}(\mathbf{v}_t||\mathbf{f}_t) = \frac{1}{N}\sum_{t=1}^{N}\sum_{j=1}^{n}\sum_{i=1}^{n}\left\{[\mathbf{o}_t]_j\,[\mathbf{v}_t]_i \log\frac{[\mathbf{v}_t]_i}{[\mathbf{f}_t]_i}\right\} \tag{5.33}$$

In which $n = 2$ is the number of categories in which the event can fall. $\mathbf{v}_t$ denotes the hypothetical distribution of the truth at instance t, which, like a perfect observation, is either $(1,0)^T$ if the event in fact did not occur or $(0,1)^T$ if the event truly occurred. $E_{\mathbf{o}_t}$ is the expectation operator with respect to the probability distribution $\mathbf{o}_t$. In this case, the

Kullback-Leibler divergence $\mathrm{D}_{KL}(\mathbf{v}_t||\mathbf{f}_t)$ reduces to the logarithmic score (Good, 1952), which is also known as the Ignorance score (Roulston and Smith, 2002). These scores are simply minus the logarithm of the probability attached to the event that truly occurred.

$$\mathrm{D}_{KL}(\mathbf{v}_t||\mathbf{f}_t) = -\log\left[\mathbf{f}_t\right]_{k(t)} \tag{5.34}$$

Where $k(t)$ is the category in which the true outcome of the event fell at instance t. Because $\mathbf{o}_t$ is the best estimate of the unknown true outcome, we can use the expectation $E_{\mathbf{o}_t}\mathrm{D}_{KL}(\mathbf{v}_t||\mathbf{f}_t)$ to evaluate the forecast, which can also be written as the right hand expression in Eq. 5.35. In information theory, this expression is often defined as the cross-entropy between $\mathbf{o}_t$ and $\mathbf{f}_t$, hence it is referred to in this thesis as the cross-entropy score XES_t

$$\mathrm{XES}_t = E_{\mathbf{o}_t}\mathrm{D}_{KL}(\mathbf{v}_t||\mathbf{f}_t) = -\sum_{i=1}^{n}\left[\mathbf{o}_t\right]_i \log\left[\mathbf{f}_t\right]_i \tag{5.35}$$

This measure can be interpreted as the expected remaining uncertainty relative to the truth, when the forecasts are evaluated in the light of the observations, which are assumed to be reliable probability estimates of the truth. The difference with the divergence score becomes clear from figure 5.8. For a series of forecasts, the cross-entropy score is defined as $\mathrm{XES} = \frac{1}{N}\sum_{t=1}^{N}\mathrm{XES}_t$.

Decomposition of cross-entropy

Figure 5.8 shows that the relation between all the components allows for several decompositions. The relation between DS and XES can be written as

$$\mathrm{XES} = \mathrm{DS} + \frac{1}{N}\sum_{t=1}^{N}\mathrm{H}(\mathbf{o}_t) \tag{5.36}$$

The estimated remaining uncertainty in the forecasts relative to the truth (XES) is equal to the average uncertainty relative to the observations (DS) plus the average uncertainty that the observations represent, relative to the estimated truth (the second term on the right hand side of Eq.5.36).

Another natural decomposition for the XES is the original decomposition of DS for perfect observations as presented in Weijs et al. (2010b). For uncertain observations, the three components presented there add up to the XES instead of to the DS; see also figure 5.8. The decomposition of the cross-entropy score (XES) therefore reads

$$\mathrm{XES} = \mathrm{REL}_{\mathrm{XES}} - \mathrm{RES}_{\mathrm{XES}} + \mathrm{UNC}_{\mathrm{XES}} \tag{5.37}$$

$$-\frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{n}\left[\mathbf{o}_t\right]_i \log\left[\mathbf{f}_t\right]_i = \frac{1}{N}\sum_{k=1}^{K} n_k \mathrm{D}_{KL}(\bar{\mathbf{o}}_k||\mathbf{f}_k) - \frac{1}{N}\sum_{k=1}^{K} n_k \mathrm{D}_{KL}(\bar{\mathbf{o}}_k||\bar{\mathbf{o}}) + \mathrm{H}\left(\bar{\mathbf{o}}\right) \tag{5.38}$$

Note that the resolution and reliability components are equal to those of the DS decomposition in Eq. 5.31.

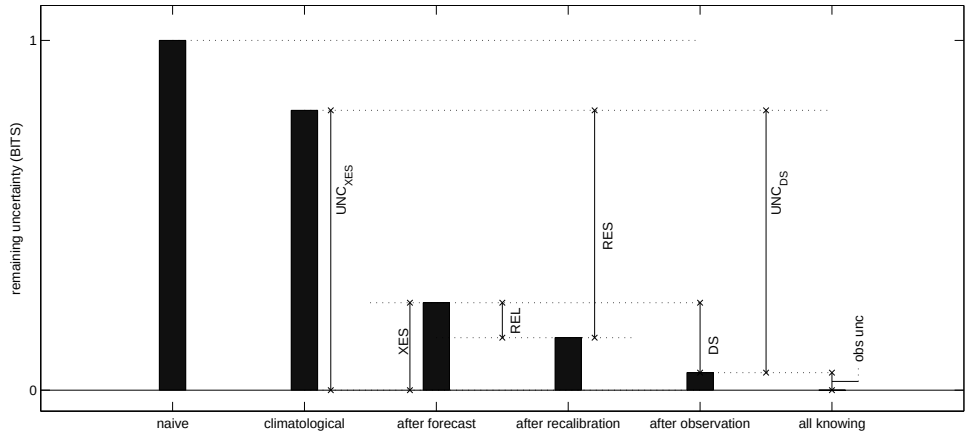


Figure 5.8: The relations between the components and the scores presented in this chapter are additive. The bars give the average remaining uncertainty about the truth (measured in bits) for various (hypothetical) stages in the forecasting process. The naive forecast is always assigning 50% probability of precipitation (complete absence of information), the climatological forecast takes into account observed frequencies. This climatological uncertainty ($\mathrm{UNC_{XES}}$) can be reduced to (XES) by believing the forecasts $\mathbf{f}$. If these are not completely reliable, the uncertainty can be further reduced with REL by recalibration. After observation, there is still some uncertainty ('*obs unc*') about the hypothetical true outcome, given that observations are not perfect. Only for an all knowing observer, the uncertainty is reduced to 0. The resolution (RES) is the information that could maximally be extracted from the forecasts by perfect calibration. The divergence score (DS) and the new uncertainty component ($\mathrm{UNC_{DS}}$) measure the uncertainty after forecast and in the climate, relative to the observations.

5.6.4 Example application

As an illustration of the new term in the decomposition, the scores were calculated for a real data set of binary probabilistic rainfall forecasts of the Dutch Royal Meteorological Institute (KNMI). The observed rainfall amounts were transformed into probabilistic uncertain observations using a very simple uncertainty model. The purpose of this exercise is merely to illustrate the concepts in this section. The forecasts that are evaluated are the forecast probabilities of a daily precipitation of 0.3 mm or more. This is the same dataset that was used in Weijs et al. (2010b). In that paper the rainfall amounts x_t, which were given with a precision of 0.1 mm, were converted to binary observations with a simple threshold filter: if $x_t \geq 0.3 \Rightarrow \mathbf{o}_t = (0, 1)$, if $x < 0.3 \Rightarrow \mathbf{o}_t = (1, 0)$. In this section, a random measurement error is assumed to make $\mathbf{o}_t$ probabilistic and account for the uncertainty in the observation.

The model of the uncertainty in the observation is Gaussian. The observed rainfall amount becomes a random variable with a normal probability density function

$$g_{obs}(x) = N(\mu_{obs}, \sigma_{obs}) \tag{5.39}$$

with mean μ_{obs} and standard deviation σ_{obs}. Because in this case we deal with a binary predictand, the pdf of the observation can be converted to a binary probability mass

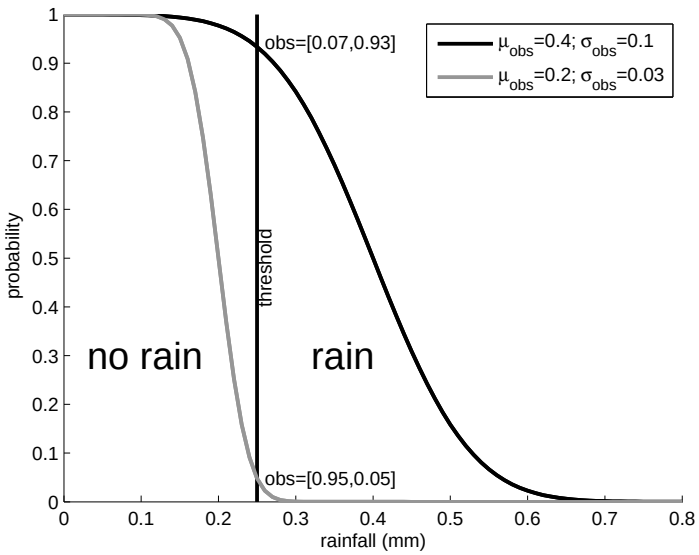


Figure 5.9: The measurement uncertainty in the precipitation measurement leads to a probabilistic binary observation. In the example, a simple Gaussian measurement uncertainty is assumed. The measurement distributions are centered around the measurements and have a constant standard deviation.

function $\mathbf{o}_t = (1 - o_t, o_t)$ by using

$$o_t = 1 - G_{obs}(T) = 1 - \int_{-\infty}^{T} g_{obs}(x)dx \tag{5.40}$$

in which $G_{obs}(T)$ is the cumulative distribution function of the observation, evaluated at threshold T. This conversion is illustrated in figure 5.9.

The decompositions of the the DS and XES scores for the entire dataset were calculated for a range of different standard deviations σ_{obs} for the measurement uncertainty. In figure 5.10 it can be seen that while the average observation uncertainty grows, the divergence score improves (decreases), but the cross entropy score (XES) deteriorates. This indicates that if the observation uncertainty is reflected by this model, the best estimate for the information loss compared to the truth is higher than would be assumed neglecting observation uncertainty. Not taking the observation uncertainty in account would lead to an overestimation of the forecast quality in this specific case. A closer analysis reveals that most of the deterioration is caused by observations of 0 mm, which start to give significant probability to rain when the standard deviation grows beyond 0.1 mm. As many of the 0 mm observations are during cloudless days and in fact almost certain, we might reconsider the simple Gaussian uncertainty model .

When the uncertainty model is changed to have no uncertainty for 0 mm observations, the decomposition changes significantly (see dashed lines in fig. 5.10), leading to a cross-entropy score that is almost constant (sometimes even slightly decreasing) with increasing standard deviation. For this particular error model, the uncertainty in the observations thus hardly affects the estimation of the forecast quality. This gives us confidence that as

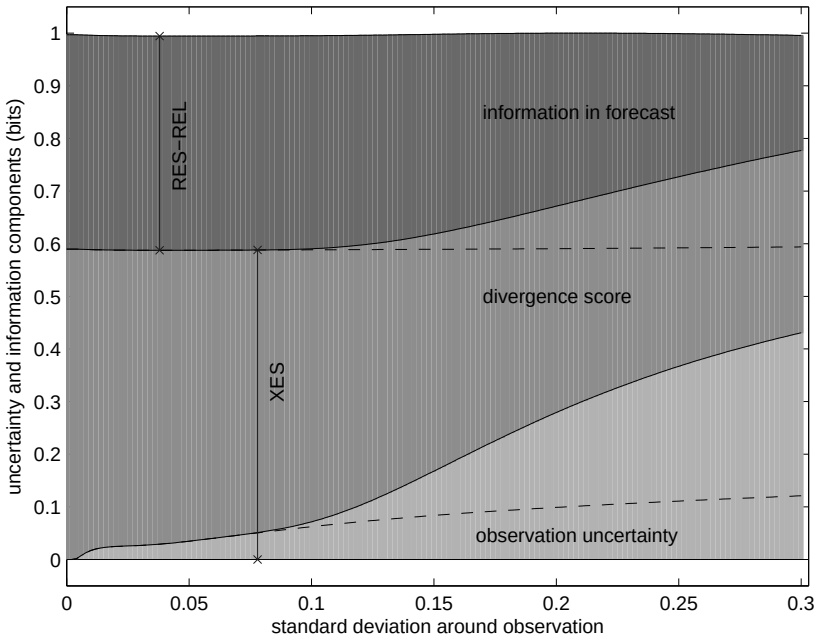


Figure 5.10: The resulting decompositions (reliability not shown) as a function of measurement standard deviation. The growing XES with standard deviation, indicates that for the homoscedastic Gaussian observation uncertainty model, forecast quality is lower than would be estimated assuming perfect observations. The dashed lines show the decomposition for the same observation uncertainty model, with the exception that measurements of 0 mm are assumed to be certain. The almost constant XES in that case indicates that the estimation of forecast quality is robust against those observation uncertainties.

long as the 0 mm observations are certain, the estimate of forecast quality is robust against Gaussian observation errors with standard deviation up to 0.3 mm. Although in that case there is significant observation uncertainty (lower dashed line, fig. 5.10) that lowers the divergence score, the changes in the cross entropy score for individual forecasts cancel each other out. Not surprisingly, the robustness of forecast quality estimates depends very much on the characteristics of the observation uncertainty. Further experiments are necessary to determine how to formulate realistic observation uncertainty models and how this can benefit verification practice.

5.6.5 Discussion: divergence vs. cross-entropy

For the divergence score (DS), worse observations lead to better scores for forecast quality, because the quality is evaluated relative to the observations. This might be considered undesirable, especially when the performances for two locations with similar climates are compared, while the quality of the observations is not the same. On the other hand, the divergence score has the advantage of not making explicit reference to a truth beyond the observations, which might be philosophically more appealing.

If the cross-entropy score (XES) is used as a scoring rule, the score estimates the quality of the forecasts at reducing uncertainty about the truth. This quality may be estimated differently in the light of observation uncertainty, but should not be relative to it. The skill might both be overestimated and underestimated in the presence of observation uncertainty. This depends on the nature of the errors, which should there be modeled to the best possible extent. The XES allows a better comparison between the quality of different forecasts. In other words, the benchmark to compare the forecasts to is the truth. Because the uncertainty of the forecasts relative to this benchmark can only be evaluated if we would know the truth, we can only estimate its expected value. In contrast, in the divergence score (DS), the benchmark to which the forecasts are compared are the probabilistic estimations of the truth. The remaining uncertainty with respect to these estimates, the observations, can be calculated exactly. Summarizing, the divergence score is the exact divergence from an estimate of the truth (the observation), while the cross-entropy score is an estimated (expected) divergence from the exact truth.

5.7 Deterministic forecasts cannot be evaluated

Can a forecaster be completely sure about something that in the end does not happen and still get credit for his forecast? This does not appear natural, but it often turns out to happen in practice. For example, a deterministic flow forecast of $200\,\mathrm{m}^3/\mathrm{s}$ is considered quite good, when $210\,\mathrm{m}^3/\mathrm{s}$ is observed. Apparently, it is already expected that some error will occur and a forecast that is $10\,\mathrm{m}^3/\mathrm{s}$ off is considered to be not that bad. Hydrological models are per definition simplifications of reality. Often, they describe relations between macrostates, like averaged rainfall, mass of water in the groundwater reservoir, and flow through a river cross-section. Similar to problems in statistical thermodynamics, having limited information about what really goes on inside a hydrological system on a microscopical level, our forecasts on a macroscopical level can never be perfect (Weijs, 2009; Grandy Jr, 2008). What can be said about the real world on the basis of a model is therefore inherently erroneous to some extent, or should be stated in terms of probabilities.

How then, should deterministic forecasts be evaluated? Literally taken, a deterministic (point value) forecast states: "the outcome is x". Implicitly, such a forecast asks to be evaluated from a black and white view: the forecast is either wrong or right. The divergence score also reflects this. If the forecast is right, the perfect score of 0 will be attained. If the forecast is wrong, however, a penalty of infinity will be given. If one such a forecast is given, the forecaster can look for another career , because even a future series of perfect forecasts can not average out the infinite penalty. The decomposition shows that the reliability component is responsible; See Fig. 5.4. Although the deterministic forecasts usually contain information about the observed outcomes, given that the resolution (correct information) is positive and removes some of the uncertainty, this is completely annihilated by the reliability term (the wrong information). The discrepancy between the information (reduction of uncertainty) that the forecasts contain and the information conveyed by the messages that constitute the forecasts is so large that the expected surprise about the

truth of a person taking the forecast at face value goes to infinity. The fact that deterministic forecasts are still used in society (and unfortunately sometimes even preferred), while they explode uncertainty to infinity, seems to present a paradox. In this section, two possible interpretations are proposed that offer a solution to this paradox.

5.7.1 Deterministic forecasts are implicitly probabilistic (information interpretation)

Fortunately, in practice, almost no person takes deterministic forecasts at face value. The fact that a user does not take the forecast literally can be seen as recalibration of the forecast ("unconscious statistical post-processing") by that user. The user bases his internal probability estimates on the forecast, but adjusts the probabilities given by the forecaster based on his own judgment, instead of literally copying the forecaster's statements. For a deterministic forecast, this means reallocating some probability to outcomes that the forecast did not speak about. This reallocation improves the reliability of the internal probability estimates of the user on which he bases his actions. We can thus see this as the user eliminating the wrong information from the forecast.[3] The user can do the recalibration based on previous experience with the forecasts, common sense and can also add information from his own observations. The user of the forecast can think "if the forecaster says the water level will be 10 cm under the embankment, he implicitly also forecasts a little that overtopping will occur". Note that the example of Grand Forks in (Krzysztofowicz, 2001) shows that not all users do this. Mathematically this recalibration is equivalent to also attaching some probability to overtopping. However, it is not the task of a user to guess what the forecaster wanted to say. The forecaster has the task of summarizing different sources of information and expert knowledge into a forecast that various users can base their decisions on. Consistency requires that the forecaster communicates his judgments to the user (Murphy, 1993). If he deems it possible that 210 m^3/s will flow through the river instead of his best estimate 200 m^3/s, then the forecaster should also communicate a probability for this outcome to the user.

The forecaster may also present the deterministic forecast as being an expected value or mean. This suggests an underlying probabilistic forecast. However, when taking the information-theoretical viewpoint, communicating an expected value means nothing without additional statements regarding the probability distribution. The principle of maximum entropy (PME) (Jaynes, 1957) states that when making inferences based on incomplete information, the best estimate for the probabilities is the distribution that is consistent with all information, but maximizes uncertainty. In this way, the uncertainty is reduced exactly by the amount the information permits, but not more. The resulting distribution thus gives an exact representation of the information actually conveyed by the forecast. Maximizing entropy with known mean and variance, gives a Gaussian distribution, maximizing uncertainty about the velocities of gas molecules with known

3. Note that eliminating wrong information is different from adding information. If a user takes the forecasts as true, but partial information and is rational (following Bayesian probability logic), no future information can update the zero probability. This is another argument against assigning zero probability to anything.

total kinetic energy gives the Maxwell-Boltzmann distribution (Jaynes, 2003; Cover and Thomas, 2006). When PME is applied to expected value forecasts, however, the maximum entropy forecast distribution that is consistent with the information given by the forecaster is uniform between minus and plus infinity. It is the complete opposite end of the spectrum compared to the previous literal interpretation of the deterministic forecast: from claiming total certainty to claiming total uncertainty.

In the case of streamflow forecasts, the user can still get a less nonsensical forecast distribution by combining the information in the forecast with the common sense notion that streamflows in rivers are nonnegative. This extra constraint turns the PME forecast distribution for a known expected value into an exponential distribution (Cover and Thomas, 2006).

This brings back the question who ought to specify these constraints, which constitute information. The fact that the user can reduce the maximum entropy by adding this common sense constraint actually means that the forecaster failed to add this information. Note that the forecaster should be best equipped to give probability estimates and these should be summarized in such a way that no information is lost, but also all uncertainty is represented (cf. consistency).

As was argued in the introduction, predictions only make sense when they are testable, i.e. can be evaluated. One way to evaluate deterministic forecasts with information measures is to convert them to probabilistic forecasts by looking at the joint distribution of forecasts and observations. The conditional distributions of observations for each forecast value can then be seen as probabilistic forecast distributions. It is important to note however, that the probabilistic part of such a forecast is derived from data that includes the observations. Such a forecast is thus evaluated against the same data that is used as the basis of its own uncertainty model, which is clearly undesirable.

Also without explicit conversion to a probabilistic forecast, the uncertainty model becomes explicit when a series of deterministic forecasts is evaluated. A penalty (objective) function for a deterministic forecast can be interpreted as an uncertainty (information) measure for a corresponding probabilistic forecast. For example, a deterministic forecast evaluated with root mean squared error implicitly defines a Gaussian forecast probability density function. An important consequence of this insight is that the way to evaluate a deterministic model actually defines (i.e. forms) the probabilistic part of a total model, consisting of a separate deterministic and probabilistic part. The objective function (which is a likelihood measure) should therefore be stated a priori, as it forms part of the model that is put to the test against observations.

While estimating the error model from the data may under some conditions be acceptable in calibration, for (independent) evaluation of forecasts it is unacceptable, because it uses the data against which it is evaluated. A correct approach would be to explicitly formulate a parametric error model, and find its parameters in the calibration. The combination of the hydrological model and the error model can subsequently be used to make probabilistic predictions, which can be evaluated with the divergence score in an independent evaluation

period. The error models are not restricted to Gaussian distributions, but can take more flexible forms. Such an approach is taken in Schoups and Vrugt (2010).

As a last consideration, it must be stressed that even if an error model is properly formulated and added to the deterministic "physical" part, the resulting model still represents a false dichotomy between true behavior of the system and the error, as was argued by Koutsoyiannis (2010). A more consistent approach would be to explicitly make the probabilistic part of the model an integrated part of the physical reality it is supposed to simplify. Such approaches can lie in studying the time-evolution of chaotic systems (Koutsoyiannis, 2010) or in applying the principle of maximum entropy in combination with macroscopic constraints, as for example suggested by Weijs (2009) and Koutsoyiannis (2005a).

Concluding, from the information-theoretical viewpoint, several reasons come to light why deterministic forecasts should in fact be considered to be implicitly probabilistic. The problem with these forecasts is that they leave too much of the probabilistic interpretation to the user. It might be considered ironic that the users who are claimed not to be able to handle probabilistic forecasts and are for that reason provided with deterministic forecasts are the ones who have to rely most on their ability to subconsciously make probability estimates based on the limited information in the deterministic forecast.

5.7.2 Deterministic forecasts can still have value for decisions (utility interpretation)

A second, independent interpretation of deterministic forecasts that justifies their existence is their usefulness, even to users who do not make subconscious probability estimates. Even though a reservoir operator might be infinitely surprised if he has taken a deterministic inflow forecast of $200\,\mathrm{m}^3/\mathrm{s}$ at face-value and he finds out the inflow was $210\,\mathrm{m}^3/\mathrm{s}$, his loss is not infinite. The operator might spill some water, but all is not lost.

The difference between surprise and loss is due to the fact that most decision problems are not equal to placing stakes in a series of horse races. Such a horse race is the classical example where information can be directly related to utility, see Kelly (1956) and Cover and Thomas (2006) for more explanation. Kelly showed that when betting on a series of horse races, where the accumulated winnings can be reinvested in the next bet, the stakes the gambler should put on each horse should be proportional to the estimated winning probabilities. In a single instance of such a horse race, all money not bet on the winning horse is lost, so the only probability that is important for the results is the one attached to the winning horse. If zero probability (and thus no bets) were put on the winning horse, then the gambler loses all his capital and has no chance of future winnings. In contrast, for decision problems like reservoir operation, an operator blindly believing in an inflow into his reservoir of $200\,\mathrm{m}^3/\mathrm{s}$ and optimally preparing only for that flow, will automatically also be quite well prepared for $210\,\mathrm{m}^3/\mathrm{s}$. Conversely, the preparation on a predicted event, which influences the utility of an outcome, may depend on the entire forecast distribution and not just on the probability of the event that materializes. This makes the loss function non-local (locality is discussed in sections 5.4.2 and 6.6.2).

Another difference with the horse race is that the total amount of value at stake in hydrological decision making usually does not depend on the previous gains, while the results for the horse race assume that the gambler invests all his previously accumulated capital in the bets. The gambler therefore wants to maximize the product of rates of return over the whole series of bets, while for a reservoir operator, each period offers a new opportunity to gain something from the water, even in case he spilled all his stored water in the previous month. The reservoir operator is interested in the total sum of gains. This is comparable with a gambler whose spouse allows him/her to bet a fixed amount of money each week (Kelly, 1956) and then spends it all in the bar on the same evening without possibility of reinvesting in the next bet. Assuming a utility that increases linearly with the consumption of beer bought with the winnings, the best decision is to bet all money on the one horse with the best expected return[4]. Again, one loss is not fatal for the whole series of bets. The gambler can still hope for better luck next week. The evaluation of the value of deterministic forecasts is therefore not as black and white as evaluation of the information they contain.

The evaluation of deterministic forecasts in this interpretation is thus connected to a decision problem. Decisions can be taken as if the forecasts are really certain, and still be of positive value, although generally less value than probabilistic forecasts; see e.g. Philbrick and Kitanidis (1999); Krzysztofowicz (2001); Pianosi and Ravazzani (2010); Zhao et al. (2011). The loss functions for evaluating deterministic forecasts can be seen as functions that map the discrepancy between forecast value and observed value to a loss of the decision based on the wrong forecast, compared to a perfect forecast. In the utility interpretation, evaluating deterministic forecasts with mean squared error implicitly defines a decision process in which the disutility is a quadratic function of the distance between forecast and observation. In that case, a series of forecasts that has the smallest MSE has most utility or value for the user.

5.8 Conclusions

Analogously to the Brier score, which measures the squared Euclidean distance between the distributions of observation and forecast, an information-theoretical verification score was formulated, measuring the Kullback-Leibler divergence between those distributions. More precisely, the score measures the divergence from the distribution of the event after the observation to the distribution that is the probability forecast. This "divergence score" is a reinterpretation of the Ignorance score or logarithmic score, which was previously not defined as a Kullback-Leibler divergence. Extending the analogy to the useful and well-known decomposition of the Brier score, the divergence score can be decomposed into uncertainty – resolution + reliability. This decomposition can be interpreted as "the remaining uncertainty is the climatic uncertainty minus the the true information plus the

4. Note that in the reservoir case, unlike the horse race, we also need to take into account the influence of the current decision on the future returns through the state; see chapter 7.

wrong information". For binary events, Brier score and its components are second order approximations of the divergence score and its components.

The divergence score and its decomposition generalize to multi-category forecasts. A distinction can be made between nominal and ordinal category forecasts. Scores based on the cumulative distribution over ordinal categories can be seen as combinations of binary scores on multiple thresholds. How the scores for all thresholds should be weighted relative to each other depends on the user of the forecast. Scores on cumulative distributions are therefore not exclusively dependent on physical observations, but contain subjective weights for the different thresholds. Two possible formulations of a ranked divergence skill score have been formulated. The first equally weighs the skill scores relative to climate, while the second equally weights the absolute scores. The second ranked divergence skill score is equal to the existing ranked mutual information skill score for the case of perfectly calibrated forecasts, but additionally includes a reliability component, measuring miscalibration.

In forecasting, a distinction can be made between information and useful information in a forecast. The latter can not be evaluated without a statement about context in which the forecast will be used. The first is only dependent on how the forecasts relate to the observations and is objective. Therefore, in the author's opinion, information should be the measure for forecast quality. It can be measured using the logarithmic score, which now can be interpreted as the Kullback-Leibler divergence of the forecast from the observation. Useful information or forecast value, on the other hand, is a different aspect of forecast 'goodness' (Murphy, 1993) that should be evaluated while explicitly considering the decision problems of the users of the forecast.

The Brier score can be used as an approximation for quality or as an exact measure of value under the assumption of a group of users with uniformly distributed cost-loss ratios. It is argued that these two applications should be clearly separated. In case one wants to assess quality, information-theoretical scores should be preferred. If an approximation is sufficient, the Brier score could still be used, with the advantage that it is well understood and extensive experience exists with the use of it. However, when extreme probabilities have to be forecast, the differences might become significant and the divergence score is to be preferred on theoretical grounds.

In case value is to be measured, an inventory of the users of the forecasts should be made to assess the total utility. When explicitly investigating the user base, a better estimator for utility than the Brier score can probably be defined. Using the Brier score as a surrogate for forecast value, implicitly assuming the emergent utility function is appropriate for a specific type of forecasts, is clearly unsatisfactory. In this respect, it is important to stress that also the divergence score does not measure value, but quality. Only in a very unrealistic case (a bookmaker offering fair odds) a clear relation exists between the two. In chapter 6, it is argued that for practitioners in meteorology and hydrology, quality is most likely of more concern than value, because the latter is in fact evaluating decisions rather than forecasts.

When evaluating forecasts by comparing them with observations that are affected by measurement uncertainty, the observation uncertainty can influence the evaluation of the forecast quality. The observations should then be represented by probability distributions. When extending the use of the divergence score to the case of uncertain observations, cross entropy score is a more intuitive measure for intercomparison of forecasts at locations with different observational uncertainty. The divergence score (DS) can be interpreted as a measure for the remaining uncertainty relative to the observation. The cross entropy score can be seen as the expected remaining uncertainty with respect to a hypothetical true outcome. Both scores can be decomposed into uncertainty, resolution and reliability. The difference in the decompositions is in the uncertainty component. For the case of the cross-entropy, it represents the climatic uncertainty relative to the truth, for the case of the divergence score it represents the climatic uncertainty relative to the observation. The difference between the two uncertainty components is equal to the difference between the cross-entropy and divergence scores, and corresponds to the average observational uncertainty. If the observations are assumed perfect, which is usually the case in verification practice, both scores and decompositions are equal.

Starting from the observation that deterministic forecasts are still commonly used and evaluated, but are worthless from an information-theoretical viewpoint, the conclusion can be drawn that these forecasts are either implicitly probabilistic or should be viewed in connection to a decision problem. In both interpretations, the evaluation depends on external information that is not provided in the forecast. Deterministic forecasts leave too much interpretation to the user, if seen as implicit probabilistic forecasts, or make too many assumptions on the user if they are evaluated using another utility measure. On the one hand, forecasting can be seen as a communication problem in which uncertainty about the outcome of a random event is reduced by delivering an informative message to a user. On the other hand, forecasting can be seen as an addition of value to a decision problem. Any measure that is not information only becomes meaningful when it is interpreted in terms of utilities.

Science is required to make testable predictions. Forecasts should therefore be stated in terms that make it clear how to evaluate them. Deterministic and interval forecasts fail this criterion, because additional assumptions on utility and probability have to be made during evaluation of the forecast. Probabilistic forecasts can be evaluated using information theory. To facilitate the use of the score and its decomposition, scripts that can be used in Matlab® and Octave are available on the website `http://www.hydroinfotheory.net`

Chapter 6

Some thoughts on modeling, information and data compression

> *"If arbitrarily complex laws are permitted, then the concept of law becomes vacuous, because there is always a law!"*
> - Hermann Weyl, 1932: The Open World, summarizing Gottfried Wilhelm Leibniz, 1686: Discours de métaphysique V, VI[1]

Abstract - This chapter[2] concerns the relation between information and models. Models play an important role in optimal water system operation. They mainly serve as tools for making predictions. These predictions are of paramount importance for making good decisions, but also pure science is required to make testable predictions. While the previous chapter dealt with evaluating such predictions, this chapter goes one step further and looks into the question how such evaluations contribute to improvements in the predictions and the models that make them. Furthermore, the question is addressed whether the purpose of a model should play a role in its calibration. The process of learning from data is analyzed from the intuitive information-theoretical point of view, starting from the strong connection between modeling, data compression and description length. Algorithmic information theory is introduced as a little known, but fundamental framework within philosophy of science. The findings lead to a reflection on the role and importance of understanding in science. From this analysis, some important recommendations follow for the practice of formulating models.

1. Original french text: ... Car supposons, par exemple, que quelqu'un fasse quantité de points sur le papier à tout hasard, comme font ceux qui exercent l'art ridicule de la géomance. Je dis qu'il est possible de trouver une ligne géométrique dont la notion soit constante et uniforme suivant une certaine règle, en sorte que cette ligne passe par tous ces points, et dans le même ordre que la main les avait marqués. ... Mais quand une règle est fort composée, ce qui lui est conforme passe pour irrégulier. ...
2. Contains material from:
   - S.V. Weijs, G. Schoups, and N. van de Giesen. Why hydrological forecasts should be evaluated using information theory. *Hydrology and Earth System Sciences* ,14 (12), 2545–2558, 2010
   - S.V. Weijs, Interactive comment on "HESS Opinions 'A random walk on water' " by D. Koutsoyiannis *Hydrology and Earth System Sciences Discussions*, 6, C2733-C2745, 2009

Science	Data compression
real world	data generating algorithm
observations	file to be compressed
theory	decompression algorithm
errors/remaining uncertainty	compressed file
complexity of theory	size of decompr. algorithm

Computation	Physics
computer	physical system
computation	motion
input	intial state
rules	laws of physics
output	final state

Table 6.1: Analogy between science and data compression (left) and between physical systems and computation (right, according to Deutsch (1998))

6.1 Introduction

Science and data compression have the same objective: by discovering patterns in (observed) data, they can describe them in a compact form. In the case of science, we call this process of compression "explaining" and the compact form a "theory" or physical "law". The similarity of these objectives leads to strong parallels between philosophy of science and the theory of data compression; see table 6.1. A formal description of these ideas was put forward in the "formal theory of inductive inference" by Solomonoff (1964), who was among others inspired by the grammars of Chomsky (1956) and the foundations of probability according to Carnap (1950). Together with similar ideas, independently developed by Kolmogorov (1968) and Chaitin (1966), Solomonoffs theory was the start of the field of "algorithmic information theory" (AIT). The theory offers formal definitions of complexity and algorithmic probability and gives an explication[3] of Occam's razor, which states that the simplest explanation is the best. Because this information theory uses formal theories of computation, the parallels between physical systems and computers must also be noted (see right of table 6.1). Algorithmic information theory complements the information theory of Shannon (1948). Together, these information theories offer a view on the parallels between data compression and model inference. Although the theories of Shannon and Solomonoff start from quite different perspectives, they often lead to remarkably similar results. While Shannon's theory is useful to deal with predictions, algorithmic information theory also considers the models that produce them, making it a good basis for inference. This chapter reviews the implications of these theories, which link data compression to model inference and appear to be largely unknown in the hydrological community. An exploratory real world application is presented, where data-compression is applied to hydrological time series to reveal their information content for inference; see also Weijs and van de Giesen (2011).

6.1.1 The principle of parsimony

A specially designed data compression algorithm can in principle compress a given series to one bit, if the corresponding decompression algorithm contains the whole series and outputs this series if it encounters the bit in the compressed file. This is equivalent to having built a model of the data that is in fact nothing more than restating the same

3. Explication is a term introduced by Carnap (1950). It is the process of making a pre-scientific idea (explicandum) into something precise (explicatum).

data. The model has become as complex as the original data was. When there is unseen data, it is not possible to compress or decompress it with the same algorithms that were developed for the seen data. Therefore, in representing the new data, both the one bit file and the specific decompression algorithm for that file is needed, giving a total file size that is not shorter than the original data. Equivalently, a model of such complexity can reproduce the very specific structure in that data, but can not make predictions applying the found structure to new data. This is known as overfitting and is an important problem throughout science and hydrology is no exception.

In hydrological modeling, the problem is often that limited data are available about complex processes. Therefore, models that perform well in calibration often lack predictive accuracy for unseen data. Generally, more complex models offer higher flexibility in fitting the calibration data, but do not always offer good predictions, while simple models lack precision both in describing the calibration data and predicting new events. A model should therefore be complex enough to describe the data, but not too complex; see e.g. Dooge (1997); Koutsoyiannis (2009); Schoups et al. (2008). Qualitatively this can be formulated in the form of Occam's razor ("The simplest explanation is the best") or the principle of parsimony. Quantum physicist David Deutsch states a similar principle in terms of explanations (Deutsch, 1998):

> "do not complicate explanations beyond necessity, because if you do, the unnecessary complications themselves remain unexplained."

Of course this definition leaves us with the question how to define explanation. Some discussion on this topic follows in section 6.5. Jaynes (2003) puts it this way:

> "Do not introduce details that do not contribute to the quality of your inferences"

and the data compression view would be:

> "Do not increase the size of your compression algorithm beyond the gains in compression that you achieve"

Independently of how the principle is stated, there seems to be such a thing as an optimal model complexity. Several methods have been proposed to determine this optimum. Schoups et al. (2008) compare several of these methods applied to hydrological prediction. Examples of model complexity control methods are cross validation, the Akaike information criterion, AIC, Akaike (1974) and the Bayesian information criterion, BIC, Schwarz (1978), which are all formalizations of the principle of parsimony. More background about several model selection methods can be found in the books by Burnham and Anderson (2002); Li and Vitanyi (2008); Vapnik (1998); Grünwald (2007); Rissanen (2007).

6.1.2 Formalizing parsimony

In the Bayesian framework, comparing model predictions with observed data gives the likelihood, i.e. the conditional probability of the observations, given the model. By multiplying the prior probabilities of parameters and models with the likelihoods, we obtain predictive distributions. The prior parameter estimates are usually based on expert knowledge and

other side information, which is sometimes referred to as "soft information" (Winsemius et al., 2009). These distributions and model structures are thus already conditioned on large amounts of information from observations. Each prior distribution over possible models and parameters is the posterior distribution of previous gains in knowledge. In that sense, "standing on the shoulders of giants" might be translated as "updating distributions of well-informed people". When we follow the growth of knowledge backwards, ultimately we end up with the problem of specifying a prior over models, before we have seen any data. This also becomes relevant if we want to base predictions purely on data, such as for example in the study of artificial intelligence. Intuitively, the prior probability of a model, before seeing any data, has an inverse relation with its complexity. This is a reflection of the principle of parsimony or Occam's razor. In algorithmic information theory, introduced in section 6.2, we will see that complexity is formalized as the minimal length to describe a program (i.e. model) and has a relation to probability. It therefore quantifies the principle of parsimony.

The length of a program that describes the observations is strongly related to compression. If we can produce data with a short program, we can store the program instead of the data and save space. For example, the highly detailed and seemingly complex fractal figure on the cover of this thesis can be stored by the simple algorithm of the Mandelbrot-set and some information regarding coordinates and coloring. If a model (program) does not perfectly describe the data, but only gives high probability to it, then some extra space is needed to encode and store which observation actually occurred, given the predictions of the model. This is also how several data compression algorithms work. In section 6.3 it will be shown that, using an optimal coding scheme, this extra storage per observation converges to the divergence score treated in the previous chapter.

The next section briefly presents some of the main ideas of algorithmic information theory and gives some references for further reading. Subsequently, the relation between the divergence score and data compression is explained in detail. Section 6.4 presents a practical experiment in which hydrological data are compressed with some well-known data compression algorithms to find upper bounds for their information content for hydrological modeling. Subsequently, in section 6.5 the data-compression perspective is used to look critically at what we call understanding. When hydrological models are used to support risk based water system operation, they often have specific purpose other than just understanding. In section 6.6, the question is addressed whether these models should be trained to minimize risk rather than to minimize uncertainty. Some recommendations for the practice of modeling conclude the chapter.

6.2 Algorithmic information theory, complexity and probability

This section introduces algorithmic information theory (AIT), which stems from computer science and early work on artificial intelligence, but has far wider implications. Before

going into the theory, it is useful to briefly review the Bayesian perspective on model inference. It provides the link between probabilistic forecasts, as treated in chapter 5 and the quality of models that produce them. Subsequently, after introducing Turing's formal theory of computation, the connection between description length and probability will become clear when AIT is introduced.

6.2.1 The Bayesian perspective

As was argued in chapter 5, probabilistic forecasts can be judged by how small they make the remaining uncertainty about the truth. For perfect observations, this is achieved by minimizing the divergence score.

$$\min \left\{ \mathrm{DS} = \frac{1}{N} \sum_{t=1}^{N} D_{KL}(\mathbf{o}_t || \mathbf{f}_t) = \frac{1}{N} \sum_{t=1}^{N} \left(-\log[f_t]_{j(t)} \right) \right\} \tag{6.1}$$

where $j(t)$ is the index of the event that occurred at instance t. This minimization is equivalent to maximizing the likelihood, i.e. maximizing the total probability of the data (D), given the model (M)

$$\max \left\{ p(D|M) = \prod_{t=1}^{N} [f_t]_{j(t)} \right\} \tag{6.2}$$

However, in hydrological model inference, we want to find a probable model, given the data, instead of the other way around. When doing model inference, i.e. induction, the probability of interest is thus $p(M|D, I)$ instead of $p(D|M)$. These two probabilities are related through Bayes rule (using the notation of Jaynes (2003))

$$p\left(M_i|D, I\right) = p\left(M_i|I\right) \frac{p\left(D|M_i, I\right)}{p\left(D|I\right)} \tag{6.3}$$

where $p\left(M_i|D, I\right)$ stands for the probability of model i, given the data D and prior information I. Therefore, a prior probability for a model (a combination of model structure and parameter values) is needed. This is captured by $p(M_i|I)$. Note that $p(D|I)$ serves as a normalizing constant and can be eliminated if a fixed set of models is compared. To find a plausible model, it both has to explain the data well and have a high a priori probability, i.e. be consistent with knowledge gathered from previous experiments and observations.

Ideally, to prevent false certainty, the search should be done over the entire space of models that have nonzero prior probability and nonzero likelihood, but this is impossible in practice. Note that many hydrological modeling studies that attempt Bayesian model inference start from a single model structure and convert prior to posterior parameter distributions using the likelihood based on calibration data. Unless there really is evidence that allows the identification of single model structure that is uniquely consistent with prior information, this approach will generally underestimate uncertainty. Recent developments focus on considering multiple model structures, e.g. Fenicia et al. (2008, 2010). However, if after this analysis a single model is chosen that is supposed to represent best

the true catchment behavior, a logical error is committed, because all models but the best are ruled out without evidence to do so. This is an example of overfitting by searching the model-space. Multi-model prediction schemes like Bayesian model averaging Duan et al. (2007) prevent this and are more consistent with Bayesian probability as the logic of science, as expounded in Jaynes (2003).

6.2.2 Universal computability

Algorithmic information theory makes use of the concept of program length. A model is seen as an algorithm or program that runs on a computer. This might seem like a very specific, arbitrary and limited definition, but in fact it is not. It is beyond the scope of this thesis to treat the foundations of computability theory in detail, but the main results will now be mentioned; see Boolos et al. (2007) for more background. One important basis for the ideas in AIT is the Church-Turing thesis, which states that any effectively computable function can be computed by a model of computation Turing (1937) introduced: the Universal Turing Machine (UTM). This thesis has not been proven, but no counterexample has been found so far and it widely believed to be true. Furthermore it was shown that all possible UTMs can simulate each other, given unlimited resources (memory and time). All binary modern age computers are examples of UTMs, although in practice they are bounded by finite memory. They can thus simulate any other Turing machine, and be simulated by any universal Turing machine. A UTM reads symbols from an input tape one by one and changes its internal state according to the symbols read from the tape. Furthermore it can move the (infinite) tape right or left and (over)write symbols on the tape. The first part of the tape can be viewed as a program that tells it how to simulate other computers. An example of a very small UTM has only 7 states (working memory) and 4 different symbols on the tape. Shannon (1956) showed that it is possible to exchange states for symbols on the tape and gave examples of UTM's with just two states or just binary symbols on the tape. Such simple computers, given sufficient time and tape length (memory), can thus simulate any other computer and perform all possible calculations.

Another result from Turing (1937) is that the so-called halting problem is unsolvable. This means that for some programs, the only way to determine whether they will ever halt is to run those programs and test them. For example, one can think about programs with "while-loops" in them, that continue until some condition that follows from the calculation in the loop is fulfilled. This limitation is related to the incompleteness theorems of Gödel (1931) and has substantial implications. As we will see in this chapter, it also means that it is ultimately impossible to say whether we have found the best model.

Given the fact that all known physical processes are Turing-computable, but quantum processes take extreme resources to simulate, Deutsch (1985, 1998) expanded the Turing-Church thesis to state that "every finitely realizible physical system can be perfectly simulated by a universal model computing machine operating by finite means". This analogy is also depicted in Tab. 6.1 and means that any computer is a physical system and any physical system can be regarded as a computer. Physical processes can thus be simulated

on a binary computer and, conversely, computations can be performed by physical systems (cf. e.g. quantum computers). In hydrological terms it means that any hydrological process can in principle be simulated with arbitrary precision by a binary computer.

6.2.3 Kolmogorov complexity, patterns and randomness

Using the thesis that any computable sequence can be computed by a UTM and that program lengths are universal up to an additive constant (the length of the program that tells one UTM how to simulate another), Kolmogorov (1968) gave very intuitive definitions of complexity and randomness; see also (Li and Vitanyi, 2008) for more background. Kolmogorov defined the complexity of a certain string (i.e. data set, series of numbers) as the length of the minimum computer program that can produce that output on a UTM and then halt. Complexity of data is thus related to how complicated it is to describe. If there are clear patterns in the data, then they can be described by a program that is shorter than the data itself. The majority of conceivable strings of data cannot be "compressed" in this way. Data that cannot be described in a shorter way than naming the data itself is defined as random. This is analogous to the fact that a "law" of nature cannot really be called a law if its statement is more elaborate than the phenomenon that it explains (see Tab. 6.1). A problem with Kolmogorov complexity is that it is incomputable, but can only be approached from above. This is related to the unsolvability of the halting problem: it is always possible that there exists a shorter program which is still running (possibly in an infinite loop) that might eventually produce the output and then halt. A paradox that would arise if Kolmogorov complexity were computable is the following definition known as the Berry paradox: "The smallest positive integer not definable in under eleven words".

6.2.4 Algorithmic probability and Solomonoff induction

A few years before Kolmogorov published his paper on complexity, Solomonoff (1964) published a "Formal theory of inductive inference", which has many parallels with the ideas of Kolmogorov. He realized that predictions always rely on finding patterns in past data to predict the next symbol or number in the sequence. A pattern that describes the data, but is very complex usually turns out not to be a good predictor, because it is "overfitted" to the data. Universal Turing machines provide an excellent means to formalize this. A program for a UTM that starts with the known data as output, but is almost as long as the data, is less likely to give correct values for subsequent data points. In other words, if a simple pattern can be recognized in the data, it is more likely to be the process that generated the data than a complex pattern. This is also related the fact that programs for a Turing machine that execute and then halt cannot be the prefix of other programs. This means that other programs cannot start with symbols that form a halting program, because the machine would halt before the other symbols are read.

Solomonoff used the requirement for prefix-free programs to define algorithmic probability. For example, if computer programs would be generated by random coinflips, one quarter of these programs will start with the symbols 01. If "01" would be a halting program,

it would represent one quarter prior probability of all existing programs. Generally, we can associate an algorithmic probability to all halting programs of $2^{-|M_i|}$, where $|M_i|$ is the length of the ith program M_i on Turing machine M. This probability $p\,(M_i)$ could thus be seen as an universal prior for model i with respect to machine M. These priors can be combined with the likelihood of different models with respect to the data, to yield probabilistic predictions. In Solomonoff (1964), several variations of this concept are described which turn out to be equivalent. Solomonoff (1978) gives a sharp upper bound for the prediction error of these predictions and proves that it outperforms any other universal prediction method up to a machine-dependent additive constant. The following equation, using probabilistic models, is perhaps the most intuitive formulation for Solomonoff's universal prediction from data

$$P_M(a_{n+1}|a_1, a_2, \ldots, a_n) = \sum_i 2^{-|M_i|}\ S_i\ M_i(a_{n+1}|a_1, a_2, \ldots, a_n) \tag{6.4}$$

in this equation $P_M(a_{n+1}|a_1, a_2, \ldots, a_n)$ is the probability that machine M predicts for outcome a_{n+1}, given all previous observed data. $M_i(a_{n+1}|a_1, a_2, \ldots, a_n)$ is the forecast probability that model M_i (the ith program) predicts, S_i is the total probability that the model assigned to the observed data (the likelihood, cf. Eq. 6.2) and $2^{-|M_i|}$ is the algorithmic prior probability for model M_i. We can see from the formula that the predictive probability uses a sum over the outcomes of all possible models, weighted by their prior probabilities. It can thus be seen as an instance of Bayes' rule (cf. Eqs. 6.4 and 6.5)

$$P_M(D_{n+1}|D_1, \ldots, D_n, M) = \sum_i p\,(M_i)\, p\,(D_1, \ldots, D_n|M_i)\, p\,(D_{n+1}|D_1, \ldots, D_n, M_i) \tag{6.5}$$

Equation 6.5 gives the probability for all possible values of datapoint D_{n+1}, given all previous data for the reference machine M. When using more complex computers or languages, some functions will be relatively shorter (more probable) and others longer (less probable). The computer is thus a way to represent prior information about which basic patterns we find in nature. Note that when the output to be produced gets longer, the programs also get longer, making the program length cost of one computer simulating another relatively less important. This is exactly equivalent with Bayes' rule, where the prior becomes less important when more data are available.

The reason why Solomonoff induction is not the answer to all problems in science is that, like Kolmogorov complexity, it is incomputable due to the insoluble halting problem. It should thus be seen as golden standard that shows the limits of what is possible and as a guideline to develop methods that approach it and are computable. A perfect method of inference is thus per definition impossible with finite computation resources.

6.2.5 Computable approximations to automated science

Computable approximations to Solomonoff induction can be achieved by limiting or penalizing the time spent in a program or the memory it uses; see e.g. Levin search. Another approach is to perform the summation over a subset of functions (programs) that are computable. One could for example try ever larger neural networks. Some other approaches,

like minimum description length (Rissanen, 2007; Grünwald, 2007), take only the best model instead of a sum over all models. While these methods may not be optimal, they provide computable approximations that will converge when given enough data. In principle, the capabilities of machines to do inductive inference will increase with the growing computing power available. In combination with the growing archives of measured environmental data, see e.g. Lehning et al. (2009) in "the fourth paradigm", this will make the role of computers in science even more important. For example, in Schmidt and Lipson (2009) an algorithm discovered laws of motion for a double pendulum by searching the space of all possible models by genetic programming and in the same issue of Science, robot scientist Adam independently generated and tested scientific hypotheses (King et al., 2009).

6.3 The divergence score: prediction, gambling and data compression

The divergence score presented in chapter 5 as a measure for forecast quality has an interpretation in data compression and gambling. The gambling interpretation is due to Kelly (1956) and in appendix C it is explained how informative forecasts lead to good gambling returns. In the previous section it was also stated that the divergence score has an interpretation as the minimum average description length per observation. This section explains the basic background of this data compression analogy from the viewpoint of the information theory of Shannon (1948), starting from the strong analogy with gambling.

Analogously to the gambling problem where high stakes are put on the most likely events, yielding the highest returns, data compression seeks to represent the most likely events (most frequent characters in a file) with the shortest codes, yielding the shortest total code length. As is the case with dividing the stakes, also short codes are a limited resource that has to be allocated as efficiently as possible. When required to be uniquely decodable, short codes come at the cost of longer codes elsewhere. This follows from the fact that such codes must be prefix free, i.e. no code can be the prefix of another one. This is formalized by the following theorem of McMillan (1956), who generalized the inequality (Eq. 6.6) of Kraft (1949) to all uniquely decodable codes.

$$\sum_i A^{-l_i} \leq 1 \tag{6.6}$$

in which A is the alphabet size (2 in the binary case) and l_i is the length of the code assigned to event i. In other words, one can see the analogy between gambling and data compression through the similarity between the scarcity of short codes and the scarcity of large fractions of wealth. Just as there are only 4 portions of $\frac{1}{4}$ of the wealth available (you cannot divide a pie in five quarters), there are only 4 prefix-free binary codes of length $\log_2 4 = 2$ (see table 6.2, code A). In contrast to fractions of wealth, which can be chosen freely, the code lengths are limited to integers. For example, code B in the table uses one code of length 1, one of length 2 and two of length 3, we can verify that it sharply

	occurrence frequencies			codes		expected code lengths per value					
event	I	II	III	A	B	A_I	B_I	A_II	B_II	A_III	B_III
1	0.25	0.5	0.4	00	0	0.5	0.25	1	0.5	0.8	0.4
2	0.25	0.25	0.05	01	10	0.5	0.5	0.5	0.5	0.1	0.1
3	0.25	0.125	0.35	10	110	0.5	0.75	0.25	0.375	0.7	1.05
4	0.25	0.125	0.2	11	111	0.5	0.75	0.25	0.375	0.4	0.6
total	H=2	H=1.75	H=1.74			2	2.25	2	1.75	2	2.15

Table 6.2: Assigning code lengths proportional to minus the log of their probabilities leads to optimal compression. Code B is optimal for distribution II, but not for the other distributions. Distribution III has no optimal code that achieves the entropy bound.

satisfies Eq. 6.6, using $D = 2$ (notice also the analogy to Fig. 3.1 on page 41), we find $1 * 2^{-1} + 1 * 2^{-2} + 2 * 2^{-3} = 1 \leq 1$

In table 6.2, it is shown how the total code length can be reduced, assigning codes of varying length depending on occurrence frequency. As shown by Shannon (1948), if every value could be represented with one code, allowing for non-integer code lenghts, the optimal code length for an event i is $l_i = \log(1/p_i)$. The minimum average code length is the expectation of this code length over all events, H bits per sample, where H can be recognized as the entropy of the distribution (Cover and Thomas, 2006), which is a lower bound for the average description length.

$$H(\mathbf{p}) = E_{\mathbf{p}}\{l\} = \sum_{i=1}^{n} p_i \log \frac{1}{p_i} \tag{6.7}$$

However, because in reality the code lengths have to be rounded to an integer number of bits, some overhead will occur. The rounded coding would be optimal for a probability distribution of events

$$q_i = \frac{1}{2^{l_i}} \forall i, \tag{6.8}$$

such as frequencies II in table 6.2. In this equation, q_i is the i^{th} element of the probability mass function $\mathbf{q}$ for which the code would be optimal and l_i is the code length assigned to event i. The overhead in the case where $\mathbf{p} \neq \mathbf{q}$ is $D_{KL}(\mathbf{p}||\mathbf{q})$, yielding a total average code length of

$$H(\mathbf{p}) + D_{KL}(\mathbf{p}||\mathbf{q}) \tag{6.9}$$

bits per sample. In general, if a wrong probability estimate is used, the number of bits per sample is increased by the Kullback-Leibler divergence from the true to the estimated probability mass function.

For probability distributions that do not coincide with integer ideal code lengths, the algorithm known as Huffman coding (Huffman, 1952) was proven to be optimal for value by value compression. It finds codes of an expected average length closest to the entropy-bound and is applied in popular compressed picture and music formats like jpg, tiff, mp3 and wma. For a good explanation of the workings of this algorithm, the reader is referred to Cover and Thomas (2006). In table 6.2, code A is optimal for the naive probability distribution and code B is optimal for the distribution II. Both these codes achieve the

entropy bound. Code B is also an optimal Huffman code for the distribution III (last column in table 6.2). Although the expected code length is now more than the entropy, it is impossible to find a shorter code. The overhead is equal to the Kullback-Leibler divergence from the true distribution (III) to the distribution for which the code would be optimal.

$$D_{KL}(III||II) = D_{KL}((0.4, 0.05, 0.35, 0.2)\,||\,(0.5, 0.25, 0.125, 0.125)) = 0.4106$$

If the requirement that the codes are value by value (one code for each observation) is relaxed, blocks of values can be grouped together to approach an ideal probability distribution. When the series are long enough, entropy coding methods like Shannon and Huffman coding using blocks can get arbitrarily close to the entropy bound (Cover and Thomas, 2006).

6.3.1 Dependency

If the values in a time series are not independent, however, the dependencies can be used to achieve even better compression. This high compression results from the fact that for dependent values, the joint entropy is lower than the sum of entropies of indi idual values. In other words, average uncertainty per value decreases, when all the other values in the series are known, because we can recognize patterns in the series, that therefore contain information about themselves. Hydrological time series often show strong internal depen encies, leading to better compression and bett r prediction. Consider, for example, the case where you are asked to gamble on (or assign code lengths to) the streamflow value on May, 12, 1973. In one case, the information offered is the dark-colored climatological histogram (Fig. 6.1 on the right), in the second case, the time series is available (the left of the same figure). Obviously, the e pected compression and expected return for the bets are better in the second case, which shows the value of exploiting dependencies in the data. The surprise ($-\log P_{\text{true value}}$) upon hearing the true value is 3.72 bits in case the guessed distribution was assumed and 4.96 bits when using the climate as prior. These surprises are equivalent to the divergence scores treated in the previous chapter.

Another example are the omitted characters that the careful reader may (not) have found in the previous paragraph. There are 48 different characters used, but the entropy of the text is 4.3 bits, far less than log(48)=5.6, because of for example the relatively high frequencies of the space and the letter “e”. Although the entropy is more than 4 bits, the actual uncertainty about the missing letters is far less for most readers, because the structure in the text is similar to english language and that structure can be used to predict the missing characters. On the one hand this means that english language is compressible and therefore fairly inefficient. On the other hand this redundancy leads to more robustness in the communication, because even with many typographical errors, the meaning is still clear. If english were 100% efficient, any error would obfuscate the meaning.

In general, better prediction, i.e. less surprise, gives better results in compression. In water resources management and hydrology we are generally concerned with predicting one

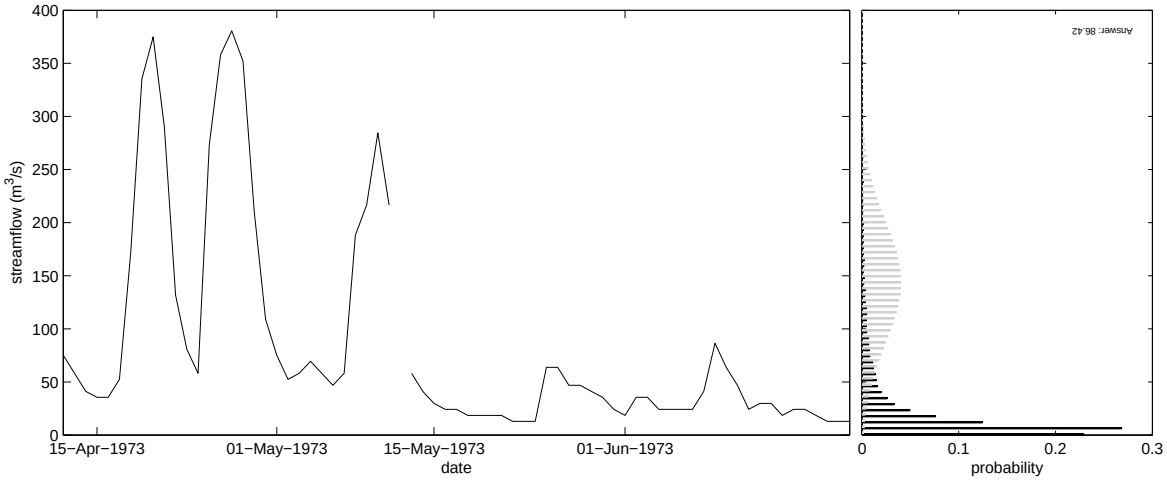


Figure 6.1: The missing value in the flow time series can be guessed from the surrounding values (a guess would for example be the grey histogram). This will usually lead to a better guess than one purely based on the occurrence frequencies over the whole 40 year data set (dark histogram) alone.

series of values from other series of values, like predicting streamflow (Q) from precipitation (P) and potential evaporation (E_p). In gambling we would call precipitation and evaporation series side information. In terms of data compression, knowledge of P and E_p would help compressing Q, but would also be needed for decompression. When P, E_p and Q would be compressed together in one file, the gain compared to compressing the files individually is related to what a hydrological model learns from the relation between these variables. Similarly, we can try to compress hydrological time series to investigate how much information those compressible series really contain for hydrological modeling.

6.4 A practical test: "Zipping" hydrological time series

In this section, a number of compression algorithms will be applied to different datasets to obtain an indication of the amount of information they contain. Most compression algorithms use entropy-based coding methods such as introduced in the previous section, often enhanced by methods that try to discover dependencies and patterns in the data, such as autocorrelation and periodicity.

The data compression perspective indicates that formulating a rainfall-runoff model has an analogy with compressing rainfall-runoff data. A short description of the data will include a good model about it. However, not all patterns found in the data should be attributed to the rainfall-runoff process. For example, a series of rainfall values is highly compressible due to the many zeros (a far from uniform distribution), the autocorrelation, and the seasonality. These dependencies are in the rainfall alone and can tell us nothing about the relation between rainfall and runoff. The amount of information that the rainfall contains for the hydrological model is thus less than the number of data points multiplied

by the number of bits to store rainfall at the accurate precision. This amount is important because it determines the model complexity that is warranted by the data (Schoups et al., 2008). In fact, we are interested in the Kolmogorov complexity of the data, but this is incomputable. A crude practical approximation of the complexity is the filesize after compression by some commonly available compression algorithms. This provides an upper bound for the information in the data.

If the data can be regenerated perfectly from the compressed (colloquially referred to as zipped) files, the compression algorithm is said to be lossless. In contrast to this, lossy compression introduces some small errors in the data. Lossy compression is mainly used for various media formats (pictures; video; audio), where these errors are often beyond our perceptive capabilities. This is analogous to a model that generates the observed values to within measurement precision. This section gives one example of lossy compression, but will be mainly concerned with lossless compression. Roughly speaking, the file size that remains after compression, gives an upper bound for the amount of information in the time series. Actually, also the code-length of the decompression algorithm should be counted towards this file size (cf. a self-extracting archive). In the present exploratory example the inclusion of the algorithmic complexity of the decompression algorithm will be left for future research. The compression algorithms will be mainly used to explore the difference in information content between different signals.

6.4.1 Data and Methods

The time series used

The algorithms are tested on a real world hydrological dataset from Leaf River (MS, USA) consisting of rainfall, potential evaporation and streamflow. See e.g. Vrugt et al. (2003) for a description of this data set. As a reference, various artificially generated series where used. The generated series consist of 50000 values, while the time series of the Leaf River dataset, contains 14610 values (40 years of daily values). The following series where used in this experiment

constant contains only 1 value repeatedly. Intuitively, this file contains the least possible amount of information.

linear contains a slowly linearly increasing trend, ranging from 0 at the beginning of the series to 1 at the end of the series.

uniform_white is the output from the Matlab® function "rand", it is uncorrelated white noise with a uniform distribution between 0 and 1.

Gaussian_white is the output from the Matlab® function "randn", it is uncorrelated white noise with a normal distribution and is scaled between 0 and 1.

sine_1 is a sinusoidal wave with a wavelength spanning all 50000 values, ranging between 0 and 1.

sine_100 is a sinusoidal wave with a wavelength spanning 1/100 of 50000 values, ranging between 0 and 1. It is therefore a repetition of 100 sine waves.

Leaf_P is a daily rainfall series from the catchment of Leaf river (1948-1988).

Leaf_Q is the corresponding daily series of observed streamflow in Leaf river .

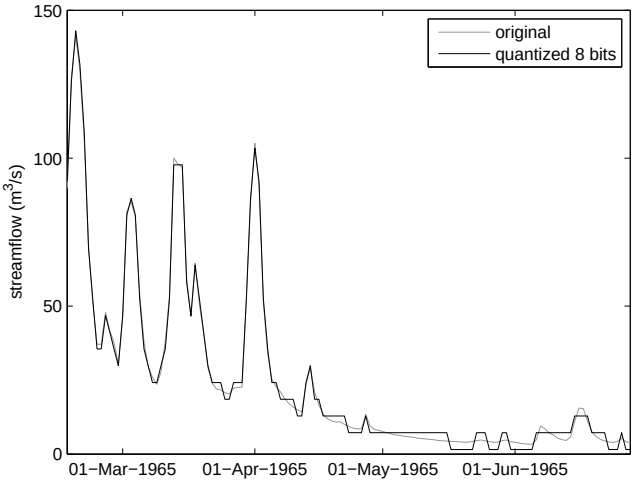


Figure 6.2: The effect of quantization. Because the errors are absolute, the largest relative errors occur in the low-flow periods, such as selected for this figure.

Quantization

Due to the limited amount of data, quantization is necessary to make correct estimates of the distributions, which are needed to calculate the amount of information and compression. This is analogous to the maximum number of bins permitted to draw an informative histogram. Although the quantization constitutes a loss of information, it does not affect the results, as they are all measured relative to the quantized series. All series were first quantized to 8 bit precision, using a simple linear quantization scheme (eq. 6.10). Using this scheme, the series were split into $2^8 = 256$ equal intervals and converted into an 8 bit unsigned integer (an integer ranging from 0 to 255 that can be stored in 8 binary digits).

$$x_{\text{integer}} = \lfloor 0.5 + 255 \frac{x - \min x}{\max x - \min x} \tag{6.10}$$

These can be converted back to real numbers using

$$x_{\text{quantized}} = (\frac{\max x - \min x}{255}) x_{\text{integer}} + \min x \tag{6.11}$$

Because of the limited precision achievable with 8 bits , $x_{\text{quantized}} \neq x$. As can be seen in figure 6.2, the quantization leads to rounding errors, which can be quantified as a signal to noise ratio (SNR). The SNR is the ratio of the variance of the original signal to the variance of the rounding errors.

$$\text{SNR} = \frac{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2}{\frac{1}{n}\sum_{t=1}^{n}\left(x_t - x_{t,\text{quantized}}\right)^2} \tag{6.12}$$

Because the SNR can have a large range, it is usually measured in the form of a logarithm, which is expressed in the unit decibel: $\text{SNR}_{dB} = 10 \log_{10}(\text{SNR})$.

Compression algorithms

The algorithms that were used are a selection of commonly available compression programs and formats. Below are very short descriptions of the main principles and main features of each of the algorithms used and some references for more detailed descriptions. The descriptions are sufficient to understand the most significant pattern in the results. It is beyond the scope of this thesis to describe the algorithms in detail.

ARJ Uses LZ77 (see LZMA) with sliding window and Huffman coding.

WAVPACK is a lossless compression algorithm for audio files.

JPG_LS The Joint Photography Experts Group created the JPEG standard, which includes a range of lossless and lossy compression techniques. Here the lossless coding is used, which uses a Fourier-like type of transform (Discrete cosine transform) followed by Huffman coding of the errors).

JPG_50 Is the the result of the JPG format in lossy mode. After the discrete cosine transform, it discards the higher frequencies (wavelet coefficients), which results in a loss of small-scale detail.

HDF_RLE HDF (hierarchical data format) is a data format for scientific data of any form, including pictures, time series and metadata. It can use several compression algorithms, including run length encoding (RLE). RLE replaces sequences of reoccurring data with the value and the number of repetitions. It would therefore be useful to compress pictures with large uniform surfaces and rainfall series with long dry periods.

PPMD a variant of Prediction by Partial Matching, implemented in the 7Zip program. It uses a statistical model for predicting each value from the preceding values using a variable sliding window. Subsequently the errors are coded using Huffman Coding.

LZMA The Lempel-Ziv-Markov chain algorithm combines the Lempel-Ziv algorithm, LZ77 (Ziv and Lempel, 1977), with a Markov-Chain model. LZ77 uses a sliding window to look for reoccurring sequences, which are coded with references to the previous location where the sequence occurred. The method is followed by range coding. Range coding (Martin, 1979) is an entropy-coding method which is mathematically equivalent to arithmetic coding (Rissanen and Langdon, 1979), it has less overhead than Huffman coding.

BZIP2 Uses the Burrows and Wheeler (1994) block sorting algorithm in combination with Huffman-Coding.

PNG Portable Network Graphics (PNG) uses a filter based on prediction of one pixel from the preceding pixels. Afterwards, the prediction errors are compressed by the algorithm "Deflate" which uses dictionary coding (matching repeating sequences) followed by Huffman coding.

TIFF A container image format that can use several compression algorithms. In this case PackBits compression was used, which is a form of run length encoding.

6.4.2 Results

Lossless compression results

As expected, the filesizes after quantization are exactly equal to the number of values in the series, as each value is encoded by one byte (8 bits) and stored in binary raw format. From the occurrence frequencies of the 256 unique values, the entropy of their distribution was calculated. Normalized with the maximum entropy of 8 bits, the fractions in row 3 of table 6.3 give an indication of the entropy bound for the ratio of compression achievable by value by value entropy encoding schemes such as Huffman coding, which do not use temporal dependencies.

The signal to noise ratios in row 4 give an indication of the amount of data corruption that is caused by the quantization. As a reference, the uncompressed formats BMP (Bitmap), WAV (Waveform audio file format), and HDF (Hierarchical Data Format) are included, indicating that the file size of those formats, relative to the raw data, does not depend on what data are in them, but does depend on the amount of data, because they have a fixed overhead that is relatively smaller for larger files.

The results for the various lossless compression algorithms are shown in rows 7-17. The numbers are the percentage of the file size after compression, relative to the original filesize (a lower percentage indicates better compression). The best compression ratios per time series are highlighted. From the result it becomes clear that the constant, linear and periodic signals can be compressed to a large extent. Most algorithms achieve this high compression, although some have more overhead than others. The uniform white noise is theoretically incompressible, and indeed none of the algorithms appears to know a clever way around this. In fact, the smallest file size is achieved by the WAV format, which does not even attempt to compress the data and has a relatively small file header (meta information about the file format). The Gaussian white noise is also completely random in time, but does not have a uniform distribution. Therefore the theoretical limit for compression is the entropy bound of 86.3 %. The WAVPACK algorithm gets closest to the theoretical limit, but also several file archiving algorithms (ARJ, PPMD, LZMA BZIP2) approach that limit very closely. This is because they all use a form of entropy coding as a backend (Huffman and Range coding). Note that the compression of this non-uniform white noise signal is equivalent to the difference in uncertainty expressed by the bars “naive” and “climate” in Fig. 5.8 on page 111.

The results for the hydrological series firstly show that the streamflow series is better compressible than the precipitation series. This is remarkable, because the rainfall series has the lower entropy. Furthermore is can be seen that for the rainfall series, the entropy-bound is not achieved by any of the algorithms, presumably because of the overhead caused by the occurrence of 0 rainfall more than 50 percent of the time, see Eqs. 6.8 and 6.9 on page 130. Further structure like autocorrelation and seasonality can not be used sufficiently to compensate for this overhead. In contrast to this, the streamflow series can be compressed to well below the entropy bound (27.7% vs. 42.1%), because of the strong autocorrelation in the data. These dependencies are best exploited by the PPMD

dataset	constant	linear	uniform white	Gaussian white	sine 1	sin 100	Leaf Q	Leaf P
filesize	50000	50000	50000	50000	50000	50000	14610	14610
$\frac{H}{\log N}$	0.0	99.9	99.9	86.3	96.0	92.7	42.1	31.0
SNR	NaN	255.0	255.6	108.0	307.4	317.8	42.6	39.9
Uncompressed formats								
BMP	102.2	102.2	102.2	102.2	102.2	102.2	407.4	407.4
WAV	100.1	100.1	100.1	100.1	100.1	100.1	100.3	100.3
HDF_NONE	100.7	100.7	100.7	100.7	100.7	100.7	102.3	102.3
Lossless compression algorithms								
JPG_LS	12.6	12.8	110.6	94.7	12.9	33.3	33.7	49.9
HDF_RLE	2.3	2.7	101.5	101.5	3.2	92.3	202.3	202.3
WAVPACK	0.2	1.9	103.0	87.5	2.9	25.6	38.0	66.2
ARJ	0.3	1.0	100.3	88.0	3.1	1.9	33.7	40.0
PPMD	0.3	2.1	102.4	89.7	3.6	1.4	27.7	36.4
LZMA	0.4	0.9	101.6	88.1	1.9	1.2	31.0	37.8
BZIP2	0.3	1.8	100.7	90.7	3.0	2.3	29.8	40.5
PNG	0.3	0.8	100.4	93.5	1.5	0.8	40.2	50.0
GIF	2.3	15.7	138.9	124.5	17.3	32.0	38.8	45.9
TIFF	2.0	2.4	101.2	101.2	2.9	91.2	201.5	201.5
Lossy compression								
JPG_50	10.0	10.1	150.6	114.9	10.2	20.4	26.3	59.3

Table 6.3: The performance, as percentage of the original file size, of well known compression algorithms on various time series. The best results per signal are highlighted.

algorithm, which uses a local prediction model that apparently can predict the correlated values quite accurately. Many of the algorithms cross the entropy bound, indicating that they use at least part of the temporal dependencies in the data.

Lossy compression results

Apart from the lossless data compression, which can give upper limits for the total amount of information in a time series, lossy compression can also be of interest for hydrological modeling. Instead of representing the series exactly, lossy compression retains the subjectively most important features of the data and discard the other detail, depending on the amount of compression required. This approach has some parallels with the idea of rainfall multipliers to account for uncertainty in measured rainfall, such as applied by Kavetski et al. (2006) and various related papers. The loss of detail can also be seen as the addition of a noise and is comparable to the effect of the quantization discussed earlier, which is in fact also a form of lossy data compression.

As an example of lossy compression, the picture format JPG was used to store the various time series. As can be seen from the table, the compression ratio for the streamflow series outperforms all lossless compression algorithms. Of course, this strong compression comes at the cost of errors introduced by the JPG compression algorithm. The errors are shown in figure 6.3 and follow from the loss of finer details by eliminating terms from the discrete cosine transform used in JPG. The signal to noise ratio (SNR) over the entire series was

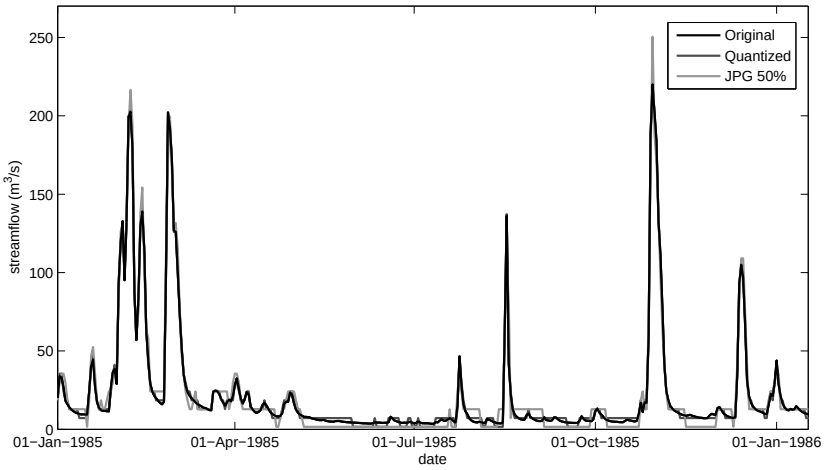


Figure 6.3: The effect of lossy compression on the signal. It introduces a noise but is not significantly more than the quantization noise.

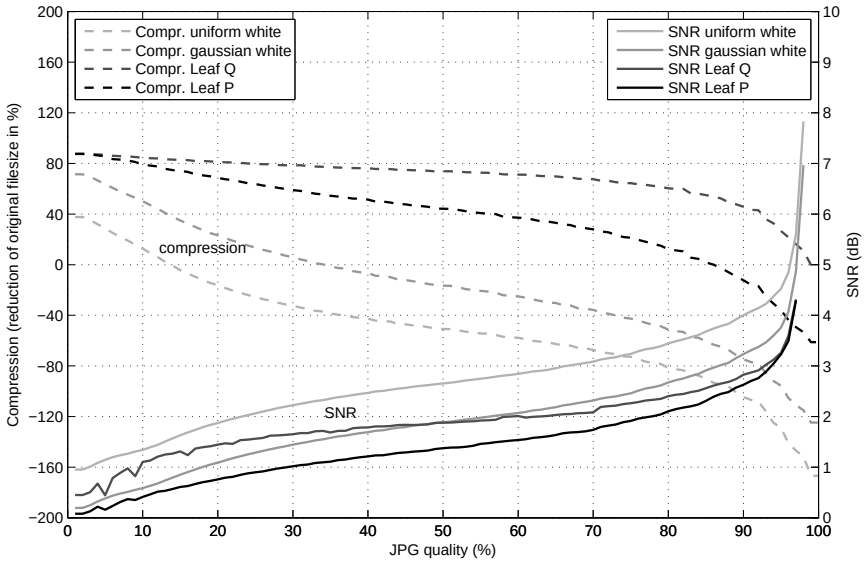


Figure 6.4: The dependency of compression (negative indicates a grown file size) and signal to noise ratio (logarithmic) on the JPG quality setting in "lossy mode". Especially for the discharge signal (Q), large compression can be achieved if a small noise is allowed. For high quality, the compression is negative.

75.6 in the case of JPG with 50% quality setting. This SNR is the noise that is added to the signal already quantized to 8 bit precision. The signal to noise ratio with respect to the original signal is 76.3, which is remarkable, because this seems to indicate that in this case the two corruptions of the signal cancel each other out somewhat. Further results indicating the tradeoff between quality and compression are shown in figure 6.4.

statistic	P	Q	Qmod	errQ	Q|Qmod	perm_Q	perm_errQ
entropy (% of 8 bits)	31.0	42.1	44.9	38.9	26.4	42.1	38.9
best compression (%)	36.4	27.7	25.8	31.5	N.A.	45.4	44.1
std. dev. (range=256)	11.7	11.6	10.4	4.95	N.A.	11.6	4.95
Autocorrelation ρ	0.15	0.89	0.95	0.60	N.A.	<0.01	<0.01

Table 6.4: Information-theoretical and variance statistics and compression results (remaining file size %) for rainfall-runoff modeling.

6.4.3 Compressing with hydrological models

In the previous paragraph single time series were compressed to obtain an indication of their information content. Given the connection between modeling and data compression, a hydrological model should in principle be able to compress hydrological data. This can be useful to identify good models in information-theoretical terms, but can also be useful for actual compression of hydrological data. Although a more detailed analysis is left for future work, this section contains a first test of estimating the performance of hydrological models using data compression tools.

The hydrological model HYMOD was used to predict discharge from rainfall for the Leaf River dataset; see e.g. Vrugt et al. (2009) for a description of model and data. Subsequently, the modeled discharges were quantized using the same quantization scheme as the observed discharges. An error signal was defined by subtracting the modeled (Qmod) from the observed (Q) quantized discharge. This gives a signal that can range from -255 to +255, but because the errors are sufficiently small, ranged from -55 to +128, which allows for 8 bit coding. Because the observed discharge signal (Q) can be reconstructed from the precipitation time series (P), the model, and the stored error signal (errQ), the model could enable compression of the dataset consisting of P and Q. In table 6.4, the entropies of the signals are shown. The second row shows the resulting filesize as percentage of the original filesize for the best compression algorithm for each series (PPMD or LZMA).

The table also shows the statistics for the series where the order of the values was randomly permuted (perm_Q and perm_errQ). As expected this does not change the entropy, because that depends only on the histograms of the series. In contrast, the compressibility of the signals is significantly affected, indicating that the compression algorithms made use of the temporal dependence for the non-permuted signals. The joint distribution of the modeled and observed discharges was also used to calculate the conditional entropy $H(Q|Q_{\mathrm{mod}})$. It must be noted, however, that this conditional entropy is probably underestimated, as it is based on a joint distribution with 255^2 probabilities estimated from 14610 value pairs. This is the cost of estimating dependency without limiting it to a specific functional form. The estimation of mutual information needs more data than Pearson correlation, because the latter is limited to a linear setting and looks at variance rather than uncertainty. In the description length, the underestimation of $H(Q|Q_{\mathrm{mod}})$ is compensated by the fact that the dependency must be stored by the entire joint distribution. If representative for the dependence in longer data sets, the conditional entropy gives a

theoretical limit of compressing Q with knowledge of P and the model, while not making use of temporal dependence.

A somewhat unexpected result is that the errors seem more difficult to compress (31.5 %) than the observed discharge itself (27.7 %), even though the entropy is lower. Apparently the reduced temporal dependence in the errors (lag-1 autocorrelation coefficient $\rho = 0.60$), compared to that of the discharge ($\rho = 0.89$), offsets the gain in compression due to the lower entropy of the errors. Possibly, the temporal dependence in the errors becomes to complex to be detected by the compression algorithms. Further research is needed to determine the exact cause of this result, which should be consistent with the theoretical idea that the information in P should reduce uncertainty in Q. The Nash-Sutcliffe Efficiency (NSE) of the model over the mean is 0.82, while the NSE over the persistence forecast ($Q_{\mathrm{mod}}(t) = Q_{t-1}$) is 0.18 (see Schaefli and Gupta, 2007), indicating a reasonable model performance. Furthermore, the difference between the conditional entropy and the entropy of the errors could indicate that an additive error model is not the most efficient way of coding and consequently not the most efficient tool for probabilistic prediction. The use of for example heteroscedastic probabilistic forecasting models (e.g. Pianosi and Soncini-Sessa, 2009) for compression is left for future work.

6.4.4 Discussion and conclusions

This section presented an initial attempt at using data compression as a practical tool in the context of learning from data and estimating the information content of hydrological signals. To the author's knowledge, this is the first time that this approach is used in hydrology. The present study is limited in scope, and more elaborate studies would be interesting. The results show that hydrological time series contain a large amount of redundant information, due to their far from uniform distributions and temporal dependencies. Model complexity control methods that use the number of observations should therefore be corrected for the true information content. Compression tools that search for patterns in individual time series can be helpful for this task. It would be interesting to develop compression tools specifically aimed at hydrological time series, that also give an overview of the various patterns found in the data.

A hydrological model actually is such a compression tool. It makes use of the dependencies between for example rainfall and streamflow. The patterns that are already present in the rainfall reduce the information that the hydrological model can learn from: a long dry period could for example be summarized by one parameter for an exponential recession curve in the streamflow. The information available for a rainfall runoff model could theoretically be estimated by comparing the filesize of compressed rainfall plus the filesize of compressed streamflow with the size of a file where rainfall and streamflow are compressed together, exploiting their mutual dependencies. We could denote this as:

$$\text{learnable info} = |\mathrm{ZIP}(P)| + |\mathrm{ZIP}(Q)| - |\mathrm{ZIP}(P,Q)| \tag{6.13}$$

where $|\mathrm{ZIP}(X)|$ stands for the filesize of a theoretically optimal compression of data X, which includes the size of the decompression algorithm. A good benchmark for this

could be a self-extracting archive, i.e. the filesize of an executable file that reproduces the data on some given operating system on a given computer. This brings us back to the ideas of algorithmic information theory, which use program lengths on Turing machines. The shortening in description length when merging input and output data, i.e. the compression progress, could be seen as the amount of information learned by modeling. The hydrological model that is part of the decompression algorithm embodies the knowledge gained from the data.

6.5 Prediction versus understanding[4]

Data compression just looks for patterns, i.e. correlations; dependencies, in the data. It does not worry about cause and effect, nor does it take into account whether a certain model works for the right reasons. Model performance is reduced to one single benchmark, the description length, and there is no distinction between different aspects to the quality of a model. The single number that summarizes model performance does not allow for model diagnostics, nor does it reflect possibly different views to what constitutes a good model, depending on its use. The mechanistic approach seems to obviate the need for art in modeling, such as advocated by Savenije (2009). The process of hypothesis forming and testing has no importance in the search for patterns. In the following, it is argued that none of these objections is fundamentally a problem. For practical reasons, given the present state of the art, human intellect is still vital, but this does not preclude the possibility of more mechanistic science in the future. The following section addresses some of these issues and is at points deliberately provocative to stimulate discussion and present a different view on some of the current discourse in hydrology.

6.5.1 Hydrological models approximate emergent behavior

Hydrological systems are high-dimensional, complex systems. They consist of an extremely large number of water molecules bouncing against each other and against those of soil and vegetation particles. Elementary forces are acting upon them and photons are hitting their atoms. Even though it is impossible to predict the paths of individual water molecules with accuracy, the macro-states of a large number of molecules interacting is surprisingly predictable. Although in some cases this might be seen as some form of (biological/ecological) self-organization, fundamentally this predictability follows from the calculus of probabilities. Given a large number of equally probable microstates, probability in a complex system often concentrates in a small number of possible macro-states. An example is the sum of the outcomes of a large number of dice. The uncertainty about the precise microstate (the outcomes of all individual dice) is equal to the sum of uncertainties of the individual dice, but the uncertainty about the sum is far less. Also in a hydrological system, the macro-states, which are for example sums or averages such as

4. This section is based on a comment on an early version of Koutsoyiannis (2010), which is recommended for related discussions.

the water storage and vegetation cover, are far more predictable than the micro-states, such as the position of all water molecules and the activity of the individual stomata in the vegetation leaves. Compare also the relatively simple behavior of flow in a river and the complex flow of water through all its pathways in the catchment. We could see this as an emergence of predictability from randomness (cf. Koutsoyiannis (2010) emergence of randomness from determinism). This very fundamental mechanism of emergence, both visible in evolutionary processes and the movement of systems towards maximum entropy, is the reason why we can make hydrological predictions in the first place.

The realization that hydrological processes are emergent behavior from complex interactions is also a reason why forecasts should be probabilistic rather than deterministic. Conceptual hydrological models can only be seen as approximations of complex hydrological systems. This argument is complementary to the arguments of consistency and testability given in the previous chapter, and to the notion that systems with feedbacks often show chaotic behavior, limiting predictability to a certain time horizon (Lorenz, 1963; Koutsoyiannis, 2010). When seen as emergent behavior from a complex system, the limited predictability is analogous to random fluctuations of the sum of a very large number of dice, or, as Grandy Jr (2008) puts it: "Effects of the microscopic dynamical laws can only be studied at the macroscopic level by means of probability theory".

One might feel that modeling emergent behavior is unsatisfactory, especially when probability replaces truth and description of behavior replaces explanation. A statement that is often heard is that "a model might give the right answers for the wrong reasons". However, it can be argued that there is no way to determine what the "right reasons" are, except comparing other answers that the model gives with observations to determine whether they are right. The following section addresses the question whether there is in fact a fundamental difference between a short description and explanation.

6.5.2 Science and explanation as data compression?

The analogy between modeling and data compression can also be applied to the whole of science in general. When using the theory of inductive inference of Solomonoff (1964), the principles of parsimony and good model performance are sufficiently formal to be implemented in an automatic machine. This would yield the best predictions, given all historical observations. An interesting question would therefore be whether the current state of science would follow with a high probability if such an inference machine would be fed with all observed data that humanity has recorded so far. Would it yield the best possible predictions and would its shortest programs be our best theories?

Some authors, e.g. Deutsch (1998), make the distinction between prediction and explanation. The question that naturally arises is then what defines a good explanation apart from consistency with observations and a requirement for parsimony. Sometimes it is required that explanations match intuition or are understandable. It could also be posed, however, that intuition and understanding emerge from seeing patterns in observations too. This would mean that they emerge from trying to make good predictions.

Taking the data-compression view, physics can never really "explain" a phenomenon in an other way than just describing it in a more compact, accurate, or generalized way, which all lead to shorter programs. In other words, we try to codify observed behavior in laws as much as possible (leave little noise) and a shorter code is preferred over a larger one. Laws should therefore be as general as possible and re-use of laws for various problems is encouraged. In science, a good set of models is the set of models that best describes all observations so far and with the smallest total complexity. Whenever it is possible to unify two parts of physics into one and it yields a shorter program, progress has been made. Sometimes this progress is visible in terms of prediction of thus far unobserved phenomena. This progress is cashed at the moment when it is observed, for example if the Higgs Boson is found. Another way to advance science is to observe phenomena that cannot be explained by the current set of models, forcing the models to become more complex in order to make good predictions of the phenomena causing these new observations, but can also lead to discoveries of new patterns.

Of course, scientific progress is not limited to fundamental physics and explaining the unexplained. Sometimes the emergent behavior of a complex system can be described in a much shorter way than the reductionist explanation, avoiding the need to specify the full micro-state. Especially because the full micro-state is impossible to observe anyway, it has no value trying to predict it. What constitutes a good model inherently depends on the location of the observer and his access to information; see also the very interesting paper about subjectivity of theories of everything by Hutter (2010). Modeling relations between macro-states is what we mainly try to do in hydrology, referred to by Koutsoyiannis (2010) as "overstanding", but in a way also in Newtonian mechanics. The difference is that the latter almost perfectly describes the emergent behavior in many conditions.

When we infer from a limited data set, a "physically based" model should in principle be able to outperform a purely data-driven model, because it has more prior information in it. The model structure has become very likely on the basis of all past observations. A model structure violating a physical law should be punished through all historical observations that confirmed that law.[5]Although a physically based model might not have a better fit to the data than a data-driven model of similar complexity, the historical observations of e.g. mass balance should give it a higher probability. This makes the use of physically based models consistent with the data-compression view, which dictates that, in absence of prior information, predictive performance and complexity are enough to fully describe the merit of a model. Another way to view this is that not the complexity of the model itself, but the amount of complexity added to accepted scientific knowledge should be penalized in model complexity control. Theories are thus ways to conveniently describe our knowledge and combine that with new observations. However, ultimately theories are not fundamental to science. According to Solomonoff, we do not need theories at all and everything is just about predictions.

5. Note that when for example a hydrological model violates the mass balance, this can be usually interpreted as a flux across the boundary that is not explicitly modelled. Only when a model attempts to model all in- and outgoing mass-fluxes and does not conserve mass, we would regard it as contradicting the mass balance and give it infinitesimal prior probability.

6.5.3 What is understanding?

If science is just about compressing observations to get good predictions, where is the understanding? We could pose that what is usually seen as understanding is nothing more than seeing analogies between the mathematical relations that give good predictions and our intuition based on observations in everyday life. We intuitively understand conservation of mass because we see it everyday. We understand the movement of molecules in a gas because it is analogous to bouncing marbles in some way. We understand the concept of waves by picturing the movement of ripples when we throw a stone in the water. Understanding is thus nothing more than picturing the predictive model in terms of similar relations in our observable world. This is also the reason why "Nobody understands quantum mechanics" (Feynman, 1965). If the way the world behaves at a certain scale does not have analogous counterparts on a human-observable scale, it is impossible to understand, given this notion of understanding. However, physicists sufficiently familiar with the phenomena might simply expand their intuition to achieve understanding.

If we see understanding as matching intuition, we could observe that understanding is often overrated. It is often stated as an objective as such. From an aesthetic point of view it is desirable that a model is understandable in the sense that it has analogies to observable behavior on a human scale. An example is to picture a catchment as a series of interconnected buckets. However, in many cases, the system we try to model simply does not behave in such a way. In those cases, the understandable model structure compromises prediction accuracy and is not closer to how the actual system "works" than a black-box model fitted to the data. The advantage of conceptual models is that knowledge that has been gained from past observations, like conservation of momentum and mass, is easily added to the model in the form of constraints on the structure. Data that has been previously transformed into knowledge of laws (patterns that are general) is helping predictions. So again, the overall goal of good predictions already captures the benefits of physically understandable parameters. Also, by keeping the formulation of the relations for prediction restricted to the known physical laws, we do not unnecessarily extend the description length of our total scientific knowledge with extra relations, that are only usable in one specific hydrological system.

It is important to notice that in hydrology, we are always predicting macroscopical variables. In this case intuitive understanding (through "overstanding") can sometimes still be achieved if one realizes that there are analogies with many systems in nature where an "intention" of the system emerges from randomness. Examples are adaptations and optimality that emerge from evolution, free will that apparently emerges from self-organization of processes in the brain, and maximum entropy distributions that emerge from microscopic randomness, deterministic dynamics and macroscopic constraints. Especially under idealized assumptions, the relations between the macro-quantities can sometimes be of the same form as some of the relations between micro-quantities, which even enhances the feeling of understanding.

However, it is dangerous to generalize this kind of understanding. Heterogeneity, for example, can completely change the relation between the macro-states into a form that

has no relation with similar processes on micro-scale. Fitting a model structure that still assumes that the form of the relation is "understandable" in terms of simple mechanics will yield poor predictions and thus is poor science.

Also from a more practical point of view, we could ask ourselves what the use of understanding, models and theories is. It could also be posed that predictions are the fundamental goal and understanding, explanation, theories and models are just the means to achieve good predictions. Understanding could be seen as an emergent goal from the overall objective of good predictions[6]. Especially when there is a clear decision problem associated with a modeling exercise, good predictions are clearly the objective. The question could then be asked why we would want to use information and probability as an objective in model inference rather than directly try to optimize the utility associated with the decision problem at hand. This question is addressed in the next section.

6.6 Modeling for decisions: understanding versus utility

6.6.1 Information versus utility as calibration objective

Utility-based evaluation of predictions is inevitably connected to a particular user with a decision problem and therefore cannot be done without explicit consideration of the different users of those predictions. Moreover, an obvious question that arises is whether it is desirable to base the evaluation on the value to a particular user or group of users. In that case, the evaluation becomes an evaluation of the decisions of those users rather than of the predictions themselves or of the hydrological model that produced them. The difference between information and utility is analogous to that between quality and value, as treated in Murphy (1993) and chapter 5. This difference is particularly important if the results of the evaluation are used in a learning or calibration process. In that case, two effects can occur by using utility (i.e. value) instead of information as a calibration objective:

- The model learns from information that is not there (treated in Sect. 6.6.2).
- The model fails to learn from all information that *is* there (treated in Sect. 6.6.3).

6.6.2 Locality and philosophy of science: knowledge from observation

In the data compression-view of science, knowledge consists of compressed observations. Knowledge should thus come from observations. Furthermore we saw that predictions about macroscopical variables in complex systems should in principle be probabilistic. This leads to certain requirements on the way these predictions are evaluated in a calibration process. Two fundamental requirements are propriety and locality. Especially the latter is closely linked to the difference between information and utility.

Locality is a property of scores for probabilistic forecasts (Mason, 2008; Benedetti, 2010). A score is said to be local if the score only depends on the probability assigned to (a small

6. and the objective of making good predictions emerges from evolution

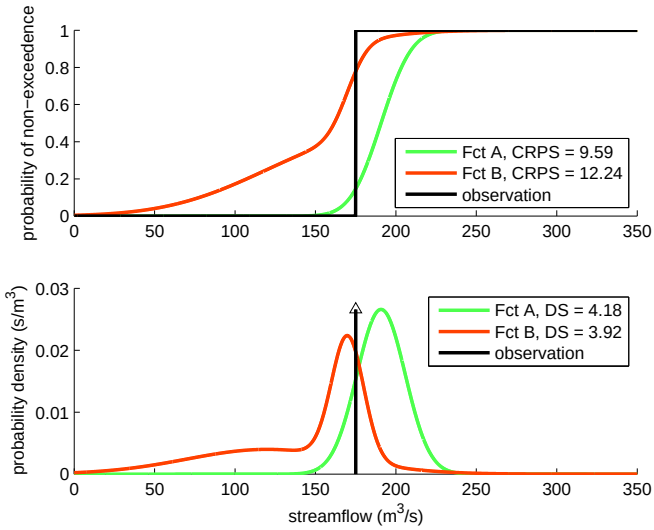


Figure 6.5: The RPS and CRPS scores measure the sum of squared differences in CDFs. Therefore they depend on probabilities assigned to events that where not observed. The divergence score only depends on the value of the PDF (the slope of the CDF) at the value of the observation. In the example, forecast A has a better (=lower) CRPS than forecast B, even though it assigned a lower probability to what was observed (resulting in a higher (=worse) DS).

region around) the event that occurred, and does not depend on how the probability is spread out over the values that did not occur. In contrast to this, non-local scores *do* depend on how that probability is spread out. Usually non-local scores are required to be sensitive to distance, which means that probability attached to values far from the observed value is punished more heavily than forecast probability that was assigned to values close to the observation. This concept of distance only plays a role in forecasts of continuous and ordinal discrete predictands: for nominal forecasts, distance is undefined. For both these types of predictands, an extension of the Brier score exists: the Ranked Probability Score (RPS) and the continuous RPS (CRPS) (see Laio and Tamea, 2007 for description and references). Both these scores are non-local, while the divergence score is local.

Fig. 6.5 shows a comparison between (non-local) CRPS and the (local) divergence score. Note that forecast B obtains a worse CRPS than forecast A, even though B gives a higher probability to what is actually observed. It can also be imagined how changes in the distribution of the lower tail of forecast B would affect the CRPS, although based on the observation no statements can be made about the merit of that redistribution of probability. Note that any preference between two forecasts that assign equal probabilities to the observed value must be based on prior information, e.g. the fact that a bimodal distribution is counter-intuitive. It is important, however, that this prior information should be included in the forecast, rather than adding it implicitly during the evaluation process.

For most decision problems, expected utility is a non-local score: a reservoir operator that attached most probability to values far from the true inflow is worse off than one that

used a forecast with most probability close to the true value, even if the probability (density) attached to the true value was the same. Therefore, non-local scores are sometimes considered to have more intuitive appeal than local scores. It might seem logical to train a forecasting model to maximize the user-specific utility it yields for the training data, which may be a non-local function.

There is, however, a serious philosophical problem with non-local scores if used in a learning (i.e. calibration) process. In principle, the knowledge a model embodies comes from observations or prior information, which in the end also comes from observation; see Fig. 6.6. By calibrating a model, the information in observations is merged with prior information, through a feedback of the objective function value to the search process (the arrows from "EVAL" to the model in Fig. 6.6). It is therefore a violation of scientific logic if the score that is intended to evaluate the quality of forecasts depends on what is stated about things that are not observed. Changes in the objective function would cause the model to learn something from an evaluation of what is stated about a non-observed event. In an extreme case, two series that forecast the same probabilities for all events that were observed, can obtain different scores based only on differences in probabilities assigned to events that were never observed (Benedetti, 2010). A similar argument in the context of experimental design was made by Bernardo (1979). If these non-local scores are used as objectives in calibration or inference (see e.g. Gneiting et al., 2005), things are inferred from non-observed outcomes, i.e. information that is not present in the observations.

6.6.3 Utility as a data filter

The use of utility in calibration can, apart from using non-existing information, also lead to learning only from part of the information that is in the observations. In that sense, the decision problem that specifies the utility acts like a filter on the information. The information-theoretical data processing inequality tells us that this filter can only decrease information (see Cover and Thomas, 2006). This filter can affect two of the three information flows to the model, depicted in Fig. 6.6: the flow from the output (1) and from the input (2) observations.

The first flow of information, from the observations of streamflow, is filtered by the "state of world" block in Fig. 6.6. By evaluating based on utility, the information in the streamflow observations only reaches the model through its effect on how the state of the world affects the utility of decisions based on the forecast. Figure 6.6 depicts a hypothetical binary evacuation decision that is coupled to a conceptual rainfall-runoff model for flood forecasting. In this simplified decision problem, the utility is only influenced by a binary decision (evacuate or not) and a binary outcome (the place floods or not). There are thus no gradations in severity of the floods that affect the damage. The calibration towards maximum utility for this decision problem will train the hydrological model to optimally distinguish flood-evacuation events. This implies that in the training, all that the hydrological model sees from the continuous observed discharges is a binary signal: flood or no flood. This constitutes at most one bit of information per observation, in the unlikely case that 50% of the observations is above the flood threshold, i.e. the climatic uncertainty is

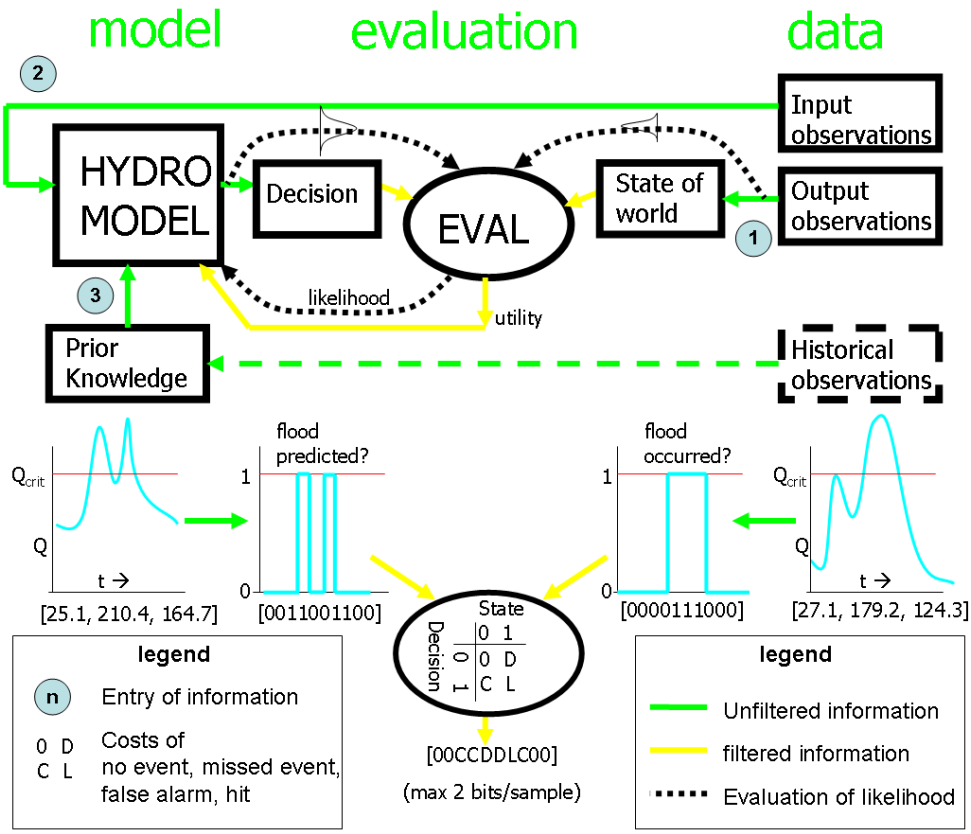


Figure 6.6: There are three routes through which information can enter the model in a learning process: the output observations (1), the input observations (2) and prior information (3). When evaluating a model based on value, the decision model that is implicitly defined by the loss function acts as a filter on the information in the observations. The figure shows the case where both the decision and the state of the world are binary, resulting in a feedback of costs to the model of only 2 bits of information per input-output observation pair (the string of characters at the bottom of the figure). The graphs in the middle show how both the predicted and the measured flows are converted to binary sequences due to the way the cost-loss model is formulated.

1 bit, while the original signal, the observed flows, i.e. real numbers, contained far more information (see Fig. 6.6).

The second flow of information to the model are the input observations. A typical example is the amount of information from rainfall observations, see Bárdossy and Das (2008), which can have influence of model performance. The amount of information that reaches the model is affected by the information filter in the "decision" block. For example, if a binary decision problem, e.g. to be or not to be in the flood zone tomorrow, is considered, the information from input observations travels through the model and subsequently through the decision model. While the model still gives a real number as output, the "decision" block maps that model output to a binary signal. The binary signal is all that enters the evaluation and can be learned from the input observations. When a model is evaluated based on a cost-loss model of a two action- two state of the world decision problem, the maximum amount of information that can be learned from each input-output observation pair is thus 2 bits. In Fig. 6.6, this information is contained in the

string "00CCDDLC00", which represents the sequence of utilities over all time steps.

The hydrological model will therefore have far less information to learn from. Given the fact that there is a balance between the available information for calibration and the complexity that a model is allowed to have (see Schoups et al., 2008), hydrological models that are trained on user-specific utility functions (e.g. this binary one) are likely to become overly complex relative to the data. They will surely achieve better utility results on the calibration data, because there is less information to fit, but are likely to perform worse on an independent validation dataset. The model that has been trained with maximum information as an objective is likely to yield better results for the validation set, even in terms of utility. Because it has the unfiltered information from the observations to learn from, it is less prone to overfitting: the complexity of a conceptual hydrological model is better warranted by the full information. The objective of optimally predicting binary flood events for evacuation decisions could benefit from more parsimonious data-driven models, e.g. linear regression models or neural networks; see Solomatine and Ostfeld (2008) for an overview. These models can make a mapping directly from predictors (e.g. precipitation, snowpack, soil moisture, past discharge) to decisions, but this complicates the use of prior information.

The third information flow in Fig. 6.6 consists of this prior information on the workings of the hydrological system, which can be valuable for improving forecasts. The information can enter in the form of prior parameter estimates, e.g. from regionalization (Bárdossy, 2007; Hundecha et al., 2008), or constraints that are captured in the model structure. Examples are constraints on mass balance and energy limits for evaporation. These constraints describe the patterns in data or "physical laws" that ultimately come from observations. Both adding too much (unwarranted assumptions) and too little (e.g too wide prior parameter distributions) information through this route deteriorates the forecasts, especially when little data are available.

The framework presented in this section shows some similarity with the ideas presented in Gupta et al. (2009, 1998, 2008). In those papers it is also argued that information can be lost in the evaluation. However, the important difference of the framework presented in this chapter compared to those ideas is that we argue that information is lost by using measures other than information (in other words, measures that do not reflect likelihood), while Gupta et al. (2008) argue that information is lost because of the low dimensionality of the evaluation measure. In our information-theoretical viewpoint, we can in principle learn all we need from the observations through a single measure, because a real number can contain infinitely many bits of information. What is learned depends only on the data and the prior information. The challenge is to give a reliable representation of prior information which will result in the right likelihood function. In principle, this is equivalent to endorsing the likelihood principle, which states that all information that the data contains about a model is in the likelihood function (as argued by Robert (2007) p.14, Jaynes (1957), p.250 and Berger and Wolpert (1988).

The divergence score corresponds to a logarithmic scoring rule (see Jose et al. (2008) for more context), which is the only scoring rule that is both local and proper (proofs can be

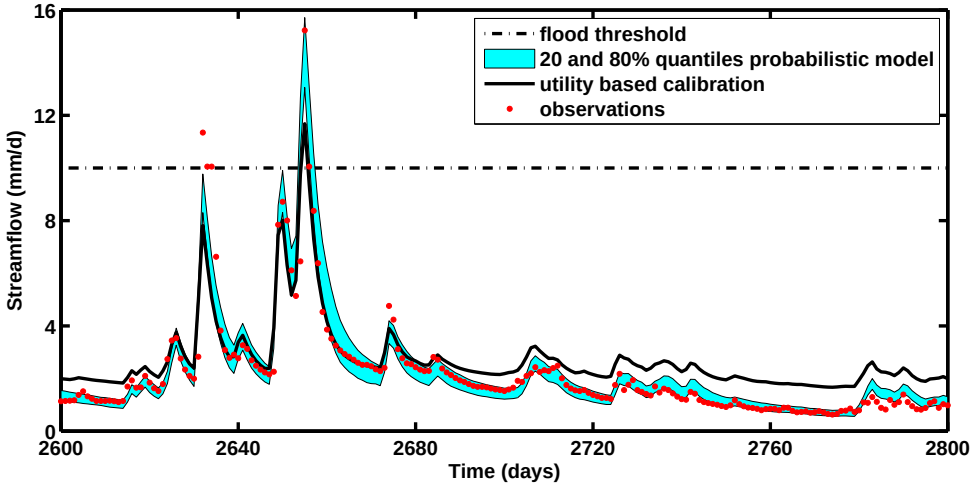


Figure 6.7: A detail of the validation results for both models, compared with observations. For the probabilistic model, the 20% and 80% quantiles are shown. With the cost-loss ratio of 0.2 used in this example, the upper quantile determines the decision for the probabilistic model. Note that the probabilistic model predicts the full distribution, the quantiles are just for visualization.

found in Bernardo, 1979 and Benedetti, 2010), where propriety is the requirement that the scoring rule can only be optimized when the forecaster does not lie. Scoring rules that are not proper can be hedged, meaning that the expected score is maximized by forecasting probabilities that are not consistent with the best estimates of the forecaster (see Gneiting and Raftery (2007) for an elaborate discussion on proper scoring rules). A utility function that includes the importance of the outcomes can be hedged by attaching more forecast probability to important events. A model that is trained on such a measure is thus encouraged to "lie". All utility functions that are not affine functions of information violate either locality or propriety, which makes them doubtful objectives for calibration.

6.6.4 Practical example

As an illustration of the information-filter effect described in section 6.6.3, a hydrological model was calibrated both based on information and on a utility function relating to the binary decision scenario similar to that depicted in Fig. 6.6. A simple lumped conceptual rainfall-runoff model was used (Schoups et al., 2010) to simulate daily streamflow given daily forcing records of rainfall and evaporation from the French Broad River basin at Asheville, North Carolina. In order to evaluate the model for unseen data, it was calibrated using 1 year of streamflow observations (1961), and validated using 9 years of streamflow observations (1970-1978).

The calibration on the information-uncertainty scale used minimization of the divergence score (i.e. remaining uncertainty) as an objective. In the continuous case this corresponds to maximizing the log-likelihood. This means that the model needs to provide explicit probabilistic forecasts. The probabilistic part used a flexible stochastic description, allowing for heteroscedasticity, autocorrelation and non-Gaussian distributions. The calibration relied on the general likelihood function presented in Schoups and Vrugt (2010).

Calibration objective	Result in calibration	Results in validation
min average cost	1.6	2.47
min divergence score	3.8	2.29

Table 6.5: The resulting average disutility per year, composed of costs for action and losses for unpredicted events, is minimized by explicitly calibrating on it, but performance in the validation period is better for the probabilistic model trained to minimize remaining uncertainty.

The calibration on the utility-risk scale employed a cost-loss utility function relating to the binary decision problem (Murphy, 1977). The flood threshold is defined at a value of 10 mm/d (streamflow divided by catchment area). Here, a cost C is associated with a precautionary action, which is taken if exceedence of the flood threshold is forecast. When a peak flow event occurs but was not predicted, a loss L occurs. For illustration purposes, values for C and L were chosen to be equal to 0.2 and 1.0, respectively.

The results in Table 6.5 show that in the validation run for this case, we indeed find that the explicit probabilistic model trained to minimize remaining uncertainty outperformed the model trained on maximum utility for the specific decision problem at hand. As expected, the large deterioration from calibration to validation seems to suggest overfitting to the filtered information. Looking at the resulting model behavior in Fig 6.7, we can tell that the model trained on utility systematically overpredicts low flows. There is nothing in the evaluation that discourages this behavior and apparently these parameters gave an advantage in fitting the floods in the calibration year. The probabilistic model is encouraged to attach high likelihood to each observation, learning from all data in the calibration. In this case, this gave an advantage in predicting the exceedence probability for the flood threshold for the unseen validation data.

We must note, however, that results might be different under less ideal conditions. For example, when the model structure is capable of representing high flows, but is inadequate for low flow situations because e.g. evaporation is not correctly represented, then the utility-based calibration might do better also in validation. We can explain this by seeing the utility function as an implicit way to add prior information. If we know a priori that the model structure misses relevant processes for low flow, then it could be reasonable to ignore the low-flow data in calibration. The analogous way to represent this in an explicit uncertainty model is to give an extra spread to the probabilistic predictions at low flows, making the model less sensitive to them. More elaborate case studies are needed to further investigate which practical factors might lead to different results and how they can be accounted for in the information-theoretical framework. Furthermore, applying this view on results in past literature, especially those relating to “informal” likelihood methods, might give new insights about prior information that is implicitly added.

6.7 Conclusions and recommendations for modeling practice

A hydrological model can be seen as a tool for prediction, a theory, a hypothesis or a compact form to code observations. In all cases, it consists of mathematical (i.e. algorithmic)

relations that represent analogies to quantities in the postulated real world. Models that concern observable quantities can be tested. When models approximate emergent behavior from macroscopic systems, such as in hydrology, predictions can never be perfect. In chapter 5 it was argued that correspondence between model predictions and observations is only testable without additional assumptions if the predictions are probabilistic. By adapting the model to mimic the observed data, the likelihood of the model can be increased. Overly complex models increase likelihood but decrease prior probability, yielding less probable predictions.

Bayesian probability theory can serve as the basis for a philosophy of science, see e.g. Jaynes (2003), and can replace limited or imprecise concepts such as verification, and falsification and corroboration advocated by Popper (1968). Knowledge of formal theories of algorithmic information theory could help focus debates about uncertainty analysis in hydrology. Algorithmic information theory provides the link between complexity and probability needed to intuitively justify and to formalize the principle of parsimony. Solomonoff's formal theory of inductive inference offers a complete formalization of prediction from data. It combines Bayesian probability, a universal prior probability for models reflecting the principle of parsimony, Turing's theory of computation, and the simultaneous use of all possible models.

The fact that the universal prediction by Solomonoff induction is incomputable is not just a practical problem of the method, but stems from the fundamental limits of computation and provability as found by Turing (1937) and Gödel (1931), which apply to any method of induction. Due to this incomputability, not only perfect predictions, but also optimal probabilistic predictions are in principle not achievable. Several practical methods for prediction can be seen as computable approximations to Solomonoff induction. The principle of minimum description length (MDL; see Rissanen (2007); Grünwald (2007)) which can be approximated by data compression methods, may be a useful practical approach to learning from data. Given that inference and data compression are analogous tasks, compression algorithms may be useful to estimate the information available for learning and the merit of a model in terms of compression progress. Conversely, good models can be used to efficiently store bulky hydrological data.

Understanding and explanation are not explicit building blocks of the information-theoretical view on science. Rather, they emerge from the objective of finding short descriptions by the fact that analogies can be used to compress the description of scientific knowledge. Understanding may be a practical requirement for the application of science, but fundamentally science is served by making good predictions, i.e. finding short descriptions of the totality of observations. When understanding is seen as a requirement for models of complex systems, this might lead to a flawed view of how the system works and will not yield good predictions.

Calibration of models is a way to find parameter sets that, together with a given model structure, minimize some discrepancy measure between model outcomes and observations. This discrepancy measure can represent a user-specific utility for a given decision problem, or can be interpreted as a likelihood or information measure that implicitly specifies

a probabilistic prediction. The likelihood principle states that all information the data contain about a model, is in the likelihood function of that model. Training on a user-specific objective function can therefore never increase the amount of information that can be learned from the observations. Apart from decreasing information by filtering it through the utility function, calibration based on utility, if it is a non-local measure, also lets the model learn from information that is not in the observations, i.e. non-existing information. This makes model calibration with objective functions that reflect user-specific utilities doubtful. Although the purpose of a model may influence its design or related data collection strategy, it should not influence its calibration. A model should learn from observation data and not from decisions based on its predictions.

Bayesian logic and the likelihood principle also dictate that the model, which defines the likelihood function, should be specified a priori, before seeing the data against which the model is tested. This seems to conflict with the idea that we can do "model diagnostics" to learn about the true behavior of the system. For example, the stepwise improvement of a model concept by repeatedly looking at the results and adding components based on missing processes, without introducing new data, introduces a high risk of overfitting. In fact such an approach used the same information multiple times, because there is no clear separation between prior knowledge and data. The space of possible model structures becomes an important degree of freedom. A Bayesian fundamentalist would then correctly object that we fit a model to the data by forming hypothesis from data that are tested on that same data.

Model complexity control methods can be used to prevent overfitting by balancing model complexity and the amount of information to be learned from the data. To successfully apply these methods, dependencies in the data should be accounted for. Data compression methods may be used to estimate a correction factor for the number of data points, which is an input to some of these methods. Also the use of black box models to mimic the data might be useful to get an idea of the data complexity and therefore the optimal model complexity.

The difference between conceptual, physically based and statistical models is in some sense gradual. Also our physical knowledge ultimately comes from fitting mathematical relations to data. For hydrological models, it is important to include process knowledge, but its certainty should not be overestimated. Although in practice, by attribution of uncertainties to parameters or states, we can still make reasonably good predictions, theoretically it makes little sense to include hypotheses beyond prior knowledge in a single model structure and attribute all uncertainty to parameter values. In principle, all models are not certain and should not be attributed 100% prior probability. One could even say that all model structures of macroscopical systems are wrong, unless they are explicitly probabilistic.

Chapter 7

Stochastic dynamic programming to discover relations between information, time and value of water

> *"Water has an economic value in all its competing uses and should be recognized as an economic good"*
> - IV^{th} Dublin principle on integrated water resources management.

Abstract - This chapter presents stochastic dynamic programming (SDP) as a tool to reveal economic information about managed water resources. An application to the operation of an example hydropower reservoir is presented. SDP explicitly balances the marginal value of water for immediate use and its expected opportunity cost of not having more water available for future use. The result of an SDP analysis is a steady state policy, which gives the optimal decision as a function of the state. A commonly applied form gives the optimal release as a function of the month, current reservoir level and current inflow to the reservoir. The steady state policy can be complemented with a real-time management strategy, that can depend on more real-time information. An information-theoretical perspective is given on how this information influences the value of water, and how to deal with that influence in hydropower reservoir optimization. This results in some conjectures about how the information gain from real-time operation could affect the optimal long term policy. Another issue is the sharing of increased benefits that result from this information gain. It is argued that this should be accounted for in negotiations about an operation policy. Some suggestions for future research involving the technique of reinforcement learning conclude the chapter.

7.1 Introduction

As stated in the principle IV of the Dublin principles of integrated water management "Water has an economic value in all its competing uses and should be recognized as an economic good" (Global Water Partnership, 2000). This report also states "In order to extract the maximum benefits from the available water resources there is a need to change perceptions about water values and to recognize the opportunity costs involved in current allocative patterns." Water derives its value partly from how it is used (see Tilmant et al.

(2008) for a more detailed discussion on the components of the value of water). A different allocation can lead to a different value of water, which also depends on the stakeholders and their preferences. Allocating water optimally corresponds to maximizing its value. Before this maximization, moral questions have to be answered, which are not considered here. Decisions about water resources have to allocate water between different uses, but also in space and time. This chapter focuses on the latter case, where a set of subsequent decisions in time have to be taken.

Most problems of water system operation are sequential decision processes, meaning that a decision in a certain time step is followed by subsequent decisions. For the operation of pumping stations in a polder, for example, every 15 minutes the current situation is reconsidered and a decision made about operating the pump in the next period. In chapter 2, model predictive control (MPC) was used to solve this type of problem. When extending MPC to finding optimal decisions under uncertainty (the multiple model predictive control method, chapter 2), the problem of interdependency between current and future information for decisions arises. Consequently, the amount of information that will become available between the current decision and the future decisions affects the optimal current decision.[1]

Dynamic Programming (DP), which was also shortly reviewed in chapter 2, offers the possibility to disaggregate the decision process in time, resulting in a series of one step decision problems. This circumvents the aforementioned interdependency problem. In this chapter, a special case of dynamic programming for systems affected by uncertainty, Stochastic Dynamic Programming (SDP), will be applied and analyzed from an information-theoretical viewpoint.

7.2 Stochastic dynamic programming (SDP)

Stochastic dynamic programming proceeds by going backwards in time, while recursively calculating the value-to-go function at time t from the value-to-go at time $t+1$. The value-to-go function $F_t^*(x)$ at time step t estimates the benefits that can be made by optimally operating the water system from time step t to the end of the planning horizon, while being in state x. Bellman's principle of optimality (Bellman, 1952) states that any optimal policy that visits a certain state has the property that the remaining decisions from that state to the end of the horizon would also form an optimal policy from that state if it were the starting point. When searching an optimal sequence of T decisions in a sequential decision process, the problem can be split into T independent subproblems. In each subproblem, the value-to-go for each possible state is computed based on the decision that maximizes the expected value-to-go for that state. This value is the sum of the benefits from the transition to the new state by that decision and the value-to-go from the state where it leaves us. Because the problem is stochastic, an expected value

1. Note that we cannot know *which* specific information will become available, but an estimate of the *amount* of information *is* possible.

over several possible future states is taken as value-to-go following a decision (see Fig. 7.1). SDP has been widely applied to the optimization of water resources. See Yakowitz (1982); Yeh (1985); Stedinger et al. (1984); Philbrick and Kitanidis (1999); Labadie (2004) for more references and e.g. Karamouz et al. (2005); Pianosi and Soncini-Sessa (2009); Tilmant et al. (2010) for more recent applications.

7.2.1 Formulation of a typical hydropower reservoir optimization problem using SDP

For hydropower production, reservoir operation is usually analyzed and optimized at a monthly time step, complemented by a daily or hourly operation that uses the results from the monthly time scale. The optimal policy of monthly releases as a function of reservoir level, current hydrological conditions and month of the year can be found using SDP. In a typical reservoir operation SDP problem, the state variables are discretized into intervals represented by characteristic values. For a single-reservoir problem, a two dimensional state is usually chosen, consisting of the current reservoir storage level and current inflow to the reservoir, which is used as a characterization of the hydrological condition (Tejada-Guibert et al., 1995). This means that for the monthly flows, it is assumed that all information that is available about the next flows is present in the current flow, which will be known at the end of the current period. The streamflow persistence is assumed to be dominated by lag-1 autocorrelation. This is called a first order Markov (Markov-1) process. Note that this assumption may be problematic when the streamflow dependence structure has more persistent characteristics, as often found in nature (Koutsoyiannis, 2005b), leading to for example undersizing of reservoirs (Hurst, 1951). Using more state variables in the Markov description may be a remedy, but leads to higher computation power requirements.

With the Markov-1 assumption, the equation for the value-to-go function in SDP, which is solved backwards in time is

$$F_t^*(S_t, Q_t) = \max_{R_t}\{(B(S_t, R_t, Q_t) + \underset{Q_{t+1}|Q_t}{E}[F_{t+1}^*(S_{t+1}(S_t, Q_t, R_t), Q_{t+1})])\} \tag{7.1}$$

subject to:

$$S_{t+1} = S_t + Q_t - R_t - L_t \tag{7.2}$$

$$S_{t+1,min} \leq S_{t+1} \leq S_{t+1,max} \tag{7.3}$$

$$R_{t,min} \leq R_t \leq R_{t,max} \tag{7.4}$$

where the asterisk stands for optimal. The optimal value-to-go function F_{t+1}^* in time step $t+1$ is a lookup table in the discrete case, stored in the previous calculation step, where for every time period, the value-to-go is given for each given combination of the two states: storage S_t and current flow Q_t. R_t is the release that is to be optimized, L_t is the are the spills, and B is the function for the immediate benefits in the current time step. $E_{Q_{t+1}|Q_t}$ denotes the expectation operator with respect to the conditional distribution of the next

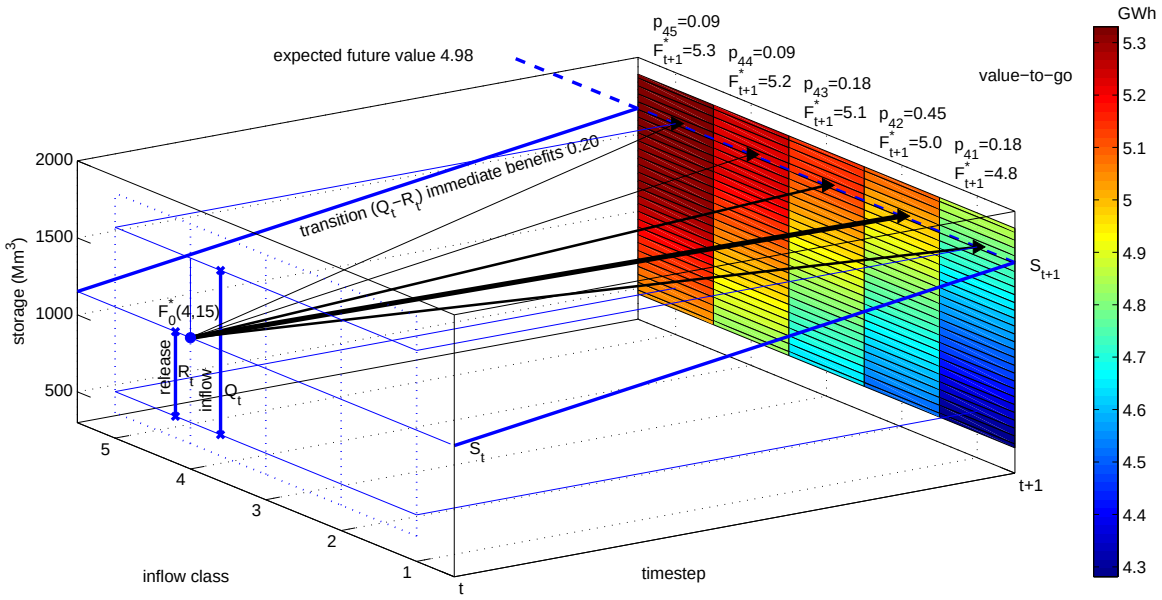


Figure 7.1: For each state, the value-to-go is computed from the optimal sum of the expected future value-to-go and the immediate benefits. The result is shown for the example reservoir used in this chapter, for the decision problem in May, with initial state corresponding to the top left of Fig. 7.5, the value-to-go is in GWh until the end of the planning period.

period's inflow, given the current. Apart from this value-to-go function at every time step, solving the optimization also yields the optimal release function R_t^* at every time step.

$$R_t^*(S_t, Q_t) = \arg\max_{R_t}\{(B(S_t, R_t, Q_t) + \underset{Q_{t+1}|Q_t}{E}[F_{t+1}^*(S_{t+1}(S_t, Q_t, R_t), Q_{t+1})])\} \qquad (7.5)$$

when solved backward until the current time step t, the optimal release R_t^* is known for the current state S_t,Q_t and can be executed. For finding a steady state policy, the policy is iterated until the yearly increase in the value-to-go function converges to a constant value for all states.

The model of the reservoir system is captured in the constraints (Eq.7.2-7.4). Next to minimum and maximum releases (Eq. 7.4) and reservoir levels (Eq. 7.3), the constraints include the mass-balance equation for the reservoir (Eq. 7.2). When a system of reservoirs is optimized, the states become vectors and a connectivity matrix can be used to specify the layout of the reservoir system (see e.g. Tilmant et al., 2007).

7.2.2 Computational burden versus information loss

Due to the discretization, F_t^* has to be computed for every combination of discrete values of the state that might be visited by the reservoir system. This has to be done for every time period. Therefore, the number of evaluations of the objective function is proportional to $T \times (N_S^{n_S} \times N_Q^{n_Q})$, where T is the number of time periods within the horizon, N_S and

N_Q are the number of discrete values for the inflow and the storage and n_S and n_Q are the number of reservoirs and the number of memory states in the model of inflows to the system. This gives rise to the "curse of dimensionality", which makes the classical SDP approach computationally intractable for systems with more than 3 or 4 reservoirs.

Ways to overcome this curse are an important topic in water resources optimization literature. One solution is aggregation of multiple reservoirs into one representative reservoir, but this can of course lead to a severe loss of information about the spatial distribution of storage and to reduction in performance of the control policy. An other method is the approximation of the value-to-go with a function. Because the function can be interpolated, less discrete points are needed in the representation or the parameters can be stored. A number of references for such approaches, e.g. cubic Hermite polynomials (Foufoula-Georgiou and Kitanidis, 1988), splines (Johnson et al., 1993), and neural networks (Bertsekas and Tsitsiklis, 1996) are given by Pianosi (2008). Pianosi also discusses other strategies to combat dimensionality such as reducing the model state in an online approach (Pianosi and Soncini-Sessa, 2009). Another promising direction is Stochastic Dual Dynamic Programming (SDDP), (Pereira and Pinto, 1991) which approximates the value-to-go function by piecewise linear functions around the states actually visited by the system. The piecewise linear nature of the value-to-go function allows formulating the dual problem to approach the value-to-go from two sides, converging to the true value. A problem of the approach is that the immediate benefit function is limited to a linear function of the state. A large advantage is that systems with a much larger state dimension can be optimized (see e.g. Tilmant and Kelman (2007), where a system with 20 reservoirs is optimized).

Naturally, all these simplifications have some negative impact on the performance. In this chapter, an information-theoretical view is given on some consequences of assuming the inflow is a Markov process, which is one of the most common assumptions in a classical SDP formulation. It is argued that in general, the real-time operation of a reservoir can use more short term information than the long term optimization supposes. This leads to an underestimation of the value of water in the long term planning. When this underestimation would depend on the storage level, also the marginal future value of water and therefore the optimal policy would change. This would mean that an estimate of the future information gain is needed to find an optimal policy. A practical demonstration of this effect would be an interesting topic for further research.

7.3 Example case description

For the research described in this chapter, a toy reservoir model was constructed as an example. The model has the dimensions of Ross-lake in the Columbia river basin (WA, USA). In the current status, the model is not aimed at improving reservoir operation, but serves as an example with realistic dimensions. For the sake of transparency, it is assumed that maximizing total hydropower is the only objective and there is no ecological flow requirement or economical discount factor. In reality, economical objective functions

Parameter	Description	Value	Unit
S_{max}	Maximum storage in the reservoir	1890	Mm^3
S_{min}	Dead storage of the reservoir	464	Mm^3
h_{min}	Minimum head on turbines	82	m
R_{max}	Maximum release through generators	453	m^3/s
$R_{max} + L_{max}$	Maximum release including spills	2400	m^3/s
R_{min}	Minimum release (environmental flow)	0	m^3/s

Table 7.1: Characteristics of the hydropower reservoir toy model.

will be more complex, but this would not contribute to the example. Table 7.1 gives an overview of some of the most important characteristics of the reservoir and figure 7.2 gives the reservoir area and the head on the turbines as a function of the storage, along with the characteristic storage values of the discretization. The storage volume was subdivided in 28 intervals and the midpoints of those intervals, along with the minimum and maximum storage, give the 30 discrete storage values. This method of discretization is known as the Savarenskiy scheme and was shown to have advantages over other discretization schemes (Klemeš, 1977).

For the inflows to the reservoir, 55 years of modeled inflows from the hindcast dataset for Columbia river were used (see chapter 4). The advantage of this data set is that it provides a complete record and represents the natural flows, without the influence of releases from upstream reservoir operations. Subsequently, for each month of the year, the inflows were discretized and the transition probabilities for the Markov-1 description were calculated from observed frequencies. The flow was discretized into 5 separate classes, where the class boundaries were chosen to represent the quantiles of 20, 40, 60 and 80%. This resulted in 5 equiprobable intervals, for which the class boundaries vary for each month. In this way, the a priori distributions have maximum entropy and knowledge of the flow class gives maximum information for the given number of classes. For each discrete class in each month, the expected value of all historical flows falling within that class is used as the representative value in the SDP optimization. For the transition probabilities used in the following analysis, empirical transition matrix was chosen rather than one based on a parametric distribution. Because no assumptions on the distribution type are made, data scarcity affects the estimation of transition probabilities. This problem is aggravated when Markov models with more states. The difficulty of estimating transition probabilities from limited data is also referred to as the "curse of modeling" (Castelletti et al., 2010).

7.3.1 Simulation and re-optimization

After deriving the optimal policy, i.e. value-to-go functions, with SDP based on the Markov-1 representation, the reservoir operation was simulated using the original, non-discretized, inflows of the dataset. In the re optimization, a one step optimization problem is solved every time step, optimizing the sum of current benefits and the future value-to-go, which is read from the tables stored by the backward optimization. This amounts to

solving the optimization problem in equation 7.5. Instead of the representative values for storage and inflow, the exact values from the current state are used during re optimization. The value-to-go from the end of period state is found by linear interpolation of the stored table (see Fig. 7.1).

For this optimization, the Matlab function "fmincon" was used. This has the advantage that next to the optimal solution and the value of the objective function, it also gives the Lagrange multipliers for all constraints as an output. These are the derivatives of the objective function with respect to the value of the constraint. For example, the Lagrange multiplier for the mass balance constraint gives the improvement in objective function value as a result of relaxing the constraint with one unit (i.e. adding one m^3 of water to the system). If the objective is stated in monetary terms, the Lagrange multiplier can be interpreted as the shadow price of water. In absence of a real market, this is the imaginary price that one of the users would want to pay for one m^3 of water in the current conditions and allocation (Tilmant et al., 2008). In the next section, other theoretical links between SDP results and water resources economics are given.

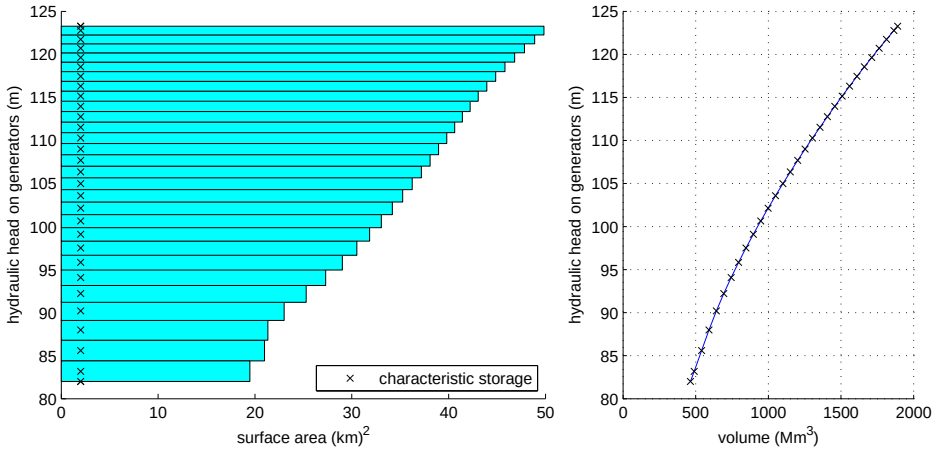


Figure 7.2: The storage volume-area-head relation for the reservoir and the storage discretization using the Savarenskiy scheme.

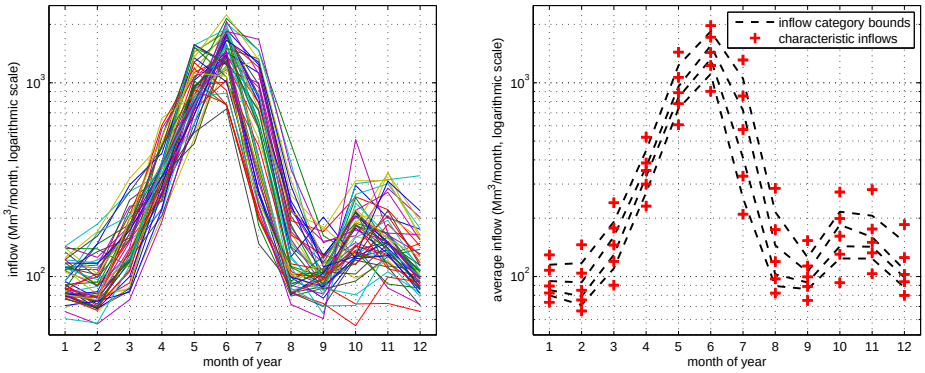


Figure 7.3: The discretization of the monthly flows (left) into 5 equiprobable classes for each month (right).

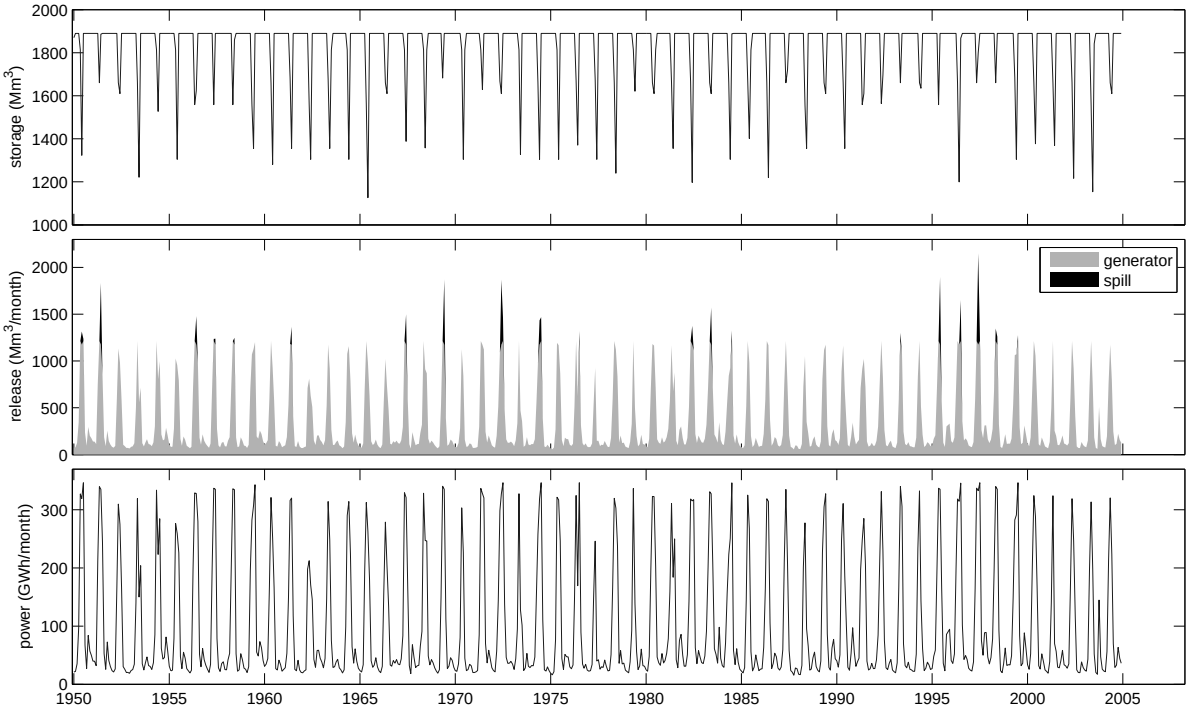


Figure 7.4: Reservoir behavior during the re-optimization. Because no firm energy production target was set in this case, energy production drops during the dry season. The reservoir is constantly kept full, unless future spills become likely.

7.4 Optimization and the value of water

In SDP, the optimal decision is found by maximizing the sum of the current and the future benefits that can be obtained from the water system. Due to the dependence of hydro-electric power generation on the reservoir level, not every unit of water has the same value, which leads to non-linearities in the optimization problem. One m^3 of water released at a higher reservoir elevation yields a higher immediate benefit. Also the value per unit of water stored for next time steps depends on the current storage level. If the reservoir is almost full, every extra unit that is stored is increasingly likely to be spilled in the future. Therefore, the marginal value of storage (the benefits of one extra m^3 stored) decreases with increasing storage for a near full reservoir in the wet season.

Both the current and the expected future benefits also depend on the current release. To maximize the total benefits, given by the value-to-go F_t^* in equation 7.1, the derivative of F_t^* with respect to the release R_t must be zero (assuming that the maximum is not constrained by R_{min} or R_{max}), so we can write

$$\frac{\partial F_t}{\partial R_t} = 0 \tag{7.6}$$

$$\frac{\partial}{\partial R_t}\left\{B(S_t, R_t, Q_t)\right\} = -\frac{\partial}{\partial R_t}\left\{\underset{Q_{t+1}|Q_t}{E}[F_{t+1}^*(S_{t+1}, Q_{t+1})])\right\} \tag{7.7}$$

The left hand side of equation 7.7 is the marginal value of using water in the current period (which is assumed to be certain). The right hand side represents the marginal value of saving water for the future, or the opportunity cost of using the water in the current period. This marginal value is the derivative of the expected future value of the water in storage.

At the optimal release, the value of one m^3 of water in the reservoir is thus equal to the derivative of the value-to-go with respect to the release (see Fig. 7.5). Also, the immediate value of using one extra m^3 of water is then equal to the opportunity cost of not having that m^3 of water for the future. Furthermore, the marginal value is equal to the Lagrange multiplier for the mass-balance constraint in the re-optimization (Tilmant et al., 2008). Optimization can thus provide interesting information about water resource economics, as it gives insight in the trade-offs between different water uses and different time-periods. See, for example the analysis in Tilmant et al. (2008) about a large reservoir system in Turkey. Conversely, it can also be observed that the expected future value of water determines the optimal release. An accurate estimate of this future value requires an accurate model of the future decisions and therefore an accurate model of future information availability for those decisions.

7.4.1 Water value in the example problem

For the example reservoir, an SDP optimization was performed to determine the value-to-go functions and optimal release policy. Analysis of the value-to-go function and the Lagrange multipliers of the mass balance in a simulation of the re-optimization revealed interesting information about the marginal value of water as a function of time, current reservoir level and current inflow. In this example, only the objective of energy production was considered. Therefore, the marginal value of water can be expressed in terms of the energy produced per m^3 of water, e.g. Wh/m^3, thereby avoiding the need to make a price assumption. This focuses the analysis on the hydrological influence on water value and rules out the influences of the socio-economical system connected to the water resource. In a real market situation, the price per kWh is also dependent on supply and demand and may vary with the season. This may significantly influence release decisions and can even lead to typically anthropogenic weekly cycles in downstream river flows.

The high inflows in late spring and summer lead to a decrease in the marginal value or shadow price of water (see Fig. 7.6). Because the release is already at a maximum and the reservoir will most likely be replenished by July, extra water received in June is likely to be spilled in the future, which decreases the expected future value. In several instances (more than 25% of the years in June; see Fig. 7.6), the shadow price even drops to zero, corresponding to an inevitable future spill. In terms of the Lagrange multipliers this means that relaxing the mass balance constraint with one m^3 of water (adding one m^3 to the system) will not give any extra benefits, because the generators already operate at full capacity and even with this release, the reservoir will be full at the end of the period. The extra water therefore can not lead to extra immediate or future benefits and will be spilled. In March, just before the melting season, however, water is still scarce and the

value of water is high. Any extra water is directly used to spin the turbines and contribute to power production, while being at the highest reservoir elevation level.

The value of water depends on the end of period storage in the reservoir, but also on what inflow is expected in the further future. For the Markov-1 model of inflow persistence, the current inflow class determines the estimated distribution for the next period's inflow and therefore can have an effect on the value of storage. In figure 7.8, the marginal value of storing water as a function of storage level is plotted for all months. The values are the derivatives of the value-to-go functions resulting from the SDP optimization. The five lines correspond to the five inflow classes illustrated in Fig. 7.3, where 1 is the lowest inflow class.

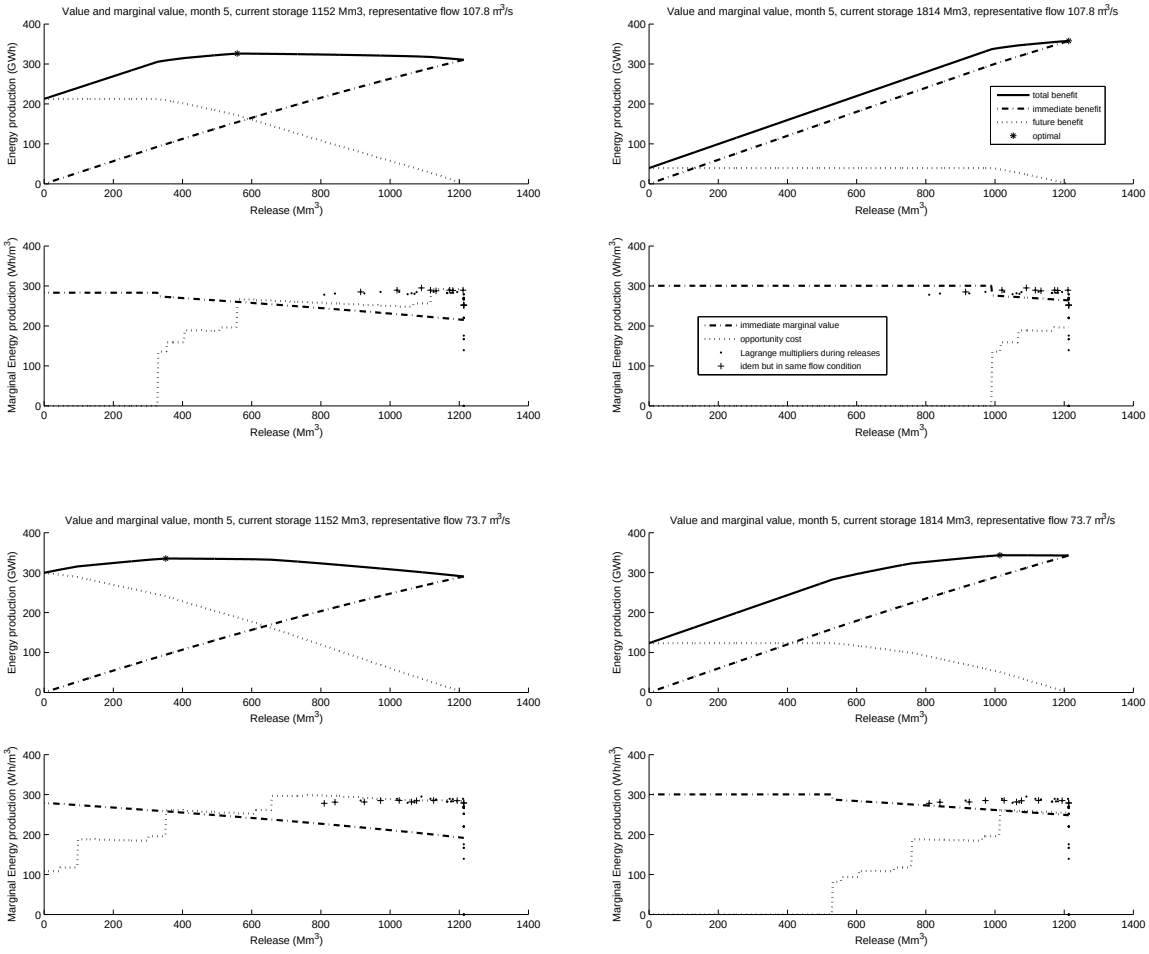


Figure 7.5: The value and opportunity cost of releases and marginal value of water for combinations of high inflow (top), low inflow (bottom), low storage (left) and high storage (right). The value curves are shifted to depict the extra value relative to the minimum.

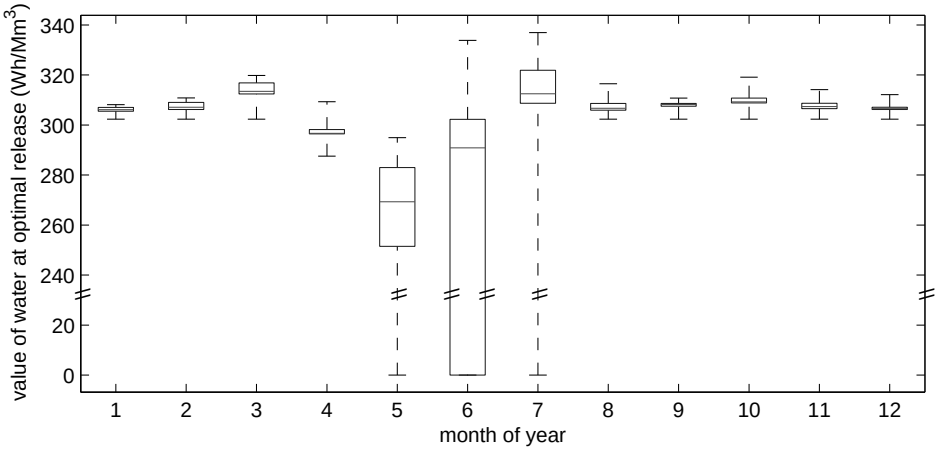


Figure 7.6: The value of water at the release decisions found by re-optimization during simulation, as function of the month of the year. The box plots show the median, quartiles and minimum and maximum values. The abundance of water in the melting season decreases its marginal value.

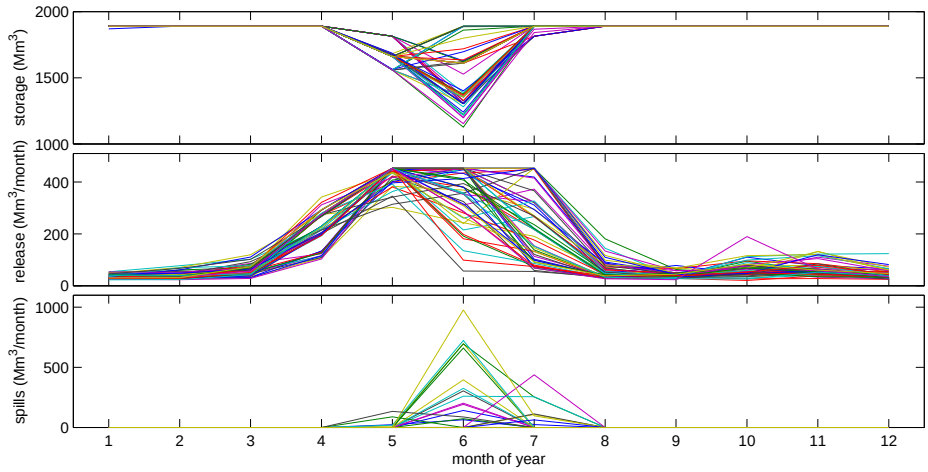


Figure 7.7: The reservoir behavior as a function of the month of the year.

7.5 Interdependence of steady state solution and real-time control

Real time operation of reservoirs makes use of actual information about inflows at a high temporal and spatial resolution in the near future, using information in high-dimensional state vectors of for example distributed hydrological models. In contrast, the long-term optimal steady state policy relies on a Markov-1 description of inflow uncertainty and persistence at a coarser time-resolution. However, the short and the long term optimization are not independent. Firstly, the future value function as function of the state, which can be obtained from the long term optimization, might be needed as a boundary condition at the end of the time horizon for the short-term optimization. Secondly, improvements in operation due to the use of actual information in real time operation, as compared to a steady state policy based on just climatic information, need to be reflected in the rewards

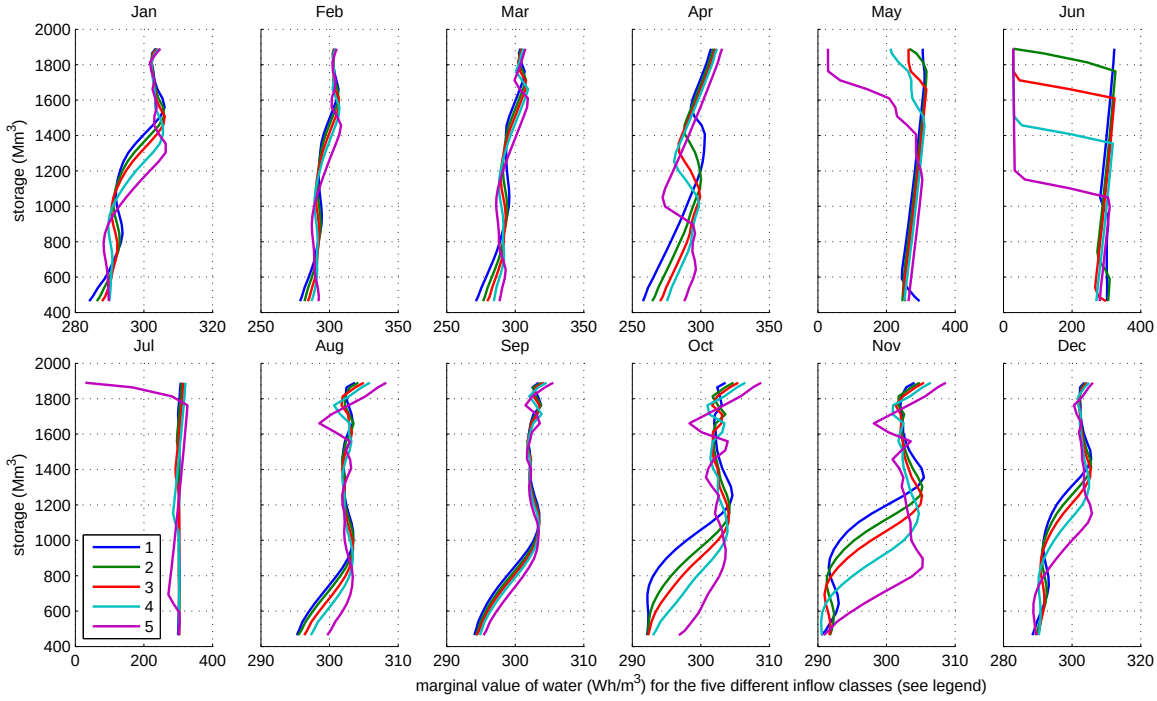


Figure 7.8: The marginal value of water as a function of the end of period storage. The value is affected by the current inflow through the inflow persistence. This is accounted for by the different lines for the inflow classes 1 (low flow) to 5 (high flow) depicted in figure 7.3. Usually the marginal value increases with storage due to higher head on the generators, but for an overly filled reservoir, the marginal value decreases due to the increasing probability of future spills, especially in the melting season.

attributed to actions in the long term optimization.

The first influence, of the long-term on the short-term operations, takes place through the end of the horizon for the short term planning. The real time optimization only has to take into account events that take place within the information control horizon, specified in chapter 2. When the information prediction horizon is shorter than the information control horizon, the actions towards the end of the information control horizon matter for the current action, but there is no predictive information to base them on. These actions can therefore be assumed to follow the long-term steady state policy. The information that these actions contain that is relevant for the current action to optimize is contained within the value-to-go function directly preceding the first action of the steady state policy. Inclusion of this information in the short-term optimization can be achieved by using the value-to-go of the long-term policy to reward the final state of the short-term optimization. This is in fact making use of Bellman's principle of optimality again, ensuring that the total benefits are maximized.

The second influence, of the short-term on the long-term operations, is due to the fact that the uncertainties under which the decisions in the long term policy are supposed to be taken are not reflecting the true uncertainty under which decisions are taken. In the long-term policy, the decision is presumed to be only based on the information which is

in the states of the Markov chain, which usually includes the actual reservoir level, the actual inflow and the month of the year. In reality, the probability distribution for the state transitions in the near future are conditioned on other, external information, which can be summarized in a forecast. The uncertainty under which decisions are taken is thus decreased ("conditioning reduces entropy"). To account for this effect, Stedinger et al. (1984) proposed to use the best forecast for the inflow as a state variable instead of the previous month's inflow and reported an improvement in the performance of the resulting policy. A possibility to also take the uncertainty in this forecast into account is offered by the technique of Bayesian Stochastic Dynamic Programming (BSDP, Karamouz and Vasiliadis (1992); Kim and Palmer (1997)). In this method, the transition probabilities for the inflows are continuously updated to reflect the forecast information. The conditional distribution of forecasts, given observed inflows is used to account also for forecast uncertainty. A limitation of the method is that it can only take into account the predictive power for one timestep ahead.

7.6 How predictable is the inflow? - entropy rate and the Markov-property

A process satisfies the Markov property if all information that the past carries about the next state is captured in the current state, which may be a vector that contains the flow from previous time steps. This can be written in terms of the familiar information measures mutual information (Eq. 7.9) and conditional entropy (Eq. 7.10)

$$P(X_{t+1}|X_t) = P(X_{t+1}|X_t \dots X_1) \tag{7.8}$$

$$I(X_{t+1}; X_t, X_{t-1} \dots X_1) = I(X_{t+1}; X_t) \tag{7.9}$$

$$H(X_{t+1}|X_t, X_{t-1} \dots X_1) = H(X_{t+1}|X_t) \tag{7.10}$$

Consequently, the information that X_t contains about X_{t+h} decreases with increasing h and all information is transferred through the intermediate states. A first order discrete Markov process can be completely described by a transition matrix specifying the conditional probabilities of ending up in the next states, given each possible current state.

For a time-homogeneous Markov process, information about the current state fades away in time, until a stationary distribution μ is obtained (under the conditions of irreducibility and aperiodicity; see Cover and Thomas (2006)). The stationary distribution is often also referred to as the "climatic" distribution, and can be found by solving

$$\mu M = \mu \tag{7.11}$$

where M is the transition probability matrix for the Markov chain. If current information is available, the distribution will diverge from the uniform stationary distribution for some time period. When the current state is precisely known ($X_t = x$), the entropy $H(X_{t+h}|X_t = x)$ will increase with lead time h and the Kullback-Leibler divergence

$D_{KL}(\mu||X_{t+h}|X_t = x)$ will asymptotically go to zero, meaning that there is no extra information relative to climatology. This was defined as the information prediction horizon in chapter 2; see also Fig. 2.6 on page 28.

In the model used in the SDP optimization, a different transition probability matrix is estimated for each month. This results in a cyclostationary process, which can have a different stationary distribution for each month. Because each month of the year is individually split into 5 equiprobable classes, the information measures between the different months can be interpreted in the same way as if the flows form a stationary process.

An important characteristic of a stationary Markov process is the entropy rate H', which gives the average uncertainty of each new value. Because the distribution of the next inflow depends on the current inflow, this uncertainty is less than the marginal entropy ($\log 5$ in this case). The entropy rate for a first order Markov process is simply the conditional entropy $H(X_{t+1}|X_t)$. This conditional entropy is equal to the marginal entropy minus the mutual information

$$H'(X_1 \ldots X_T) = H(X) - I(X_t, X_{t+1}) \tag{7.12}$$

In Fig. 7.9, the mutual information between the inflows in subsequent months is plotted for different lead times and different initial months. The mutual information is calculated in two different ways. The first assumes that the inflow is a Markov process and uses the product of the transition matrices in the different timesteps to calculate the resulting transition probabilities for multiple timesteps ahead. The second method directly estimates the joint distribution based on the series of pairs x_t, x_{t+h}. If the first method yields a lower mutual information than the second, the persistence is stronger than explained just by lag-1 dependence and predictability is underestimated by the Markov-1 model. Because of the limited data, the estimation of transition probabilities is coarse and leads to spurious mutual information, which would even be found in a random data-set (a bootstrap revealed that 10% of randomly drawn data sets yield a mutual information of over 0.37 bits). Due to this curse of modeling, no conclusions can be drawn about "beyond Markov predictability" for this case.

Notwithstanding the practical difficulties presented by this curse, we can generally expect that inflow forecasts will in reality be based on more information than just the current inflow. In theory the predictability can be assessed by looking at the mutual information $I(X_{t+h}, Y_t)$, in which Y_t is the probability distribution of the vector of all potentially relevant information available at time t.

7.7 The influence of information on the marginal value of water

Generally, more predictable inflows lead to better decisions. This translates into an increase in the value that can be obtained from the water in the reservoir. A real-time control system that uses more information than the states in the Markov chain of the

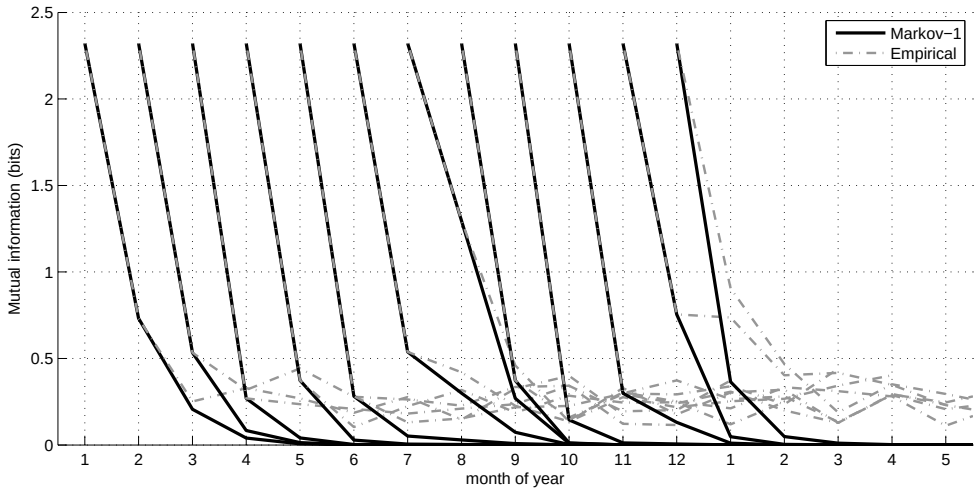


Figure 7.9: The mutual information between the inflow classes in different months calculated from the Markov matrices and the empirically estimated mutual information, based on the joint distribution of classes at various time-lags.

steady state optimization capture, will therefore increase the value of water. In the case of the hydropower objective in the example, this translates into an operation strategy that maintains a higher average elevation in the reservoir or prevents more spills. Given that such a system is used also to make future decisions, they will improve, leading to more benefits from the same amount of water. It is therefore likely that also the future value of water will increase. In this section it is investigated how these differences in value might affect the optimal decision.

The value-to-go (and therefore the decision) at time step t depends on all information that is currently available about the future behavior of the water system. This behavior does not only include the inflows, which are not controllable, but also the future decisions about releases. These releases depend on the information that will be available at the time those future decisions are made. Therefore, the optimal release in the current time step depends not only on our best probability estimates on the future inflows themselves, but also on our best estimates about how much information will be available at the time of future decisions. A model about the future growth in information is therefore necessary.

The Markov chain model of inflows in a typical SDP problem is such a model of future information, which is assumed to be captured by the state. To enable the disaggregation in time that forms the basis of dynamic programming, the external influences need to satisfy the Markov property and be part of the state, or be completely random. To satisfy this condition, all states of the rainfall runoff model and even the dependency structure in the rainfall needs to be modeled as states in the Markov process. Due to the curse of dimensionality, the number of states of the Markov process is often assumed to be low, even if in reality more information on the future inflows is available and used in real-time. Although this enables a fast solution, it also underestimates the amount of information that is available for future decisions. An underestimation of the value that can be generated from the water is the result.

A condition for this underestimation to influence the optimal operation is that also the derivative of the value function changes. A systematic fixed underestimation of future benefits for all reservoir levels would not have an effect on operation. If the real-time information has an additional value for operation that depends on the reservoir level, also the *marginal* value and thus the optimal decision is affected. Although it is beyond the scope of the present case study, it is conjectured that the marginal value is indeed affected, because accurate forecast information seems to have most additional value for operations while the reservoir is nearly full. This would result in an increased marginal value of storage and decisions that favor higher reservoir elevation levels most of the time, with occasional drawdowns for well-predicted inflow peaks.

To achieve a steady state policy that accounts for the information gain of the true real-time operations, a higher order model would be needed. The computational effort required for such an approach is often prohibitive. An alternative approach is to correct the value functions by an empirical factor, that is learned from the actual operations. This can for example be achieved using reinforcement learning based approaches (Sutton and Barto, 1998), which will be shortly discussed in the recommendations of this chapter.

7.8 Sharing additional benefits of real-time information between stakeholders

Apart from depending on the current reservoir level, the additional value of future information may also be different for different users. This can have an effect on negotiations between these users. If one of the users gains more than others from the effects of real-time information that is used in the actual decisions, a negotiated off-line Pareto-optimal solution needs to be renegotiated online to share the additional benefits. This section discusses this issue in more detail.

Optimization by SDP can be a useful tool in negotiations between different stakeholders in a water system, because it can find optimal solutions for a given objective, taking into account uncertain natural influences. Without optimization tools, these solutions can be difficult to identify. A release policy that results from a SDP optimization could therefore present a win-win solution for two conflicting objectives compared to the suboptimal status quo. Solutions like these, which identify new benefits that can be shared, can stimulate negotiations between stakeholders; see Soncini-Sessa et al. (2007) for possible methodologies. One way to stimulate negotiations is to find optimal policies for several differently weighted objectives. The set of solutions thus obtained forms a Pareto-front. The Pareto-front contains the set of solutions where no objective can be improved upon without deteriorating one of the other objectives. All solutions dominated by the (i.e. above the) Pareto front can be improved upon for one of the stakeholders without a loss for one of the others. It therefore makes sense to only negotiate about the Pareto-optimal policies.

Identifying the Pareto-front involves several optimization runs. Simplifications are often required to make the computations tractable. A typical way to do this is to reduce state

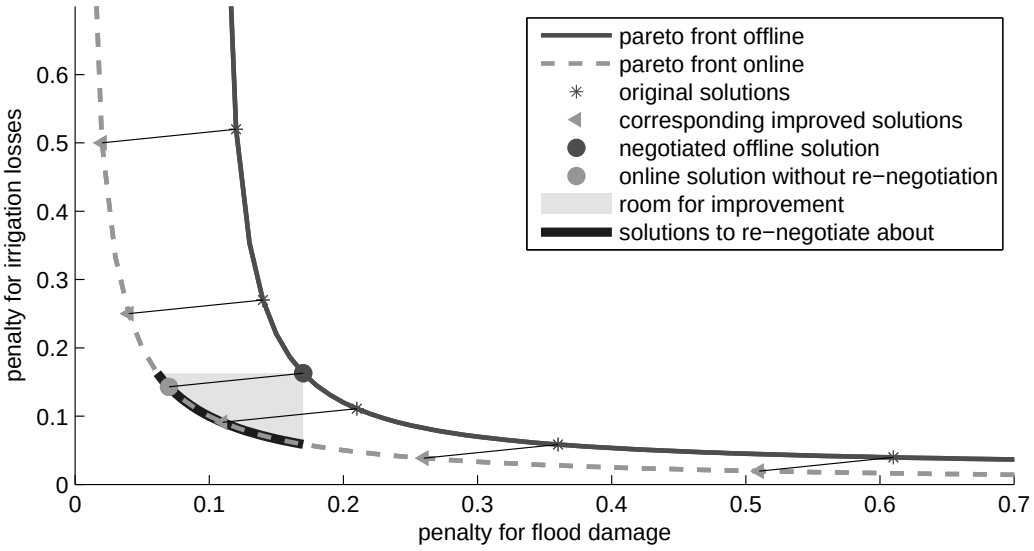


Figure 7.10: Hypothetical example of a shifted Pareto-front as a result of online operation. The lines indicate the online performance of the off-line policy with the original weights. The way to share the benefit from real-time operation can be part of new negotiations.

dimension of the SDP model, for example by replacing a Markovian inflow model with white noise, as was done in Pianosi and Soncini-Sessa (2009). In the project described in that paper, this simplified SDP model was used to present possible Pareto optimal solutions to stakeholders with conflicting objectives of flood protection in Lake Verbano and irrigation downstream. Several rounds of negotiations resulted in an objective function with fixed weights for various objectives. Subsequently, the operations were improved by a new real-time operation system using a heteroskedastic inflow model based on current information about the actual state of the catchment; see Pianosi and Soncini-Sessa (2009) .

The additional information that is used in the online optimization has a value due to better meeting the objectives. Therefore, the Pareto-front found by off-line optimization, which is used in the negotiation, is improved upon by the real-time operations; see Fig. 7.10. The improvement that can be made by using more real-time information should in principle benefit at least one of the stakeholders, without making the other worse off, theoretically leading to a new, improved, Pareto front. The online operations, based on the off-line policy with the negotiated weights, result in a solution that dominates the off-line solution and also lies on the new online Pareto front.

However, this new solution presents only one of a range of possible new solutions that provide a Pareto improvement over the negotiated off-line solution. When the weights for the off-line policy are unchanged, the benefits of the online information will mostly go to the stakeholder with the interests most sensitive to the real-time information. This is usually the objective associated with the shortest timescale. For example, in a reservoir both used for irrigation and flood prevention, the real-time information about current and expected rainfall in the catchment, which is not taken into account in the off-line

optimization, can have large benefits for flood protection. Pre-releases based on measured precipitation lead to a significant reduction in flood peaks downstream and the maximum level in the reservoir Pianosi and Soncini-Sessa (2009).

The additional benefits to irrigation, on the other hand, are not very significant, as the real-time information about how long upcoming droughts will last is usually limited. The real-time optimization mostly benefits flood protection, but the stakeholder with an irrigation interest might feel entitled to a share of these benefits. This can be realized by partly compensating the real-time gains for flood protection by selecting another off-line policy from the Pareto front that benefits irrigation at the cost of seemingly increased flood risk. The real-time information in combination with this change in off-line policy results in a balanced win-win solution compared to the negotiated off-line Pareto optimal solution.

Although just a thought-experiment, this reasoning shows that next to the information provided by models, also a model of the information flows can contribute to reach more equitable win-win solutions and to negotiate about the right options. Again, a model of the amount of information on which the actual decisions are based seems to be necessary to fully appreciate the value of water.[2] The previous case dealt with the future value, while this case concerned the value for the various stakeholders. Real world case studies are needed to determine the practical significance of this effect in various situations.

7.9 Reinforcement Learning to approximate value functions

In high-dimensional cases, it might be possible to use techniques that empirically estimate the value-to-go functions, rather than to try to explicitly compute them based on simplified models. A promising approach is reinforcement learning; see Sutton and Barto (1998) for a good introduction. Promising first applications in water system operation have been described by Bhattacharya et al. (2003), Lee and Labadie (2007) and Castelletti et al. (2010). In reinforcement learning, an agent interacts with its environment (see Fig. 7.11), by taking actions a, which are mapped from the current state s by a policy π. The environment, depending on the state and the action, reacts by giving a reward r to the agent. The objective of the agent is to maximize the total reward over a finite horizon or the total discounted reward over an infinite horizon. This maximization is pursued by learning from the interaction and finding out which state-action combinations result in the highest overall rewards. An important aspect is the trade-off between exploitation and exploration. On the one hand, the agent wants to benefit from good action-state combinations found so far, on the other hand it must explore other actions to be able to learn improved actions.

There are many parallels between dynamic programming (DP) and reinforcement learning (RL). Also in reinforcement learning, value functions are implicitly or explicitly calculated

2. Theoretically, the increased value of water due to the online information is closer to the actual value of the water and should therefore be the basis of the negotiations.

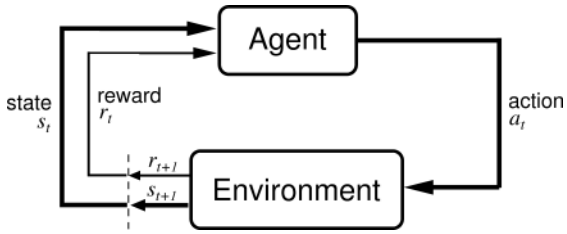


Figure 7.11: The agent-environment interaction in reinforcement learning (source: Sutton and Barto (1998))

to summarize the expected rewards associated with one state or one state-action combination. In contrast to DP, where value functions are calculated based on an explicit model of the environment, in RL these value functions are estimated from a growing experience, without the requirement for knowledge of the environment except for the state. Where in (S)DP the optimal policy is an exact mapping from the state to the optimal action, in RL a policy is a mapping from the state to a probability of an action. A high entropy probability distribution of the actions means a high exploration. If applied on the same problem, an ideal RL algorithm will converge to the same optimal policy as DP, gradually reducing towards a zero entropy action distribution (i.e. after the learning is complete, there is no exploration, only exploitation).

In practice, RL algorithms usually only reach an approximation of the optimal solution. The large advantage however, is that they can handle problems of far larger state-dimension, because instead of an exhaustive calculation of the value functions for all possible states, it focuses attention on the states that look promising from previous experience. Because experimenting on a real reservoir in usually not desirable, a simulation model can be used as the learning-environment. Furthermore, if real-time information is used to improve decisions, but this information is not captured in the state of the long term SDP optimization, that state ceases to satisfy the Markov property. Even with such an imperfect model, RL can still approximate the best policy. Where DP approaches rely on a perfect model, RL discovers an approximate model in the learning process of interacting with the environment. In this way it partly overcomes the "curse of modeling".

In the case considered in this chapter, a real-time operation module could be considered as part of the environment for the agent that optimizes the long-term policy. The actions of this agents are sent to the environment and used in the real-time optimization. The rewards, which are calculated in the environment, depend on the performance of the real-time optimization and the states are a summary of the information that the real-time system plus the long-term policy use as an input. In this setup, the corrections to the value function due to new real-time information are learned automatically by the RL algorithm, which can be initialized with the value-functions from the original SDP solution. The correction to the value function made by the reinforcement learning algorithm will thus be an approximation of the additional value added by the real-time information, which is too complex to calculate exactly.

7.10 Conclusions and recommendations

The value of water can be made more explicit by formulating an optimization problem including the different uses of water. Because of the disaggregation in time, SDP can explicitly represent the trade-off between immediate and future benefits. The time-varying value of water found by SDP also includes the effect of the uncertainty in the inflows, which makes water less valuable. Conversely, more information increases the benefits that are obtained from using the water optimally. This makes the value of water dependent on information available for future decisions.

An optimization model by SDP includes a simple model of the information that will be available for future decisions. The model is a Markov-chain, in which all information available for the decision is captured in the state. Due the exponential dependence of computational burden on the state dimension and the difficulty of estimating reliable transition probabilities for higher order models, often models of low state dimension are chosen to represent inflow predictability. Such low-dimensional states do often not capture the complete information available for actual decisions.

It is conjectured that especially for operations at shorter timescales and at high reservoir elevation levels, additional real-time information can have significant value for operation. Not taking this value into account in the long-term policy may lead to lower than optimal lake levels in a hydropower-only setting. In a multi-objective setting with irrigation and flood control, it may also lead to an over-emphasis on the objectives of flood related stakeholders, as they are profiting most from the real-time information. This may need to be accounted for in negotiations.

It is in principle impossible to fully separate the short and long term planning, unless the state truly has the Markov property. The long term influences the end-of-horizon value-to-go function for the short term, while the short term information influences the benefits in the state transitions in the long term. When for real-time operations, forecasts are used that are based on models with high dimensional state and high-dimensional observations, it is impossible to include this into a model of future information for the decisions.

Reinforcement learning may be a promising way forward to address the problem identified in this chapter. In fact, using such techniques amounts to admitting the infeasibility of an exact solution and trying to find an approximate solution to the full problem, rather than finding a full solution to the approximate problem. This might facilitate taking into account empirically the complex dynamics and value of information, which are overlooked by SDP solutions using reduced models.

Chapter 8

Conclusions and recommendations

> *"We have the duty of formulating, of summarizing, and of communicating our conclusions, in intelligible form, in recognition of the right of other free minds to utilize them in making their own decisions."*
> - Ronald Fisher

The initial focus of this research was developing methods for risk based water system operation. Information soon presented itself as central concept underlying this task. Impressed by the coherence and broad applicability of Shannon's information theory and the "Jaynesian" viewpoint on probability and information, the author shifted the focus somewhat to specifically study the role and flow of information in this process. The problem of risk based water system operation can be restated as "Rational water management decisions with incomplete information" and approached using theorems from the interlinked fields of decision theory, control theory and information theory. Between data and decisions are predictions[1], which summarize the information extracted from the data to enable informed decisions. Some of the work in this thesis therefore focused more specifically on information in predictions. Before reviewing the more concrete methodological contributions of this thesis and giving some recommendations for practice and future research, the first section presents some collateral insights about information that were obtained during the research. To the author personally they provide the most rewarding outcome of the Ph.D. research process.

8.1 Conclusions at the conceptual level

8.1.1 The nature of information

When the word information is used in daily conversation, it is often implicitly restricted to mean useful information: an informative message is required to cause surprise, but also to contain meaning. Information theory is just concerned with surprise, which is related to changes in probability, and not with the meaning (usefulness; utility) of information. Information about some event is the reduction in uncertainty about its outcome. As

1. although these are not always explicit; see e.g. reinforcement learning, where actions follow directly from observations

explained in chapter 3, it is important to distinguish between perceived uncertainty, which can be evaluated ex ante, and true uncertainty, which can be evaluated ex post when the true outcome is known. The perceived uncertainty can be expressed by the entropy of the probability distribution that is adopted in the light of all information available. The true uncertainty is related only to the probability attached to the outcome that actually occurred, and can be expressed as the Kullback-Leibler divergence from the truth to the probability distribution that was adopted before knowing it.

Information is contained in data (observations; evidence), which leads to conditioning of probability distributions of related phenomena and therefore to a reduction of uncertainty about those phenomena. In order to extract this information from data, we need models (algorithms; programs), which give us predictions of the uncertain events. Once a number of predictions has been made and the corresponding outcomes observed, the information that the predictions contained about the observations can be used to evaluate the quality of the model. If enough predictions and observations are available, we can distinguish between wrong and correct information. The decomposition in chapter 5 revealed that the remaining uncertainty is equal to the original uncertainty plus the wrong information minus the correct information. Once data can be processed in a general, repeatable manner to give informative predictions that can be successfully tested with other data, we have found a pattern in the data and this can be referred to as knowledge. The knowledge is embodied by the mathematical relations of the model that provides the informative predictions.

8.1.2 The flow of information

Information that is used for human decisions ultimately stems from observations. These can be obtained directly through our senses or with the aid of measuring instruments. To enable optimal decisions, we need to extract from those observations the information that relates to the part of the state of the world that influences the impact of our decision on our utility. In this process, information flows from observation to decision. Good decisions are fed by an inflow of information from the right observations. An increased inflow of field observations can therefore improve decisions. However, we should also manage the flow correctly. On the one hand care should be taken not to lose any of the relevant information. On the other hand, we also should be careful not to add any information that is not justified. The principle of maximum entropy, or minimum relative entropy, such as applied in the ensemble weighting in chapter 4, can help ensure this. Chapter 5 presented a method for testing how much correct information from the observations ends up in the predictions (the resolution term) and how much wrong information is introduced by miscalibration (the reliability term). The remaining uncertainty about the observations, as measured by the divergence score, is also the performance measure to minimize in model calibration. The measure has the desirable property that its expectation is minimized only if the model gives a correct representation of uncertainty. Both information left behind and unjustifiably added information deteriorate the expected score, which therefore provides an incentive to correctly represent both our knowledge and our ignorance. This representation, a probabilistic forecast, in turn enables optimal decisions.

8.1.3 The value of information

Information obtains value through decisions. In hindsight, informed rational decisions are on average better than uninformed decisions. The value of new information can be expressed as the expected utility of a decision with the new information minus the expected utility of the decision with the a priori information. In the case of water resources management, information also influences the value of water through the allocation decision. Water adopts the value of the use it is allocated to. Informed decisions are based on accurate estimates of the value of each use, allocating water to its most valuable use. In sequential decision processes, such as repeated decisions about the release from a reservoir, the allocation also represents the trade-off between immediate and future use. Consequently, the value of water then also depends on the quality of future decisions, which are in turn dependent on information that will be available in the future. These complex dynamics of information are to some extent explicitly handled by stochastic dynamic programming. Often, however, the future information is too complex to model explicitly and we should resort to empirical implicit approaches (reinforcement learning, chapter 7) to account for the value of future information.

8.1.4 The necessity of probabilistic predictions

Predictions or forecasts are of paramount importance in science and engineering. In science they are the interface between theory and observation. In engineering, they can be seen as messages that communicate information to a user. However, they should also communicate the missing information or uncertainty to the user. In order to minimize the remaining uncertainty about the true outcome for the user, they have to accurately reflect this uncertainty. Deterministic forecasts violate this requirement to the largest possible extent. Strictly speaking, they increase the remaining uncertainty to infinity, unless they are perfect. As is revealed by the decomposition in chapter 5, this is caused by the large amount of wrong information that they communicate by pretending to be certain, next to the correct information they may contain.

An experienced or knowledgeable user of the forecast may be able to filter out some of the wrong information by implicitly or explicitly recalibrating the forecasts, i.e. not believing them completely. However, this moves a considerable responsibility to the user that in fact belongs to the forecaster. Ironically, the users who are claimed not to be able to handle probabilistic forecasts and are for that reason provided with deterministic forecasts are the ones who have to rely most on their ability to subconsciously make probability estimates based on the limited information in a deterministic forecast.

Another reason to use probabilistic forecasts is that they allow rational decisions for different users at the same time. While a deterministic forecast can theoretically be optimized to allow optimal decisions for one user with a specific decision problem, probabilistic forecasts contain a representation of the full information and uncertainty about the forecast quantity that is relevant for all decision makers, irrespective of their decision problem. Given that most forecasts are made by experts and communicated to a varied body of users who are all free to decide on their own problem, forecasts ought to be probabilistic.

8.1.5 Information theory as philosophy of science

Also in the context of pure science there is a compelling reason why forecasts should be informative, i.e. probabilistic. It is almost part of the definition of science that it is required to produce testable predictions. In fact, this could be interpreted as a requirement for predictions to mandate their own way to be tested. The joint framework of probability and information theory provides an opportunity to meet this requirement. All forecasts that are not probabilistic need external assumptions to determine the way they are tested. As argued in chapter 5, these either relate to utility or they implicitly specify a probability distribution, which should have been stated a priori.

Science has the task of extracting information from observations about other observable quantities. This is done by trying to find patterns in observed data and representing them in mathematical formulae. These formulae represent our scientific knowledge and are the algorithmic representation of the redundancy in the observations. All patterns that are present in the observations allow them to be represented in a theory, which can be more compactly represented than the observations themselves. As is outlined in chapter 6, the strong analogy with data-compression is evident: a theory can be seen as a compression algorithm for data. This compressibility is the reason for the very possibility of science.

The divergence score presented in chapter 5 can be interpreted as the minimum filesize attainable by the best compression algorithm to represent the observations, while knowing the forecasts, i.e. the predictors and the model. To represent all observations, both the predictors and the model need to be stored as well. The filesize for the predictors is inversely related to the generality of the model. If the filesize is large, the conditions for which to model yields its predictions need to be extensively specified, while a smaller file means a more general model. The filesize of the decompression algorithm, which represents the model, is a measure for the complexity of the theory. The combined filesize is related to all the things that are left unexplained. An explanation is thus simply a description that is shorter than the explanandum. An inverse relation between description length and prior probability of a model, as suggested by algorithmic information theory, fits into a Bayesian framework and provides a possible justification for the principle of parsimony.

A perfect theory of everything would not need an input file for predictors nor a file to store observed outcomes, given the predictions, because the predictions already are equal to the observations and the theory is so general that all input observations have become knowledge and are thus part of the algorithm. In algorithmic information theory, the filesize of that algorithm is the Kolmogorov complexity of the universe. This complexity might be infinite, if we consider information entering from parallel universes through quantum interference (see Deutsch (1998)). We are now maximally out of the scope of this thesis, so the next section returns to a more practical level.

8.2 Methodological contributions and recommendations for practice

8.2.1 On risk based water system operation

Operation of water systems should be risk based when the control problem is not certainty equivalent. In those cases, the expected degree fulfillment of the objectives, given all possible futures, should be maximized. Two conceptual time horizons relating to sensitivity for the future and to information about the future were defined that can serve as guidelines in the design of predictive controllers. Furthermore, two different multiple model formulations of MPC, presented in chapter 2, serve as limiting cases for the availability of new information for future decisions (no information versus perfect information). In theory, stochastic dynamic programming, as applied in chapter 7, offers the possibility to exactly model the availability of future information, but only at the expense of computational intractability for most real-world problems. Two routes of circumvention are available to achieve near-optimal risk based water system operation. The first route makes modifications to the problem to solve by making assumptions that allow a fast and exact solution to the modified problem. The second route finds near-optimal actions empirically by using techniques from artificial intelligence, such as reinforcement learning. In chapter 7, that technique was identified as a promising way forward for reservoir operation.

8.2.2 On weighted ensemble forecasts

The information that is contained in a forecast resides in what is actually presented. If a forecast does not specify the entire probability distribution, but just a number of summary statistics, the maximum entropy distribution consistent with those statistics should be assumed. The forecast information contained in the statistics can be added to an ensemble by adjusting the weights of an example to match the statistics, while minimizing relative entropy from the original weights. This method prevents the weighted ensemble to represent too much or too little of the forecast information. The minimum relative entropy update (MRE-update) presented in chapter 4 forms a readily applicable method for generating weighted ensembles using this principle. The information-theoretical foundation makes it theoretically superior to the existing methods for ensemble weighting. When forced to exactly match the forecast, the existing pdf-ratio method can be used as a fast solution method for the MRE-update, making use of the form of the analytical solution to the MRE optimization problem.

8.2.3 On the evaluation of probabilistic forecasts

The Brier score and other derived quadratic scores like RPS and CPRS fail to meet certain fundamental requirements for measures of forecast quality. Forecasting should be seen as a communication problem and the quality of forecasts should be evaluated using the divergence score introduced in chapter 5. The Brier score, which has been used extensively in meteorology for the last 60 years, is a second order approximation of the

divergence score and should be replaced by it, leading to notably different evaluation in the extreme forecast probabilities.

The analogy between the divergence score and its second order approximation can be extended to a well-known decomposition of the Brier score into uncertainty, reliability and resolution, allowing a re-interpretation of these components as missing, wrong and correct information. The decomposition can also be generalized to the case of uncertain observations, including the observational uncertainty as a closing term. These measures proposed in chapter 5, given their axiomatic justification, are expected to become part of standard forecast verification practice.

8.2.4 On performance measures for model inference

The measure that is chosen to evaluate a model implicitly specifies the probabilistic part of that model. This measure should therefore be specified a priori, along with the model, otherwise part of the model is formulated with knowledge of the observations to test it against. A completely specified model gives probabilistic predictions that can and should be tested using information-theoretical measures. In contrast to measures that reflect a decision-problem-specific utility, information-theoretical measures allow the model to optimally learn from all information in the observations. In principle, the purpose of a model should not influence its calibration.

The likelihood principle states that all information that data contains about a certain model is contained in the likelihood of the data given that model. When the observations are certain, the divergence score has a direct relation to the log-likelihood. The divergence score also represents the average number of bits per value that is needed for storing the data, using an optimal compression algorithm like arithmetic coding. At a fundamental level, quality of a model can be seen as how much it compresses the data. In algorithmic information theory, the principle of parsimony can be formalized as the description length of a model or the size of the decompression algorithm. Algorithmic information theory measures are in principle incomputable, but practically computable approximations exist in the form of commonly used model complexity control methods.

Chapter 6 contains some pioneering practical explorations of compressibility of hydrological time series using general purpose data compression algorithms for computer files. Further developments in this direction may be useful to estimate the amount of information extractable from hydrological time series; compare the performance of hydrological models with general purpose pattern recognition; and provide a basis for model complexity control.

8.2.5 On optimization of reservoir release policies

Estimation of the future value of water and optimal decisions about water allocation are two sides of the same coin. Real time operation of a hydropower reservoir can be viewed as an allocation problem in time. Interesting economic information can be derived from the

results of an optimization of the release policy using stochastic dynamic programming. The information-theoretical analysis and thought experiments in chapter 7 suggest that for optimal reservoir operation, a model of future information growth is necessary. Reinforcement learning is suggested as a practical approach to this problem. Furthermore, it is suggested that the discrepancy between the information in the state of common SDP formulations and the information actually available for operations may need to be accounted for in negotiations based on multi-objective SDP results.

8.3 Limitations and recommendations for further research

8.3.1 Going from discrete to continuous

The divergence score, presented in chapter 5, could be extended to forecasts of continuous variables by applying the Kullback-Leibler divergence to probability densities rather than probability masses, but the interpretation becomes more difficult and needs some further thought. The decomposition of the divergence score is limited to discrete predictands. One reason for this is that the uncertainty component, entropy, lacks a convenient extension to the continuous case, while relative entropy *does* have a well-behaved continuous counterpart. This can be understood by thinking about the remaining uncertainty about a real number. Unless a real number is precisely (e.g. theoretically) known, this remaining uncertainty can be thought of as infinite, because an infinite amount of information (e.g. digits) is needed to specify the number exactly.

In practice, the way we deal with real numbers is usually also to discretize them. This limited precision, given by measurement equipment and computer implementation of models, is usually so high that to do meaningful information-theoretical analysis, we need to discretize further or have vast amounts of data at our disposition. Strangely enough, this means that we have to throw away information in order to get useful estimates of information and information flows. In many cases, e.g. ensembles and binary forecasts, the discretization has already been made, but for practical application in a inherently continuous setting, this issue has to be further investigated.

8.3.2 Expressing prior information

In many cases, hydrology and water management require predictions about complex systems. This leads to high data requirements. Because the available data is usually limited, meaningful prediction heavily relies on prior information in the form of assumed model structures and prior parameter distributions. This information does not come out of thin air, but usually neither is justified rigorously.

As Jaynes (2003) remarked, the process translating various forms of information into probability distributions should represent fully half of probability theory, but so far has received little attention. The principle of maximum entropy is one of the few principles that can be used for this, but its scope is very limited. This restricts the applicability of,

for example, the minimum relative entropy update in chapter 4 to cases where available information can be expressed as constraints. For physical process knowledge, the current practice is usually to convert some rather vague notion of processes into a deterministic model structure, leading to false certainty. Until more principles are found for expressing various forms of incomplete information consistently, no good alternatives exist for this practice except using multiple deterministic model structures simultaneously.

8.3.3 Merging information theory with statistical thermodynamics

Water system operation needs predictions from hydrology and hydrology tries to predict the behavior of complex systems. These systems are driven by fluxes of energy and water which in their turn are driven by the chaotic dynamics of the atmosphere. In all the processes that take place, low entropy energy that eventually comes from solar radiation is continuously dissipated. Taking an integrated view on entropy production, entropy import and export, hypotheses about maximum entropy production have been formulated (see Kleidon (2004); Kleidon and Schymanski (2008); Kleidon (2010)). Along similar lines, principles of maximum energy dissipation in hydrological systems have been postulated by Zehe et al. (2010) to explain preferential flow. Also certain biological optimality principles, see e.g. Schymanski et al. (2009), seem to explain certain phenomena quite well. In the generalized maximum entropy framework of Jaynes (1957), all these forms of optimality should be translatable to high probability given limited information on macroscopical scale. When the maximum entropy distribution given some macroscopical constraint does not match the observed data, there is an indication that there are other constraints to be discovered which have a relation to the sufficient statistics for the apparent distribution of the data. Furthermore, information-theoretical investigations of non-equilibrium thermodynamics, such as the pioneering work of Dewar (2003), is likely to yield important insights in the behavior of complex hydrological systems.

8.3.4 An integrated information-theoretical framework from observation to decision

This thesis, which concerned topics in forecasting, model inference and optimal decisions under uncertainty, can be summarized as pursuing optimal information extraction from the given input data to support decisions. An interesting new dimension to this problem arises when the input data for the forecasting models is not fixed. This is the case when there is an opportunity to install new measuring equipment to extract information from the environment to improve predictions and ultimately decisions. The hydrological literature describes a number of information-theoretical approaches to optimal monitoring network design; see e.g. Alfonso et al. (2010). By considering these methods in conjunction with the approaches presented in this thesis, which concern forecast evaluation, model inference, and Bayesian decision theory, the entire flow of information from observation to decision can be considered and optimized for risk reduction. Possible applications could be assessing the value of proposed monitoring networks for decisions or designing monitoring networks for a specific decision problem.

8.3.5 Problem solved, but the solution is a problem

In this thesis we have seen that a large part of science and engineering can be framed in terms of information, probability and decisions. Some of the deepest theories about algorithmic information seem to suggest that inductive inference could in principle be completely formalized and automated, but an optimal method of inference is necessarily incomputable. Different computable approximations of optimal induction exist that represent a minimum of prior information and can thus be used to study artificial intelligence. Note that one of the latest developments in artificial intelligence is AIXI (universal artificial intelligence; Hutter (2004); Legg (2008)), which merges Solomonoff induction (advocated in chapter 6) with Bayesian decision theory and reinforcement learning (advocated in chapter 7). In this way it defines a learning agent that makes optimal risk based decisions, given the information that is received from previous interactions with the a priori unknown environment and unlimited computational resources. Time and memory bounded versions, which are conditionally optimal, also exist. In principle it can thus be regarded as the ultimate golden standard solution to risk based water system operation. The only way in which this method can be improved is making use of prior knowledge about the environment and the control problem.

In most actual problems in science and engineering there is considerable prior information available. The main challenge in the field can thus be formulated as merging human experience, sensory capabilities and pattern recognition capabilities that have evolved over $4 * 10^9$ years of interacting with the environment and evolutionary computation (we can view natural selection as adding information from the environment) with theoretically optimal decisions for a naive agent that has limited experience in interaction with the environment and relatively little computational power, but does not get bored at analyzing vast amounts of data.

Becoming good water managers was essential for our intelligence to maintain its own hardware (i.e. survive). Maybe one day we will be replaced by silicon-managing robots, but until that time we must operate our water systems rationally, in a risk-based fashion. Once again we have ventured way outside the scope of this thesis, so it is time to stop adding information to it. Although raising more questions than answering them, hopefully this thesis sparked some interest in pursuing the application of information theory and artificial intelligence in water resources research...

References

Abebe, A. J. and Solomatine, D. P. (1998). Application of global optimization to the design of pipe networks. In *Proc. Int. Conf. Hydroinformatics*, volume 98, pages 989–995.

Ahrens, B. and Walser, A. (2008). Information-based skill scores for probabilistic forecasts. *Monthly Weather Review*, 136(1):352–363.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Alfonso, L., Lobbrecht, A., and Price, R. (2010). Information theory–based approach for location of monitoring water level gauges in polders. *Water Resources Research*, 46(3):W03528.

Alfonso Segura, J. L. (2010). *Optimisation Of Monitoring Networks For Water Systems*. CRC/ Balkema press. PhD Thesis, UNESCO-IHE.

Amorocho, J. and Espildora, B. (1973). Entropy in the assessment of uncertainty in hydrologic systems and models. *Water Resources Research*, 9(6):1511–1522.

Applebaum, D. (1996). *Probability and information: An integrated approach*. Cambridge University Press.

Ariely, D. (2008). *Predictably irrational: The hidden forces that shape our decisions*. Harper, New York.

Avellaneda, M., Bu, R., Friedman, C., Grandchamp, N., Kruk, L., and Newman, J. (2001). Weighted monte carlo: A new technique for calibrating asset-pricing models. intern. *J. of Theor. and Appl. Finance*, 4(1):91–119.

Bárdossy, A. (2007). Calibration of hydrological model parameters for ungauged catchments. *Hydrology and Earth System Sciences*, 11(2):703–710.

Bárdossy, A. and Das, T. (2008). Influence of rainfall observation network on model calibration and application. *Hydrology and Earth System Sciences*, 12(1):77–89.

Barjas Blanco, T., Willems, P., Chiang, P., Haverbeke, N., Berlamont, J., and De Moor, B. (2010). Flood regulation using nonlinear model predictive control. *Control Engineering Practice*, 18(10):1147–1157.

Barnston, A. G., van den Dool, H. M., Rodenhuis, D. R., Ropelewski, C. R., Kousky, V. E., O'Lenic, E. A., Livezey, R. E., Zebiak, S. E., Cane, M. A., Barnett, T. P., et al. (1994). Long-lead seasonal forecasts–where do we stand? *Bulletin of the American Meteorological Society*, 75(11):2097–2114.

Bellman, R. (1952). The theory of dynamic programming. *Proceedings of the National Academy of Sciences of the United States of America*, 38(8):716–719.

Benedetti, R. (2010). Scoring Rules for Forecast Verification. *Monthly Weather Review*, 138(1):203–211.

Berger, J. O. and Wolpert, R. L. (1988). *The likelihood principle*. Institute of Mathematical Statistics, Hayward, CA, 2nd edition.

Bernardo, J. M. (1979). Expected information as expected utility. *The Annals of Statistics*, 7(3):686–690.

Bertsekas, D. (2005). Dynamic programming and suboptimal control: A survey from ADP to MPC. *European Journal of Control*, 11(4-5):310–334.

Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming*. Athena Scientific, Boston.

Bhattacharya, B., Lobbrecht, A. H., and Solomatine, D. P. (2003). Neural networks and reinforcement learning in control of water systems. *Journal of Water Resources Planning and Management*, 129:458–465.

Boolos, G., Burgess, J., and Jeffrey, R. (2007). *Computability and logic*. Cambridge University Press.

Botev, Z. I. (2006). A novel nonparametric density estimator. Postgraduate Seminar Series, The University of Queensland, Australia.

Brest, J., Greiner, S., Bošković, B., Mernik, M., and Žumer, V. (2006). Self-adapting control parameters in differential evolution: A comparative study on numerical benchmark problems. *IEEE Trans. Evol. Comput.*, 10(6):646–657.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Briggs, W., Pocernich, M., and Ruppert, D. (2005). Incorporating misclassification error in skill assessment. *Monthly Weather Review*, 133(11):3382–3392.

Bröcker, J. (2009). Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135(643):1512–1519.

Bröcker, J. and Smith, L. A. (2007). Scoring probabilistic forecasts: The importance of being proper. *Weather and Forecasting*, 22(2):382–388.

Burnham, K. and Anderson, D. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. Springer Verlag.

Burrows, M. and Wheeler, D. J. (1994). A block-sorting lossless data compression algorithm. Technical report, Systems Research Center, Palo Alto, CA.

Camacho, E. F. and Bordons, C. (1999). *Model predictive control*. Advanced textbooks in control and signal processing. Springer, London.

Carnap, R. (1950). Logical foundations of probability.

Castelletti, A., Galelli, S., Restelli, M., and Soncini-Sessa, R. (2010). Tree-based reinforcement learning for optimal water reservoir operation. *Water Resources Research*, 46(9):W09507.

Chaitin, G. J. (1966). On the length of programs for computing finite binary sequences. *Journal of the ACM (JACM)*, 13(4):547–569.

Chiu, C.-L. (1988). Entropy and 2-d velocity distribution in open channels. *Journal of Hydraulic Engineering*, 114(7):738–756.

Chomsky, N. (1956). Three models for the description of language. *Information Theory, IRE Transactions on*, 2(3):113 –124.

Cloke, H. L. and Pappenberger, F. (2009). Ensemble flood forecasting: A review. *Journal of Hydrology*, 375(3-4):613–626.

Cover, T. M. and Thomas, J. A. (2006). *Elements of information theory*. Wiley-Interscience, New York.

Croley, T. E. (1996). Using noaa's new climate outlooks in operational hydrology. *Journal of Hydrologic Engineering*, 1(3):93–102.

Croley, T. E. (1997). Mixing probabilistic meteorology outlooks in operational hydrology. *Journal of Hydrologic Engineering*, 2:161–168.

Croley, T. E. (2001). Climate-biased storm-frequency estimation. *Journal of Hydrologic Engineering*, 6(4):275–283.

Croley, T. E. (2003). Weighted-climate parametric hydrologic forecasting. *Journal of Hydrologic Engineering*, 8:171–180.

Dandy, G. C., Simpson, A. R., and Murphy, L. J. (1996). An improved genetic algorithm for pipe network optimization. *Water Resources Research*, 32(2):449–458.

Day, G. N. (1985). Extended streamflow forecasting using NWSRFS. *Journal of Water Resources Planning and Management*, 111(2):157–170.

Deutsch, D. (1985). Quantum theory, the Church-Turing principle and the universal quantum computer. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 400(1818):97–117.

Deutsch, D. (1998). *The fabric of reality*. Penguin Books London.

Dewar, R. (2003). Information theory explanation of the fluctuation theorem, maximum entropy production and self-organized criticality in non-equilibrium stationary states. *Journal of Physics A Mathematical and General*, 36(3):631–641.

Diaconis, P., Holmes, S., and Montgomery, R. (2007). Dynamical bias in the coin toss. *SIAM review*, 49(2):211.

Dooge, J. (1997). Searching for simplicity in hydrology. *Surveys in Geophysics*, 18(5):511–534.

Dorigo, M. and Stützle, T. (2004). *Ant Colony Optimization*. MIT Press.

Duan, Q., Ajami, N. K., Gao, X., and Sorooshian, S. (2007). Multi-model ensemble hydrologic prediction using bayesian model averaging. *Advances in Water Resources*, 30(5):1371–1386.

Duan, Q., Gupta, V. K., and Sorooshian, S. (1992). Effective and efficient global optimization for conceptual rainfall–runoff models. *Water Resources Research*, 28:1015–1031.

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, 8(6):985–987.

Faber, B. A. and Stedinger, J. R. (2001). Reservoir optimization using sampling sdp with ensemble streamflow prediction (esp) forecasts. *Journal of Hydrology*, 249(1-4):113–133.

Fenicia, F., Savenije, H., and Hoffmann, L. (2010). An approach for matching accuracy and predictive capability in hydrological model development. *IAHS-AISH publication*, pages 91–99.

Fenicia, F., Savenije, H. H. G., Matgen, P., and Pfister, L. (2008). Understanding catchment behavior through stepwise model concept improvement. *Water Resources Research*, 44:W01402.

Feynman, R. (1965). *The character of physical law*. MIT Pr.

Fiorentino, M., Claps, P., and Singh, V. P. (1993). An entropy-based morphological analysis of river basin networks. *Water Resources Research*, 29(4):1215–1224.

Foufoula-Georgiou, E. and Kitanidis, P. K. (1988). Gradient dynamic programming for stochastic optimal control of multidimensional water resources systems. *Water Resources Research*, 24(8):1345–1359.

Georgakakos, K. P. and Krzysztofowicz, R. (2001). Special issue: Probabilistic and ensemble forecasting. *Journal of Hydrology*, 249(1-4).

Global Water Partnership (2000). *Integrated Water Resources Management*. Global Water Partnership, Technical Advisory Comittee.

Gneiting, T., Raftery, A., Westveld, A., and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133(5):1098–1118.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik*, 38(1):173–198.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 14(1):107–114.

Grandy Jr, W. T. (2008). *Entropy and the Time Evolution of Macroscopic Systems*. Oxford University Press, New York.

Grünwald, P. D. (2007). *The minimum description length principle*. The MIT Press.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1-2):80–91.

Gupta, H. V., Sorooshian, S., and Yapo, P. O. (1998). Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. *Water Resources Research*, 34(4):751–763.

Gupta, H. V., Wagener, T., and Liu, Y. (2008). Reconciling theory with observations: elements of a diagnostic approach to model evaluation. *Hydrological Processes*, 22(18):3802–3813.

Hall, W. A. and Buras, N. (1961). The Dynamic Programming Approach to Water-Resources Development. *Journal of Geophysical Research*, 66(2):517–520.

Hamlet, A. F., Huppert, D., and Lettenmaier, D. P. (2002). Economic value of long-lead streamflow forecasts for columbia river hydropower. *Journal of Water Resources Planning and Management*, 128:91.

Hamlet, A. F. and Lettenmaier, D. P. (1999). Columbia River streamflow forecasting based on ENSO and PDO climate signals. *Journal of Water Resources Planning and Management*, 125(6):333–341.

Harmancioglu, N. B., Alpaslan, N., and Singh, V. P. (1992a). Application of the entropy concept in design of water quality monitoring networks. In Singh, V. and Fiorentino, M., editors, *Entropy and Energy Dissipation in Water Resources*, pages 283–302. Kluwer Academic Publishers, Dordrecht.

Harmancioglu, N. B., Singh, V. P., and Alpaslan, N. (1992b). Versatile uses of the entropy concept in water resources. In Singh, V. and Fiorentino, M., editors, *Entropy and Energy Dissipation in Water Resources*, pages 91–117. Kluwer Academic Publishers, Dordrecht.

Huang, W. and Hsieh, C. (2010). Real-time reservoir flood operation during typhoon attacks. *Water Resources Research*, 46(7):W07528.

Huffman, D. A. (1952). A Method for the Construction of Minimum-Redundancy Codes. *Proceedings of the IRE*, 40(9):1098–1101.

Hundecha, Y., Ouarda, T. B. M. J., and Bárdossy, A. (2008). Regional estimation of parameters of a rainfall-runoff model at ungauged watersheds using the "spatial" structures of the parameters within a canonical physiographic-climatic space. *Water Resources Research*, 44(1):W01427.

Hurst, H. (1951). Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, 116:770–808.

Hutter, M. (2004). *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin. 300 pages, http://www.idsia.ch/~marcus/ai/uaibook.htm.

Hutter, M. (2010). A complete theory of everything (will be subjective). *Algorithms*, 3(4):329–350.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, 106(4):620–630.

Jaynes, E. T. (2003). *Probability theory: the logic of science*. Cambridge University Press, Cambridge, UK.

Johnson, S. A., Stedinger, J. R., Shoemaker, C., Li, Y., and Tejada-Guibert, J. A. (1993). Numerical solution of continuous-state dynamic programs using linear and spline interpolation. *Operations Research*, 41(3):484–500.

Jolliffe, I. T. and Stephenson, D. B. (2003). *Forecast verification: a practitioner's guide in atmospheric science*. Wiley, Chichester, UK.

Jolliffe, I. T. and Stephenson, D. B. (2008). Proper scores for probability forecasts can never be equitable. *Monthly Weather Review*, 136(4):1505–1510.

Jose, V. R. R., Nau, R. F., and Winkler, R. L. (2008). Scoring rules, generalized entropy, and utility maximization. *Operations Research*, 56(5):1146.

Karamouz, M. and Vasiliadis, H. V. (1992). Bayesian stochastic optimization of reservoir operation using uncertain forecasts. *Water Resources Research*, 28(5):1221–1232.

Karamouz, M., Zahraie, B., and Araghinejad, S. (2005). Decision support system for monthly operation of hydropower reservoirs: A case study. *Journal of Computing in Civil Engineering*, 19(2):194–207. J. Comput. Civ. Eng.

Kavetski, D., Kuczera, G., and Franks, S. W. (2006). Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. *Water Resources Research*, 42(3):W03407.

Kelly, J. (1956). A new interpretation of information rate. *Information Theory, IEEE Transactions on*, 2(3):185–189.

Kelman, J., Stedinger, J. R., Cooper, L. A., Hsu, E., and Yuan, S. Q. (1990). Sampling stochastic dynamic programming applied to reservoir operation. *Water Resources Research*, 26(3):447–454.

Kim, Y.-O. and Palmer, R. N. (1997). Value of seasonal flow forecasts in bayesian stochastic programming. *Journal of Water Resources Planning and Management*, 123(6):327–335.

King, R. D., Rowland, J., Oliver, S. G., Young, M., Aubrey, W., Byrne, E., Liakata,

M., Markham, M., Pir, P., Soldatova, L. N., et al. (2009). The automation of science. *Science*, 324(5923):85.

Kleeman, R. (2002). Measuring dynamical prediction utility using relative entropy. *Journal of the Atmospheric Sciences*, 59(13):2057–2072.

Kleidon, A. (2004). Beyond gaia: Thermodynamics of life and earth system functioning. *Climatic Change*, 66(3):271–319.

Kleidon, A. (2010). Non-equilibrium thermodynamics, maximum entropy production and Earth-system evolution. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1910):181.

Kleidon, A. and Schymanski, S. (2008). Thermodynamics and optimality of the water budget on land: A review. *Geophysical Research Letters*, 35(20):L20404.

Klemeš, V. (1977). Discrete representation of storage for stochastic reservoir optimization. *Water Resources Research*, 13(1):149–158.

Kolmogorov, A. N. (1968). Three approaches to the quantitative definition of information. *International Journal of Computer Mathematics*, 2(1):157–168.

Koutsoyiannis, D. (2005a). Uncertainty, entropy, scaling and hydrological statistics. 1. Marginal distributional properties of hydrological processes and state scaling. *Hydrological Sciences Journal*, 50(3):381–404.

Koutsoyiannis, D. (2005b). Uncertainty, entropy, scaling and hydrological stochastics. 2. Time dependence of hydrological processes and time scaling. *Hydrological Sciences Journal*, 50(3):1–426.

Koutsoyiannis, D. (2009). Seeking parsimony in hydrology and water resources technology. *Geophysical Research Abstracts*, 11:EGU2009–11469.

Koutsoyiannis, D. (2010). HESS Opinions "A random walk on water". *Hydrology and Earth System Sciences*, 14:585–601.

Koutsoyiannis, D. (2011). Hurst-Kolmogorov dynamics as a result of extremal entropy production. *Physica A: Statistical Mechanics and its Applications*, 390(8):1424–1432.

Koutsoyiannis, D. and Economou, A. (2003). Evaluation of the parameterization-simulation-optimization approach for the control of reservoir systems. *Water Resources Research*, 39(6):1170.

Kraft, L. G. (1949). A device for quantizing, grouping, and coding amplitude-modulated pulses. Master's thesis, Massachusetts Institute of Technology. Dept. of Electrical Engineering.

Krstanovic, P. F. and Singh, V. P. (1992a). Evaluation of rainfall networks using entropy: I. Theoretical development. *Water Resources Management*, 6(4):279–293.

Krstanovic, P. F. and Singh, V. P. (1992b). Evaluation of rainfall networks using entropy: II. Application. *Water Resources Management*, 6(4):295–314.

Krzysztofowicz, R. (1999). Bayesian theory of probabilistic forecasting via deterministic hydrologic model. *Water Resources Research*, 35(9):2739–2750.

Krzysztofowicz, R. (2001). The case for probabilistic forecasting in hydrology. *Journal of Hydrology*, 249(1):2–9.

Kullback, S. (1997). *Information theory and statistics*. Dover Pubns.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Kwakernaak, H. and Sivan, R. (1972). *Linear optimal control systems*. Wiley-Interscience New York.

Labadie, J. W. (2004). Optimal operation of multireservoir systems: State-of-the-art review. *Journal of Water Resources Planning and Management*, 130(2):93–111.

Laio, F. and Tamea, S. (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences*, 11(4):1267–1277.

Landauer, R. (1961). Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, 5:183–191.

Lee, J. H. and Labadie, J. W. (2007). Stochastic optimization of multireservoir systems via reinforcement learning. *Water Resources Research*, 43:W11408.

Legg, S. (2008). *Machine Super Intelligence*. PhD thesis, Faculty of Informatics of the University of Lugano.

Lehning, M., Dawes, N., Bavay, M., Parlange, M., Nath, S., and Zhao, F. (2009). Instrumenting the earth: Next-generation sensor networks and environmental science. In *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft research.

Leung, L. Y. and North, G. R. (1990). Information theory and climate prediction. *Journal of Climate*, 3(1):5–14.

Li, M. and Vitanyi, P. M. B. (2008). *An introduction to Kolmogorov complexity and its applications*. Springer-Verlag New York Inc.

Lindley, D. (2008). *Uncertainty: Einstein, Heisenberg, Bohr, and the struggle for the soul of science*. Anchor Books, Garden City, N.Y.

Ljung, L. (1987). *System identification: theory for the user*. Prentice-Hall, Englewood Cliffs, NJ.

Lobbrecht, A. H., Sinke, M. D., and Bouma, S. B. (1999). Dynamic control of the delfland polders and storage basin, the netherlands. *Water Science and Technology*, 39(4):269–279.

Lorenz, E. (1963). Deterministic Nonperiodic Flow. *Journal of the Atmospheric Sciences*, 20(2):130–141.

Loucks, D. P. and van Beek, E. (2005). *Water resources systems planning and management an Introduction to methods, models and applications*. Unesco, Paris.

Martin, G. N. N. (1979). Range encoding: an algorithm for removing redundancy from a digitised message. In *Video & Data Recording conference*.

Mason, S. J. (2008). Understanding forecast verification statistics. *Meteorological Applications*, 15(1):31–40.

McMillan, B. (1956). Two inequalities implied by unique decipherability. *IEEE Transactions on Information Theory*, 2(4):115–116.

Merabtene, T., Kawamura, A., Jinno, K., and Olsson, J. (2002). Risk assessment for optimal drought management of an integrated water resources system using a genetic algorithm. *Hydrological Processes*, 16(11):2189–2208.

Murphy, A. H. (1970). The ranked probability score and the probability score: a comparison. *Monthly Weather Review*, 98(12):917–924.

Murphy, A. H. (1971). A note on the ranked probability score. *Journal of Applied Meteorology*, 10(1):155–156.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600.

Murphy, A. H. (1977). The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *Monthly Weather Review*, 105(7):803–816.

Murphy, A. H. (1993). What is a good forecast? an essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, 8(2):281–293.

Murphy, A. H. and Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review*, 115(7):1330–1338.

Namias, J. (1969). Seasonal interactions between the north pacific ocean and the atmosphere during the 1960's. *Monthly Weather Review*, 97(3):173–192.

Nash, J. E. and Sutcliffe, J. V. (1970). River flow forecasting through conceptual models; Part I – a discussion of principles. *Journal of Hydrology*, 10:282–290.

Negenborn, R. R., De Schutter, B., Wiering, M. A., and Hellendoorn, H. (2005). Learning-based model predictive control for markov decision processes. In P. Horacek, M. S. and Zitek, P., editors, *16th IFAC World Congress*, Prague, Czech Republic.

O'Kane, J. P. and Flynn, D. (2007). Thresholds, switches and hysteresis in hydrology from the pedon to the catchment scale: a non-linear systems theory. *Hydrology and Earth System Sciences*, 11(1):443–459.

Peirolo, R. (2010). Information gain as a score for probabilistic forecasts. *Meteorological Applications*, in print.

Pereira, M. V. F. and Pinto, L. (1991). Multi-stage stochastic optimization applied to energy planning. *Mathematical Programming*, 52(1):359–375.

Peterson, M. B. (2009). *An introduction to decision theory*. Cambridge University Press, Cambridge, UK.

Philbrick, C. R. and Kitanidis, P. K. (1999). Limitations of deterministic optimization applied to reservoir operations. *Journal of Water Resources Planning and Management*, 125(3):135–142.

Pianosi, F. (2008). *Novel methods for water reservoirs management*. PhD thesis, Politechnico di Milano.

Pianosi, F. and Ravazzani, G. (2010). Assessing rainfall-runoff models for the management of lake verbano. *Hydrological Processes*, 24(22):3195–3205.

Pianosi, F. and Soncini-Sessa, R. (2009). Real-time management of a multipurpose water reservoir with a heteroscedastic inflow model. *Water Resources Research*, 45(10):W10430.

Piechota, T. C., Chiew, F. H. S., Dracup, J. A., and McMahon, T. A. (1998). Seasonal streamflow forecasting in Eastern Australia and the El Nino–southern oscillation. *Water Resources Research*, 34(11):3035–3044.

Popper, K. R. (1959). The propensity interpretation of probability. *The British journal for the philosophy of science*, 10(37):25.

Popper, K. R. (1968). *The logic of scientific discovery*. Taylor & Francis e-Library, second edition.

Raso, L., Schwanenberg, D., van der Giesen, N., and van Overloop, P. (2010). Tree-Scenario Based Model Predictive Control. *Geophysical Research Abstracts*, 12:3178.

Rauch, W. and Harremoës, P. (1999). Genetic algorithms in real time control applied

to minimize transient pollution from urban wastewater systems. *Water Research*, 33(5):1265 – 1277.

Rissanen, J. (2007). *Information and complexity in statistical modeling*. Springer Verlag.

Rissanen, J. and Langdon, G. G. (1979). Arithmetic coding. *IBM Journal of Research and Development*, 23(2):149–162.

Robert, C. P. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Verlag, New York.

Rodriguez-Iturbe, I. and Rinaldo, A. (2001). *Fractal River Basins; chance and self-organization*. Cambridge University Press.

Roulston, M. S. and Smith, L. A. (2002). Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, 130(6):1653–1660.

Savenije, G. (2009). HESS Opinions: 'The art of hydrology'. *Hydrology and Earth System Sciences*, 13(2):157–161.

Savic, D. A. and Walters, G. A. (1997). Genetic algorithms for least-cost design of water distribution networks. *Journal of Water Resources Planning and Management*, 123(2):67–77.

Schaefli, B. and Gupta, H. V. (2007). Do Nash values have value? *Hydrological Processes*, 21(15):2075–2080.

Schmidt, M. and Lipson, H. (2009). Distilling free-form natural laws from experimental data. *science*, 324(5923):81.

Schoups, G., van de Giesen, N. C., and Savenije, H. H. G. (2008). Model complexity control for hydrologic prediction. *Water Resources Research*, 44:W00B03.

Schoups, G. and Vrugt, J. A. (2010). A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-gaussian errors. *Water Resources Research*, 46:W10531.

Schoups, G., Vrugt, J. A., Fenicia, F., and van de Giesen, N. C. (2010). Corruption of accuracy and efficiency of markov chain monte carlo simulation by inaccurate numerical implementation of conceptual hydrologic models. *Water Resources Research*, 46:W10530.

Schuurmans, W., Leeuwen, P. E. R. M. v., and Kruiningen, F. E. v. (2002). Automation of the rijnland storage basin, the netherlands. *Lowland Technology International*, 4(No. 1):13–20.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Schymanski, S. J., Sivapalan, M., Roderick, M. L., Hutley, L. B., and Beringer, J. (2009). An optimality-based model of the dynamic feedbacks between natural vegetation and the water balance. *Water Resources Research*, 45(1):W01412.

Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, 1(1):43–62.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical J.*, 27(3):379–423.

Shannon, C. E. (1956). A universal Turing machine with two internal states. *Automata studies*, pages 129–153.

Sharma, A. (2000). Seasonal to interannual rainfall probabilistic forecasts for improved

water supply management: Part 3–a nonparametric probabilistic forecast model. *Journal of Hydrology*, 239(1-4):249–258.

Shoemaker, C. A., Regis, R. G., and Fleming, R. C. (2007). Watershed calibration using multistart local optimization and evolutionary optimization with radial basis function approximation. *Hydrological Sciences Journal*, 52(3):450–465.

Singh, K. and Singh, V. P. (1991). Derivation of bivariate probability density functions with exponential marginals. *Stochastic Hydrology and Hydraulics*, 5(1):55–68.

Singh, V. P. (1997). The use of entropy in hydrology and water resources. *Hydrological Processes*, 11(6):587–626.

Singh, V. P. and Guo, H. (1995a). Parameter estimation for 3-parameter generalized pareto distribution by the principle of maximum entropy (POME). *Hydrological Sciences Journal*, 40(2):165–181.

Singh, V. P. and Guo, H. (1995b). Parameter estimations for 2-parameter pareto distribution by pome. *Water Resources Management*, 9(2):81–93.

Singh, V. P. and Guo, H. (1997). Parameter estimation for 2-parameter generalized pareto distribution by pome. *Stochastic Hydrology and Hydraulics*, 11(3):211–227.

Singh, V. P., Guo, H., and Yu, F. X. (1993). Parameter estimation for 3-parameter log-logistic distribution (LLD3) by pome. *Stochastic Hydrology and Hydraulics*, 7(3):163–177.

Singh, V. P. and Rajagopal, A. K. (1987). Some recent advances in application of the principle of maximum entropy (pome) in hydrology. *IAHS*, 194:353–364.

Singh, V. P. and Singh, K. (1985). Derivation of the pearson type (PT) III distribution by using the principle of maximum entropy (POME). *Journal of hydrology*, 80(3-4):197–214.

Solomatine, D. P. (1999). Random search methods in model calibration and pipe network design. *Water Industry Systems: Modelling and Optimization Applications*, pages 317–332.

Solomatine, D. P. and Ostfeld, A. (2008). Data-driven modelling: some past experiences and new approaches. *Journal of Hydroinformatics*, 10(1):3–22.

Solomonoff, R. J. (1964). A formal theory of inductive inference. Part I. *Information and control*, 7(1):1–22.

Solomonoff, R. J. (1978). Complexity-based induction systems: comparisons and convergence theorems. *Information Theory, IEEE Transactions on*, 24(4):422–432.

Soncini-Sessa, R., Castelletti, A., and Weber, E. (2007). *Integrated and participatory water resources management: theory*. Elsevier Science Ltd.

Sonuga, J. O. (1972). Principle of maximum entropy in hydrologic frequency analysis. *Journal of Hydrology*, 17(3):177 – 191.

Stedinger, J. and Kim, Y. O. (2002). Updating ensemble probabilities based on climate forecasts. *Proc., Water Resources Planning and Management (19-22 May, Roanoke, Virginia)(CD-Rom), Environmental and Water Resources Institute, American Society of Civil Engineers, Reston, VA*.

Stedinger, J. R. and Kim, Y. (2007). Adjusting ensemble forecast probabilities to reflect several climate forecasts. *IAHS PUBLICATION*, 313:188.

Stedinger, J. R. and Kim, Y.-O. (2010). Probabilities for ensemble forecasts reflecting climate information. *Journal of Hydrology*, 391(1–2):9–23.

Stedinger, J. R., Sule, B. F., and Loucks, D. P. (1984). Stochastic dynamic programming models for reservoir operation optimization. *Water Resources Research*, 20(11).

Stephenson, D. B., Coelho, C. A. S., and Jolliffe, I. T. (2008). Two extra components in the brier score decomposition. *Weather and Forecasting*, 23(4):752–757.

Storn, R. and Price, K. (1997). Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4):341–359.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press.

Szilard, L. (1964). On the decrease of entropy in a thermodynamic system by the intervention of intelligent beings. *Behavioral Science*, 9(4):301–310. translation from original German version (1929).

Tejada-Guibert, J. A., Johnson, S. A., and Stedinger, J. R. (1995). The value of hydrologic information in stochastic dynamic programming models of a multireservoir system. *Water Resources Research*, 31(10):2571–2579. Water Resour. Res.

Tilmant, A., Beevers, L., and Muyunda, B. (2010). Restoring a flow regime through the coordinated operation of a multireservoir system-The case of the Zambezi River Basin. *Water Resources Research*, 46(7):W07533.

Tilmant, A. and Kelman, R. (2007). A stochastic approach to analyze trade-offs and risks associated with large-scale water resources systems. *Water Resources Research*, 43(6):W06425.

Tilmant, A., Lettany, J., and Kelman, R. (2007). Hydrological Risk Assessment in the Euphrates-tigris River Basin: A Stochastic Dual Dynamic Programming Approach. *Water International*, 32(2):294–309.

Tilmant, A., Pinte, D., and Goor, Q. (2008). Assessing marginal water values in multipurpose multireservoir systems via stochastic programming. *Water Resources Research*, 44(12):W12431.

Toyabe, S., Sagawa, T., Ueda, M., Muneyuki, E., and Sano, M. (2010). Experimental demonstration of information-to-energy conversion and validation of the generalized Jarzynski equality. *Nature Physics*, 6:988–992.

Trenberth, K. E. (1997). The definition of El Nino. *Bulletin of the American Meteorological Society*, 78(12):2771–2777.

Tribus, M. (1961). *Thermostatics and thermodynamics*. D. Van Nostrand Company, Inc.

Turing, A. M. (1937). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 2(1):230.

van Andel, S. (2009). *Anticipatory Water Management: Using ensemble weather forecasts for critical events*. CRC Press/Balkema. PhD Thesis, Unesco-IHE.

van Andel, S. J., Price, R., Lobbrecht, A., and van Kruiningen, F. (2010). Modeling Controlled Water Systems. *Journal of Irrigation and Drainage Engineering*, 136:392.

van Andel, S. J., Price, R. K., Lobbrecht, A. H., van Kruiningen, F., and Mureau, R. (2008). Ensemble Precipitation and Water-Level Forecasts for Anticipatory Water-System Control. *Journal of Hydrometeorology*, 9(4):776–788.

van Overloop, P., Negenborn, R., de Schutter, B., and van de Giesen, N. (2010a). Predictive Control for National Water Flow Optimization in The Netherlands. *Intelligent Infrastructures*, pages 439–461.

van Overloop, P. J. (2006). *Model predictive control on open water systems*. PhD thesis, TU Delft, Delft.

van Overloop, P. J., Negenborn, R. R., Weijs, S. V., Malda, W., Bruggers, M. R., and De Schutter, B. (2010b). Linking water and energy objectives in lowland areas through the application of model predictive control. In *Proceedings of the 2010 IEEE Conference on Control Applications*, pages 1887–1891, Yokohama, Japan.

van Overloop, P. J., Weijs, S., and Dijkstra, S. (2008). Multiple model predictive control on a drainage canal system. *Control Engineering Practice*, 16(5):531–540.

Vapnik, V. N. (1998). *Statistical learning theory*. John Wiley & Sons, NY, USA.

Verlinde, E. (2010). On the Origin of Gravity and the Laws of Newton. *arXiv*, arXiv:1001.0785v1.

Vesterstrøm, J. and Thomsen, R. (2004). A comparative study of differential evolution, particle swarm optimization, and evolutionary algorithms on numerical benchmark problems. In *Proc. IEEE Congr. Evolutionary Computation*, pages 1980–1987, Portland, OR, USA.

Vrugt, J. A., Gupta, H. V., Bouten, W., and Sorooshian, S. (2003). A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. *Water Resources Research*, 39(8):1201.

Vrugt, J. A. and Robinson, B. A. (2007). Improved evolutionary optimization from genetically adaptive multimethod search. *Proceedings of the National Academy of Sciences*, 104(3):708.

Vrugt, J. A., Ter Braak, C. J. F., Gupta, H. V., and Robinson, B. A. (2009). Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling? *Stochastic environmental research and risk assessment*, 23(7):1011–1026.

Wand, M. P. and Jones, M. C. (1993). Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of the American Statistical Association*, 88(422):520–528.

Weijs, S. (2009). Interactive comment on "HESS Opinions 'A random walk on water' " by D. Koutsoyiannis. *Hydrology and Earth System Sciences Discussions*, 6:C2733–C2745.

Weijs, S., van Leeuwen, E., van Overloop, P. J., and van de Giesen, N. (2007). Effect of uncertainties on the real-time operation of a lowland water system in the netherlands. *IAHS PUBLICATION*, 313:463.

Weijs, S. V. (2004). 'Sturen met onzekere voorspellingen' (control using uncertain predictions), in Dutch. Master's thesis, TU Delft.

Weijs, S. V. (2007). Information content of weather predictions for flood-control in a dutch lowland water system. In *4th International Symposium on Flood Defense: Managing Flood Risk, Reliability and Vulnerability, Toronto, Ontario, Canada*.

Weijs, S. V., Schoups, G., and van de Giesen, N. (2010a). Why hydrological predictions should be evaluated using information theory. *Hydrology and Earth System Sciences*, 14(12):2545–2558.

Weijs, S. V. and Van de Giesen, N. (2011). Accounting for observational uncertainty

in forecast verification: an information–theoretical view on forecasts, observations and truth. *Monthly Weather Review*, early online release.

Weijs, S. V. and van de Giesen, N. (2011). "zipping" hydrological timeseries: An information-theoretical view on data compression as philosophy of science. *Geophysical Research Abstracts*, 13:EGU2011–8105.

Weijs, S. V., van Leeuwen, P., and van Overloop, P. (2006). The integration of risk analysis in real time flood control. In Cunge, J., Guinot, V., and Liong, S.-Y., editors, *7th International Conference on Hydroinformatics, Nice, France*, volume 4, pages 2943–2950.

Weijs, S. V., Van Nooijen, R., and Van de Giesen, N. (2010b). Kullback–Leibler divergence as a forecast skill score with classic reliability–resolution–uncertainty decomposition. *Monthly Weather Review*, 138(9):3387–3399.

Westra, S. and Sharma, A. (2010). An Upper Limit to Seasonal Rainfall Predictability? *Journal of Climate*, 23(12):3332–3351.

Wilks, D. S. (1995). *Statistical Methods in the Atmospheric Sciences: An Introduction*. Academic Press.

Wilks, D. S. (2000). On interpretation of probabilistic climate forecasts. *Journal of Climate*, 13(11):1965–1971.

Wilks, D. S. (2002). Realizations of daily weather in forecast seasonal climate. *Journal of Hydrometeorology*, 3(2):195–207.

Winsemius, H. C., Schaefli, B., Montanari, A., and Savenije, H. H. G. (2009). On the calibration of hydrological models in ungauged basins: A framework for integrating hard and soft hydrological information. *Water Resources Research*, 45(12):W12422.

Wood, A. W., Kumar, A., and Lettenmaier, D. P. (2005). A retrospective assessment of National Centers for Environmental Prediction climate model-based ensemble hydrologic forecasting in the western United States. *Journal of Geophysical Research*, 110:0148–0227.

Wood, A. W. and Lettenmaier, D. P. (2008). An ensemble approach for attribution of hydrologic prediction uncertainty. *Geophysical Research Letters*, 35(14):1–L14401.

Wood, E. F., Lettenmaier, D. P., and Zartarian, V. G. (1992). A land-surface hydrology parameterization with subgrid variability for general circulation models. *Journal of Geophysical Research*, 97(D3):2717–2728.

Xu, M., van Overloop, P. J., van de Giesen, N. C., and Stelling, G. S. (2010). Real-time control of combined surface water quantity and quality: polder flushing. *Water Science and Technology*, 61(4):869.

Yakowitz, S. (1982). Dynamic programming applications in water resources. *Water Resources Research*, 18(4):673–696.

Yeh, W. W.-G. (1985). Reservoir management and operations models: A state-of-the-art review. *Water Resources Research*, 21(12):1797–1818.

Zehe, E., Blume, T., and Blöschl, G. (2010). The principle of 'maximum energy dissipation': a novel thermodynamic perspective on rapid water flow in connected soil structures. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1545):1377.

Zhao, T., Cai, X., and Yang, D. (2011). Effect of streamflow forecast uncertainty on real-time reservoir operation. *Advances in Water Resources*, In Press, Corrected Proof.

Ziv, J. and Lempel, A. (1977). A universal algorithm for sequential data compression. *IEEE transactions on Information Theory*, 23(3):337–343.

Appendix A

Equivalence between MRE-update and pdf-ratio solutions for the normal case

We try to solve

$$\min_{q_i}\{\sum_{i=1}^{n} q_i \log(\frac{q_i}{p_i})\}$$

subject to the constraints

$$\begin{aligned}
\sum_{i=1}^{n} q_i &= 1 \\
q_i &\geq 0 \\
\sum_{i=1}^{n} q_i x_i &= \mu_1 \\
\sum_{i=1}^{n} q_i (x_i - \mu_1)^2 &= \sigma_1^2
\end{aligned}$$

this leads to the Lagrangeans $\forall i$

$$\begin{aligned}
\frac{\partial}{\partial q_i}\left\{\sum_{i=1}^{n} q_i \log(\frac{q_i}{p_i}) + \lambda_1\left(\sum_{i=1}^{n} q_i - 1\right) + \lambda_2\left(\sum_{i=1}^{n} q_i (x_i-\mu_1)^2 - \sigma_1^2\right) + \lambda_3\left(\sum_{i=1}^{n} q_i x_i - \mu_1\right)\right\} &= 0 \\
1 + \log q_i - \log p_i + \lambda_1 + \lambda_2 (x_i - \mu_1)^2 + \lambda_3 x_i &= 0 \\
\log p_i - 1 - \lambda_1 - \lambda_2 (x_i - \mu_1)^2 - \lambda_3 x_i &= \log q_i
\end{aligned}$$

where λ_1, λ_2 and λ_3 are the Lagrange multipliers correspronding to the constraints for the sum, the mean and the variance, respectively. The constraint for nonnegativity is never binding and is left out of the Lagrangean. This leads to a solution of the form

$$q_i = p_i e^{-1-\lambda_1-\lambda_2(x_i-\mu_1)^2-\lambda_3 x_i} \tag{A.1}$$

The Lagrange multiplier can subsequently be solved numerically to match the constraints.

The result of the pdf-ratio method for normal initial and target distributions has the same form as equation A.1, but not necessarily the same values for λ_1, λ_2 and λ_3, because the solution is not forced to satisfy the constraints. When the parameters of the target normal distribution that is used as an input for the pdf-ratio method are modified so that the moments of the resultant weighted ensembles match μ_1 and σ_1 exactly, the result of

the pdf-ratio method matches exactly the result of the MRE-update. The search for the two parameters of the normal distribution and the normalization constant are exactly the three degrees of freedom that are required to find λ_1, λ_2 and λ_3 numerically. An analytical solution is not possible because the constants depend on all individual x_i.

Similarly, for the Croley parametric method with the same constraints, and objective function

$$\min_{q_i} \left\{ \sum_{i=1}^{n} (q_i - p_i)^2 \right\}$$

it can be found that the solution is of the form

$$\begin{aligned} \frac{\partial}{\partial q_i} \left\{ \sum_{i=1}^{n} (q_i - p_i)^2 + \lambda_1 \left(\sum_{i=1}^{n} q_i - 1 \right) + \lambda_2 \left(\sum_{i=1}^{n} q_i (x_i - \mu_1)^2 - \sigma_1^2 \right) + \lambda_3 \left(\sum_{i=1}^{n} q_i x_i - \mu_1 \right) + \lambda_i q_i \right\} &= 0 \\ 2q_i - 2p_i + \lambda_1 + \lambda_2 (x_i - \mu_1)^2 + \lambda_3 x_i + \lambda_i &= 0 \\ 2p_i - \lambda_1 - \lambda_2 (x_i - \mu_1)^2 - \lambda_3 x_i + \lambda_i &= 2q_i \end{aligned}$$

where λ_i are the extra Lagrange multipliers which ensure that all q_i are nonnegative. The solution to the Croley method are weights that are a parabolic function of the values x_i. The quadratic objective function thus leads to a solution that is a second order (quadratic polynomial) approximation of the solution found by the MRE-update.

Appendix B

The decomposition of the divergence score

First we use the definition of the Kullback-Leibler divergence to define the total score and the resolution and reliability components and the entropy for the uncertainty component.

$$D_{KL}\left(\mathbf{v} \shortparallel \mathbf{w}\right) = \sum_{i=1}^{n} v_i \log \frac{v_i}{w_i}$$

$$DS = \frac{1}{N}\sum_{t=1}^{N} D_{KL}\left(\mathbf{o}_t \shortparallel \mathbf{f}_t\right)$$

$$REL = \frac{1}{N}\sum_{k=1}^{K} n_k D_{KL}\left(\bar{\mathbf{o}}_k \shortparallel \bar{\mathbf{f}}_k\right)$$

$$RES = \frac{1}{N}\sum_{k=1}^{K} n_k D_{KL}\left(\bar{\mathbf{o}}_k \shortparallel \bar{\mathbf{o}}\right)$$

$$UNC = \frac{1}{N}\sum_{t=1}^{N} H\left(\bar{\mathbf{o}}\right) \;=\; -\frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{n}\left\{\left[\bar{\mathbf{o}}\right]_i \log\left[\bar{\mathbf{o}}\right]_i\right\}$$

Now we can simplify the expression for

$$\begin{aligned} REL - RES &= \sum_{k=1}^{K} n_k \left\{ D_{KL}\left(\bar{\mathbf{o}}_k \shortparallel \overline{\mathbf{f}}_k\right) - D_{KL}\left(\bar{\mathbf{o}}_k \shortparallel \bar{\mathbf{o}}\right)\right\} \\ &= \sum_{k=1}^{K} n_k \sum_{i=1}^{n} \left[\bar{\mathbf{o}}_k\right]_i \left\{ \log \frac{\left[\bar{\mathbf{o}}_k\right]_i}{\left[\bar{\mathbf{f}}_k\right]_i} - \log \frac{\left[\bar{\mathbf{o}}_k\right]_i}{\left[\bar{\mathbf{o}}\right]_i} \right\} \\ &= \sum_{k=1}^{K} n_k \sum_{i=1}^{n} \left[\bar{\mathbf{o}}_k\right]_i \left\{ \log \frac{\left[\bar{\mathbf{o}}\right]_i}{\left[\bar{\mathbf{f}}_k\right]_i} \right\} \end{aligned}$$

Note that

$$DS = \frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{n} \left[\mathbf{o}_t\right]_i \left\{ \log \frac{\left[\mathbf{o}_t\right]_i}{\left[\mathbf{f}_t\right]_i} \right\}$$

With K equal to the number of different $\mathbf{f}_t$ and one bin for each $\mathbf{f}_t$ we can label both bins and outcomes by k. We label the outcomes in a bin by

$$[\mathbf{o}_t]_{k,m_k}$$

with $m_k = 1 \ldots n_k$ so

$$DS = \frac{1}{N} \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{m_k=1}^{n_k} [\mathbf{o}_{k,m_k}]_i \left\{ \log \frac{[\mathbf{o}_{k,m_k}]_i}{\left[\bar{\mathbf{f}}_k\right]_i} \right\}$$

which can be written as

$$\begin{aligned} DS &= \tfrac{1}{N} \sum_{i=1}^{n} \sum_{k=1}^{K} \left(\sum_{m_k=1}^{n_k} [\mathbf{o}_{k,m_k}]_i \left\{ \log \frac{[\mathbf{o}_{k,m_k}]_i}{[\mathbf{f}_t]_i} - \log \frac{[\mathbf{o}_{k,m_k}]_i}{[\bar{\mathbf{o}}]_i} + \log \frac{[\mathbf{o}_{k,m_k}]_i}{[\bar{\mathbf{o}}]_i} \right\} \right) \\ &= \tfrac{1}{N} \sum_{i=1}^{n} \sum_{k=1}^{K} \left(\sum_{m_k=1}^{n_k} [\mathbf{o}_{k,m_k}]_i \left\{ \log \frac{[\bar{\mathbf{o}}]_i}{[\mathbf{f}_t]_i} + \log \frac{[\mathbf{o}_{k,m_k}]_i}{[\bar{\mathbf{o}}]_i} \right\} \right) \\ &= \tfrac{1}{N} \sum_{i=1}^{n} \sum_{k=1}^{K} \left(n_k \left[\bar{\mathbf{o}}_k\right]_i \left\{ \log \frac{[\bar{\mathbf{o}}]_i}{[\mathbf{f}_k]_i} \right\} \right) + \frac{1}{N} \sum_{i=1}^{n} \sum_{k=1}^{K} \left(\sum_{m_k=1}^{n_k} [\mathbf{o}_{k,m_k}]_i \left\{ \log \frac{[\mathbf{o}_{k,m_k}]_i}{[\bar{\mathbf{o}}]_i} \right\} \right) \end{aligned}$$

we can now recognize the first term as $REL - RES$, so

$$\begin{aligned} DS - (REL - RES) &= \\ \frac{1}{N} \sum_{i=1}^{n} \sum_{k=1}^{K} \left(\sum_{m_k=1}^{n_k} [\mathbf{o}_{k,m_k}]_i \left\{ \log \frac{[\mathbf{o}_{k,m_k}]_i}{[\bar{\mathbf{o}}]_i} \right\} \right) &= \\ \frac{1}{N} \sum_{t=1}^{N} \left(\sum_{i=1}^{n} [\mathbf{o}_t]_i \left\{ \log \frac{[\mathbf{o}_t]_i}{[\bar{\mathbf{o}}]_i} \right\} \right) &= \tfrac{1}{N} \textstyle\sum_{t=1}^{N} \quad D_{KL}\left(\mathbf{o}_t \,\|\, \bar{\mathbf{o}}\right) \end{aligned}$$

Note that, with

$$\lim_{x \downarrow 0} x \log x = 0$$

and for $n = 2$,$\mathbf{o}_t \in \left\{ (1,0)^T, (0,1)^T \right\}$ we find

$$\begin{aligned} \sum_{i=1}^{n} [\mathbf{o}_t]_i \left\{ \log \frac{[\mathbf{o}_t]_i}{[\bar{\mathbf{o}}]_i} \right\} &= \sum_{i=1}^{n} \left\{ [\mathbf{o}_t]_i \log [\mathbf{o}_t]_i - [\mathbf{o}_t]_i \log [\bar{\mathbf{o}}]_i \right\} \\ &= -\sum_{i=1}^{n} [\mathbf{o}_t]_i \log [\bar{\mathbf{o}}]_i \end{aligned}$$

so

$$\sum_{t=1}^{N} \left(\sum_{i=1}^{n} [\mathbf{o}_t]_i \left\{ \log \frac{[\mathbf{o}_t]_i}{[\bar{\mathbf{o}}]_i} \right\} \right) = N \sum_{i=1}^{n} [\bar{\mathbf{o}}]_i \log [\bar{\mathbf{o}}]_i$$

so

$$DS - (REL - RES) = -\sum_{i=1}^{n} [\bar{\mathbf{o}}]_i \log [\bar{\mathbf{o}}]_i = H(\bar{\mathbf{o}}) = UNC$$

Appendix C

Relation divergence score and doubling rate in a horse race

In this example we consider a horse race, a gambler and a bookmaker. Suppose there are n possible events (horses winning). A bookmaker is offering odds r_i, which means that if horse i wins, the gambler receives r_i times the money he bet on horse i. The fraction of the gamblers wealth bet on horse i is denoted by b_i. If another horse wins, the gambler loses his stake b_i. A bookmaker is said to offer fair odds if

$$\sum_{i=1}^{n} \frac{1}{r_i} = 1$$

. After one race, the outcome of the race can be described by a vector $\mathbf{o}$. Element o_i of this vector is 1 if horse i wins and 0 otherwise. The factor by which the wealth of the gambler has grown after one race is

$$S_t = \sum_{i=1}^{n} o_i b_i r_i$$

. If the gambler reinvests all his money in each new bet, the expected factor by which the gamblers wealth has grown after T bets will be

$$S_T = \prod_{t=1}^{T} S_t$$

.The expectation of the logarithm of this factor is defined as the doubling rate W

$$W = E\{log_2 S_t\} = \sum_{i=1}^{n} \{p_i log_2(b_i r_i)$$

, in which p_i is the probability that horse i wins the race. The expected wealth after T bets can now be written as

$$S_T = 2^{TW}.$$

Kelly (1956) showed that W is maximized by following a proportional betting strategy ($b_i = p_i$). When following this strategy it is possible to express W as a difference between

two Kullback Leibler divergences (Cover and Thomas, 2006). When the bookmaker is offering fair odds, his estimates of the win probabilities can be written as $h_i = 1/r_i$.

$$\begin{aligned} W &= \sum_{i=1}^{n} \{p_i log_2(b_i h_i) \\ &= \sum_{i=1}^{n} \{p_i log_2(\frac{b_i}{p_i}\frac{p_i}{h_i}) \end{aligned}$$

$$= D_{KL}\left(\mathbf{p} \shortparallel \mathbf{h}\right) - D_{KL}\left(\mathbf{p} \shortparallel \mathbf{b}\right)$$

This means that a gambler can make money only if his probability estimate is better than the bookie's. Distribution $\mathbf{p}$ in these divergences is the true distribution. This truth must be conditioned at least on each combination of $\mathbf{h}, \mathbf{b}$.

Now let's assume that the bookmaker offers fair odds with respect to climatology. Because in this case, one of the forecasters (the bookie) always issues the same forecast, $\mathbf{p}$ is the distribution of observations $\bar{\mathbf{o}}$, conditioned on $\mathbf{b}$ only, yielding the conditional distribution of the outcome $\bar{\mathbf{o}}_k$. The bookie's estimate $\mathbf{h}$ is equal to $\bar{\mathbf{o}}$ and W can be written as

$$W = D_{KL}\left(\bar{\mathbf{o}}_k \shortparallel \bar{\mathbf{o}}\right) - D_{KL}\left(\bar{\mathbf{o}}_k \shortparallel \mathbf{f}_k\right)$$

which can be recognized as the resolution minus the reliability term. W can thus be seen as the information gain towards the truth, compared to climate.

$$W = RES - REL$$

It is also possible to condition the truth further, to arrive at a more general expression for W, which is also valid in case the bookmaker's estimate is different from climate. In this case the distributions are conditioned on every single forecast, leading to the expression

$$W = D_{KL}\left(\mathbf{o_t} \shortparallel \mathbf{h_t}\right) - D_{KL}\left(\mathbf{o_t} \shortparallel \mathbf{b_t}\right)$$

which can be recognized as the difference in divergence score between bookie and gambler. A gambler can thus make money with growth rate

$$W = DS_{bookie} - DS_{gambler}$$

Acknowledgements

The information in this thesis did not spring from nothing, but was a result of a long evolutionary process involving random brainstorms, selection, feedback and other forms of communication, love and support received from more persons than I can mention here.

First of all I want to thank my promotor, Nick van de Giesen, for providing me with the opportunity to do this research. Nick, thank you for the support, sense of humour, trust, thoughts, critical feedback and freedom you gave me. Peter-Jules van Overloop, thank you very much for convincing me to do a Ph.D. in Delft and your very generous support, both in science and personally. I could always rely on your help. Sharing an office with you and my fellow Ph.D. students Luciano Raso and Xu Min was hilarious and inspiring.

Gerrit Schoups, Ronald van Nooijen, Rolf Hut, Nico de Vos and Nick, thanks for listening to my semi-religious sermons about information theory and brainstorming about crazy ideas and experiments. You certainly made sure I reached my 10% foolishness quotum to sustain my curiosity-addiction with enough interesting questions.

Thank you Hanneke and Betty for providing regularity and organisation in this chaotic environment of scientists. The accurately timed lunch-calls and fruit you supplied kept me going. My other colleagues, especially my fellow Ph.D. students, for being a bright bunch of nice people. Martine, thank you for sharing, support and surprises. Your information will not be lost.

Thanks to the students of the dispuut water management for creating a great positive atmosphere with many motivated students and for giving me the opportunity to accompany the brilliantly organized study trip to Argentina.

All the people of the Nieuwelaan for building and sharing a home together. My present and past housemates at the Nerdhuis, for building the binary encoded π-tile floor, crazy experiments, all your great cooking, random dinner conversations, the honor of receiving the nerd of the day award on multiple occasions, and much more. Nerds 2^2 ever!

Bart Nijssen and Andy Wood of 3TIER inc., ICIMOD, KNMI, Meteoconsult, Hoogheemraadschap van Delfland and Nelen & Schuurmans for kindly providing the data used in this research. Jery Stedinger, Amaury Tilmant, Hoshin Gupta, Federico Lombardo, Demetris Koutsoyiannis, Francesca Pianosi, Vijay P. Singh, and several anonymous reviewers for their constructive comments on my papers. The members of the examination committee for their comments that helped to improve my draft thesis.

My paranymphs Alexander Bakker and Jan Jongerden, for feedback on early versions of my chapters and joining me on hitchhiking trips to my first conference in Nice and other locations. Thanks to all my other friends who made this period a happy one and to some for sharing the idea that camping at subzero temperatures is fun. Juan Villarino, por compartir parte del circuito infinito. Sharing these travels of randomness allowed me to regain my focus.

I am deeply grateful to my mother and father and brother Menno, for growing up happily, for informing or not informing about my thesis on the right moments, letting me choose my own path and teaching me love for nature and the Balkans and to never stop wondering.

During the writing this acknowledgement, I noticed a small muscle ache developing as a result of the constant smile when thinking back of all the great moments I shared with all of you. Again I want to express my deepest gratitude to you. Last and most of all, I want to thank my girlfriend Tamara for her patience, unconditional support, endurance, enthusiasm and love, which were essential ingredients for finishing this thesis. I am glad to have you on my side in our new adventure.

About the author

Steven Weijs received his first information through a partially random, but well-selected genetic code, which, notwithstanding his mother's fruitless frantic search for yoghurt during holiday in Czechoslovakia, lead to the emergence of a little creature that began receiving visual information on February 12, 1979 in Groningen. After a bit more than two years, he was joined by his younger brother. In the years that followed, he eagerly gathered more information by asking his parents and teachers progressively more foolish questions that were not always easy to answer. He was very curious about the motivation behind his father's research, which according to Steven consisted of "looking how rabbits chew".

During secondary school, Steven became interested in electronic circuits and soon he was also transmitting low-information content (a novelty in those days) through his home-made radio transmitter. He came to TU Delft to study Civil Engineering in 1997. After working as a student assistant, following M.Sc. courses in both water management and hydrology, and designing a flood defense for a town in Argentina during his internship, he obtained his M.Sc. in water resources management cum laude in 2004.

He then joined Nelen & Schuurmans Hydroinformatics, an engineering consultant, where he worked on hydrological / hydraulic modeling and the development of decision support and control systems for several Dutch water boards. After two years, in 2006, Steven returned to academia, starting his Ph.D. research at the chair of Water Resources Management. Now convinced that modeling and control will always lack enough information to fully eliminate it, he now aimed for at least trying to *understand* uncertainty. This thesis is a result of this journey.

In his spare time, Steven enjoys hiking and hitchhiking in desolate landscapes. Randomness was one of the key ingredients of his travels through South America, which taught him the beauty of uncertainty and sub-optimal decisions. Now he will embark on a slightly more organized adventure in Switzerland. After finalizing his Ph.D. Steven will join the EFLUM lab of Marc Parlange at the EPFL in Lausanne, where he will work as a post-doc after having obtained a post-doctoral grant from the AXA research fund.

Publications

Peer reviewed publications

S.V. Weijs and N. Van de Giesen. Accounting for observational uncertainty in forecast verification: an information–theoretical view on forecasts, observations and truth. *Monthly Weather Review*, early online release, 2011.

S.V. Weijs, G. Schoups, and N. van de Giesen. Why hydrological predictions should be evaluated using information theory. *Hydrology and Earth System Sciences*, 14(12):2545–2558, 2010.

R.W. Hut, S.V. Weijs, and W.M.J. Luxemburg. Using the Wiimote as a sensor in water research. *Water Resources Research*, 46(12):W12601, 2010.

S.V. Weijs, R. Van Nooijen, and N. Van de Giesen. Kullback–Leibler divergence as a forecast skill score with classic reliability–resolution–uncertainty decomposition. *Monthly Weather Review*, 138(9): 3387–3399, September 2010.

P.J. van Overloop, S. Weijs, and S. Dijkstra. Multiple model predictive control on a drainage canal system. *Control Engineering Practice*, 16(5):531–540, 2008.

S. Weijs, E. van Leeuwen, P. van Overloop, and N. van de Giesen. Effect of uncertainties on the real-time operation of a lowland water system in The Netherlands. *IAHS PUBLICATION*, 313:463, 2007.

Non-reviewed publications

P.J. van Overloop, R.R. Negenborn, S.V. Weijs, W. Malda, W., M.R. Bruggers and B. De Schutter. Linking water and energy objectives in lowland areas through the application of model predictive control. In *Proceedings of the 2010 IEEE Conference on Control Applications*, pages 1887–1891, Yokohama, Japan, 2010.

S. Weijs. Interactive comment on "HESS Opinions 'A random walk on water' " by D. Koutsoyiannis. *Hydrology and Earth System Sciences Discussions*, 6:C2733–C2745, 2009.

S.V. Weijs and N.C. van de Giesen. Information theory, uncertainty and risk for evaluating hydrologic forecasts. In *International Workshop Advances in Statistical Hydrology (STAHY), Taormina, Italy*, 2010.

S.V. Weijs and M.M. Rutten. Application of minimum relative entropy update to long term forecast of cooling water problems in the Rhine. In *8th International Conference on Hydroinformatics, Concepcion, Chile*, 2009.

S. Weijs. Information content of weather predictions for flood-control in a dutch lowland water system. In *4th International Symposium on Flood Defense: Managing Flood Risk, Reliability and Vulnerability, Toronto, Ontario, Canada*, 2007.

S.V. Weijs. The value of short-term hydrological predictions for operational management of a dutch lowland water system. In *International Conference on Water and Flood Management, Dhaka, Bangladesh*, 2007.

S.V. Weijs, P.E.R.M. van Leeuwen, and P.J. van Overloop. The integration of risk analysis in real time flood control. In Jean Cunge, Vincent Guinot, and Shie-Yui Liong, editors, *7th International Conference on Hydroinformatics, Nice, France*, volume 4, pages 2943–2950, 2006.

Abstracts

S.V. Weijs and N. van de Giesen. "zipping" hydrological timeseries: An information-theoretical view on data compression as philosophy of science. *Geophysical Research Abstracts*, 13:EGU2011–8105, 2011.

S.V. Weijs and N. van de Giesen. Evaluating reliability and resolution of ensemble forecasts using information theory. *Geophysical Research Abstracts*, 12:EGU2010-6489, 2010.

S.V. Weijs. Minimum relative entropy update of ensemble probabilities to reflect forecast information. *Geophysical Research Abstracts*, 10:EGU2008-A-01778, 2008.

S.V. Weijs. Timescales and information in early warning system design for glacial lake outburst floods. In A.G. van Os, editor, *Proceedings of the NCR-days 2007, a sustainable river system?!*, number NCR publication 32, 2007.