Technische Universiteit Delft
Faculteit Elektrotechniek, Wiskunde en Informatica
Delft Institute of Applied Mathematics

Voorspellen van veel neerslag

gebruikmakend van regressieanalyse en de ECMWF-modeluitvoer

(Engelse titel: Prediction of high precipitation using regression analysis and ECMWF model output)

Verslag ten behoeve van het
Delft Institute of Applied Mathematics
als onderdeel ter verkrijging

van de graad van

BACHELOR OF SCIENCE
in
TECHNISCHE WISKUNDE

door

PASCALLE WIJNTJES

Delft, Nederland
Juni 2017

Abstract

In deze scriptie is onderzocht of regressie, een postprocessing methode, een goede toevoeging is aan het model dat het Europees Centrum voor Weersverwachtingen op Middellange Termijn (ECMWF) ontwikkeld heeft en waarvan het KNMI de uitvoer gebruikt, specifiek om te bepalen of er minstens 5 mm neerslag op een dagdeel (00-12 UTC en 12-00 UTC) zal vallen. Hiervoor is gebruik gemaakt van logistische regressie. Voor het verifiëren van de resultaten is er gekeken naar de Brier-Scores en de reliability diagrammen van de regressiemodellen. Deze zijn vervolgens vergeleken met de resultaten van het ECMWF-model. De resultaten zijn vrijwel gelijk, maar de Brier-Scores van het originele model zijn net iets beter dan die van de regressiemodellen. De conclusie die na het onderzoek wordt getrokken is dat het gemaakte regressiemodel geen toevoeging heeft op het ECMWF-model.

Inhoudsopgave

1 Introductie **6**

2 Weersverwachtingen **8**

3 Historische data-analyse **9**

4 Post-processing analyse met logistische regressie **12**
4.1 Lineaire regressie 12
4.2 Gegeneraliseerd Lineair Model 13
4.3 Toepassen van het GLM 14
4.4 Twee-Stedenmodel 17
4.5 Verificatie van de modellen 18
4.6 Generaliseerbaarheid van het één-stad-model 23

5 Conclusie **26**

6 Discussie **26**

A Histogram van de natte dagen **29**

B Correlatiematrices, De Bilt **31**

C Correlatiematrices, Schiphol **33**

D Goodness of Fit **35**

E R-Code **36**

F Glossary **55**

1 Introductie

In Nederland valt steeds meer neerslag. In ongeveer 100 jaar is de jaarlijkse hoeveelheid neerslag in Nederland toegenomen met 26 procent, mede door klimaatverandering [1]. Daarnaast blijkt dit ook te leiden tot een intensiteitstoename van zware buien (in korte tijd veel regen). Per graad opwarming neemt de intensiteit met ongeveer 12 procent toe. Het is daarom belangrijk of het voorspellen van veel neerslag binnen een dagdeel kan worden geoptimaliseerd met een niet tijdsintensief programma.

Hier wordt naar gekeken middels een statische postprocessing methode, regressie. Dit is een techniek welke de numerieke weersverwachtingsoutput over het algemeen verbetert. Het zal de systematische fouten uit het model halen[2].

De weersverwachtingen is waarschijnlijk het eerste waar je aan denkt bij het Koninklijk Nederlands Meteorologisch Instituut (KNMI) . Dit is ook een groot onderdeel binnen het instituut en hetgeen waar ik verder mee ga in mijn scriptie. Ik ga onderzoeken of regressieanalyse een goede methode is om te voorspellen of er meer dan 5 mm neerslag zal vallen op een dagdeel (00-12 gecoördineerde wereldtijd (UTC) en 12-00 UTC). Daarnaast zal ik onderzoeken of dit ook met twee stations tegelijk kan worden voorspeld. Daarom ga ik in deze scriptie de volgende onderzoeksvraag beantwoorden: *Kan de verwachting op minstens 5 mm neerslag in een dagdeel (00-12 UTC en 12-00 UTC) gemaakt met het Europees Centrum voor Weersverwachtingen op Middellange Termijn (ECMWF) systeem worden verbeterd door gebruik te maken van regressieanalyse?* Hiervoor zal gebruik gemaakt worden van de meetstationsgegevens in De Bilt en Schiphol en zullen de volgende deelvragen in deze scriptie aanbod komen:

1. Hoe wordt een weersverwachting bij het KNMI gemaakt? (Sectie 2)

2. Is er een droog en nat seizoen in Nederland? (Sectie 3)

3. Wat is regressieanalyse en hoe kan dit hier worden toegepast? (Sectie 4 & 4.2)

4. Is een regressiemodel een goede toevoeging aan het ECMWF -systeem? (Subsectie 4.3 & 4.5)

5. Kan met de gegevens van de twee steden een regressiemodel worden opgesteld om te voorspellen of er in beide steden meer dan 5 mm neerslag valt? En een model om te voorspellen of er in minstens één van beide steden tenminste 5 mm neerslag valt? (Subsectie 4.4)

6. Is het één-stad-model generaliseerbaar? (Subsectie 4.6)

Deelvraag 2 is relevant voor het opdelen van de data en daarmee het splitsen van de regressiemodellen. Wanneer er een nat en droog seizoen in Nederland voorkomt, is het slim om één model op te stellen per seizoen.

Voor deelvraag 4 wordt er gekeken naar verschillende verificatiemethoden die veel gebruikt worden bij weersvoorspellingen. De resultaten verkregen door het regressiemodel worden vergeleken met de verificatieresultaten van het ECMWF-model.

De meetstations die in deze scriptie worden gebruikt liggen relatief dicht bij elkaar. Hierdoor wordt verwacht dat het weer in Schiphol effect heeft op het weer in De Bilt en vice versa. Daarom is het interessant om deze twee steden te combineren bij regressie (deelvraag 5) en om te kijken of een model dat is gemaakt op basis van de ene stad, ook toepasbaar is op de andere stad (deelvraag 6).

Voor het opstellen van de regressiemodellen en het uitvoeren van de voorspelling zal het programma "R"worden gebruikt. Dit programma is speciaal ontwikkeld voor statistische berekeningen en om dit te visualiseren. De codes die voor dit onderzoek zijn geschreven zijn te vinden in Bijlage E.

2 Weersverwachtingen

Het Europees Centrum voor Weersverwachtingen op Middellange Termijn (ECMWF) [3] heeft een mondiaal computermodel ontwikkeld om de toestand van de atmosfeer te verwachten. Het KNMI maakt gebruik van de uitvoer van dit model voor de lange-termijn verwachtingen (tot 15 dagen vooruit).

Voor het maken van zo'n model wordt gebruik gemaakt van numerieke methoden. De driedimensionale numerieke berekeningen gebruiken boxen die dicht bij de aarde een hogere verticale resolutie hebben dan bovenin de atmosfeer. Voor de weersverwachting worden 51 berekeningen gemaakt (het ensemble, de 51 weerverwachtingsberekeningen samen (ENS)), welke steeds verschillen in beginvoorwaarde door hier storingen aan toe te voegen. Een visualisatie van de 51 berekeningenuitkomsten is te zien in de pluimgrafieken op de website van het KNMI [4]. Het ensemble bestaat uit 51 zogeheten "leden".

Uit dit systeem komt de weersverwachtingsdata die gebruikt wordt bij dit onderzoek. Een voorbeeld van de gegevens die verkregen worden voor de neerslagverwachting is gegeven in Tabel 2.1.

datum	RRENS	RRSDEV	RR6_46	RRGT00	RRGE03	RRGE15	RRGE50	RRGE100
20140101	26.00	11.00	29.00	100.00	100.00	90.00	4.00	0.00
20140102	29.00	10.00	27.00	100.00	100.00	92.00	4.00	0.00
20140103	1.00	1.00	2.00	59.00	10.00	0.00	0.00	0.00
20140104	42.00	15.00	39.00	100.00	100.00	98.00	29.00	0.00
20140105	5.00	4.00	11.00	98.00	59.00	4.00	0.00	0.00
				⋮				

Tabel 2.1: Voorspellingsdata van de 12-uursverwachting in De Bilt.

Hieronder staat de uitleg voor elke variabele van de voorspellingsdata.

datum: De validatiedatum van de berekening. Deze datum geeft het eind van het etmaal weer. Voor 12-00 UTC betekent dit dat in de kolom een dag later staat aangegeven dan dat de berekening voor is gemaakt.

RRENS: De hoeveelheid neerslag dat gemiddeld over ENS valt in 0.1 mm.

RRSDEV: De standaarddeviatie van de hoeveelheid neerslag over ENS in 0.1 mm

RR6_46: D waarde van lid 6 wordt afgetrokken van lid 46 van het ENS in 0.1 mm.

RRGT00: Het percentage ensembleleden dat meer dan 0 mm neerslag voorspelt.

RRGE03: Het percentage ensembleleden die minstens 0.3 mm neerslag voorspellen.

RRGE15: Het percentage ensembleleden die minstens 1.5 mm neerslag voorspellen.

RRGE50: Het percentage ensembleleden die minstens 5.0 mm neerslag voorspellen.

RRGE100: Het percentage ensembleleden die minstens 10.0 mm neerslag voorspellen.

Op twee momenten van de dag wordt een verwachting gemaakt (om 00 en 12 UTC), maar in dit onderzoek wordt uitsluitend gebruik gemaakt van de verwachting die start om 12 UTC . Het etmaal wordt hier opgedeeld in twee 12-uursperiodes: 00-12 UTC en 12-00 UTC en er wordt alleen gekeken naar de korte termijn verwachtingen (tot 48 uur vooruit). Onder de eerste periode vallen daarom de +24 en +48 uursverwachtingen en onder de tweede periode vallen de +12 en +36 uursverwachtingen. Voor de neerslag wordt een tabel als in Tabel 2.1 verkregen. Van deze tabellen worden de kolommen gebruikt als onafhankelijke variabelen van de regressiemodellen. Daarnaast worden de gegevens in de RRGE50-kolom gebruikt om het regressiemodel mee te vergelijken. Meer informatie over dit systeem is te vinden op de website van het ECMWF [3] en in de *User guide to ECMWF forecast products* [5].

3 Historische data-analyse

Naast weersverwachtingsdata is er ook historische data nodig om regressie te kunnen uitvoeren. Deze zijn namelijk de observaties en daarmee de responsvariabelen voor de latere regressiemodellen. Meer informatie hierover vindt u in Sectie 4. Een voorbeeld van de historische data in De Bilt verkregen van de website van het KNMI [6] is weergegeven in Tabel 3.1.

STN	YYYYMMDD	HH	RH
260	20140101	1	-1
260	20140101	2	-1
260	20140101	3	0
260	20140101	4	0
260	20140101	5	0
	⋮		

Tabel 3.1: Historische data van De Bilt.

De kolommen in de tabel hebben de volgende betekenissen:

STN: Het meetstationnummer. 260 is voor De Bilt en 240 voor Schiphol.

YYYYMMDD: De datum waarop de gevallen neerslag is gemeten (YYYY = jaar, MM = maand, DD = dag).

HH: Het tijdstip van de dag in uren (UTC) waarop de meting is gedaan.

RH: De hoeveelheid neerslag dat is gevallen binnen het bijbehorende uur, HH, gegeven in 0.1 mm. In de kolom wordt een -1 weergegeven wanneer er minder dan 0.05 mm is gevallen.

Voordat een voorspelling kan worden gedaan of er veel neerslag zal vallen op een dag, is het belangrijk om te kijken of een bepaalde periode van het jaar in Nederland natter is dan een andere periode. Dit kan namelijk effect hebben op de vorm van het model.

De verwachting is dat er in de winter vaker neerslag valt dan in de zomer, maar dat de neerslag in de zomer intenser is. Door gebruik te maken van de historische data verkregen via het KNMI , kan dit inzichtelijk worden gemaakt. Er is gebruik gemaakt van de variabele uursom van de neerslag (in 0.1 mm) (-1 voor <0.05 mm) (RH) tussen 1 januari 1987 en 31 december 2016.

De uiteindelijke resultaten van deze analyse zijn weergegeven in Tabel 3.2 en 3.3. Voor het maken van de tabel zijn eerst de irrelevante waarden uit de dataset verwijderd. Zo zijn alle waarden gelijk aan -1 vervangen door nullen. Er wordt dus geen rekening gehouden met zeer weinig neerslag. De hoeveelheid gevallen neerslag op die dag is dan zo weinig dat het afgerond op één decimaal ook een waarde van 0.0 mm aanneemt. Naast de tabellen is de hoeveelheid neerslag van meer dan 0 mm per maand gevisualiseerd door deze in histogrammen te zetten. Deze zijn te zien in Bijlage A.

In de resultaten valt op dat in De Bilt ongeveer net zo vaak neerslag valt als in Schiphol. Ook is te zien dat in de zomermaanden er op een dag met neerslag, er gemiddeld meer millimeter valt, dan in de wintermaanden. Daarnaast valt er iets meer neerslag in de wintermaanden dan in de zomermaanden. Dit is wat verwacht werd. Toch gaat voor de rest van het onderzoek niet met een natte en droge periode worden gewerkt.

De verschillen tussen de maanden zijn voor de splitsing in verschillende periodes te klein. De standaarddeviatie in dagelijkse neerslag is ongeveer 4.5. Dit is groter dan de afwijkingen tussen de gemiddelde gevallen neerslag per dag in de maanden.

	Jan	Feb	Maa	Apr	Mei	Jun
Natte dagen	515	483	457	382	402	425
Droge dagen	415	365	473	518	528	475
Percentage nat	0.55	0.57	0.49	0.42	0.43	0.47
Neerslag	2.25	2.19	1.91	1.42	1.98	2.3
Neerslag nat	4.07	3.84	3.89	3.35	4.58	4.86
	Jul	Aug	Sep	Okt	Nov	Dec
Natte dagen	463	444	436	462	556	540
Droge dagen	467	486	464	468	344	390
Percentage nat	0.5	0.48	0.48	0.5	0.62	0.58
Neerslag	2.9	2.65	2.58	2.57	2.65	2.51
Neerslag nat	5.82	5.55	5.32	5.17	4.29	4.32

Tabel 3.2: Neerslagdagen in de Bilt tussen de jaren 1987-2016. Natte dagen: Aantal dagen met neerslag in de 30 jaar in de maand. Droge dagen: Het aantal neerslagvrije dagen over de 30 jaar in de maand. Percentage nat: Het gemiddelde percentage natte dagen in de maand. Neerslag: Gemiddelde hoeveelheid neerslag per dag in de maand [mm]. Neerslag nat: de gemiddelde hoeveelheid neerslag in de maand per dag als er neerslag valt [mm].

	Jan	Feb	Maa	Apr	Mei	Jun
Natte dagen	528	484	448	388	385	407
Droge dagen	402	364	482	512	545	493
Percentage nat	0.57	0.57	0.48	0.43	0.41	0.45
Neerslag	2.18	1.94	1.73	1.33	1.77	2.15
Neerslag nat	3.83	3.41	3.58	3.09	4.29	4.75
	Jul	Aug	Sep	Okt	Nov	Dec
Natte dagen	453	459	473	501	575	552
Droge dagen	477	471	427	429	325	378
Percentage nat	0.49	0.49	0.53	0.54	0.64	0.59
Neerslag	2.79	3.18	2.66	2.73	2.89	2.47
Neerslag nat	5.72	6.44	5.06	5.07	4.52	4.16

Tabel 3.3: Neerslagdagen in Schiphol tussen de jaren 1987-2016. Natte dagen: Aantal dagen met neerslag in de 30 jaar in de maand. Droge dagen: Het aantal neerslagvrije dagen over de 30 jaar in de maand. Percentage nat: Het gemiddelde percentage natte dagen in de maand. Neerslag: Gemiddelde hoeveelheid neerslag per dag in de maand [mm]. Neerslag nat: de gemiddelde hoeveelheid neerslag in de maand per dag als er neerslag valt [mm].

4 Post-processing analyse met logistische regressie

Nu er blijkt dat er geen duidelijk natte en droogtepatroon in Nederland is, kan er gekeken worden naar het voorspellen van de neerslag. Hierbij wordt alleen gekeken of de hoeveelheid neerslag tenminste 5 mm of minder is.

Aangezien het weersverwachtingsmodel van het ECMWF vaak met een update komt, kan niet alle weersverwachtingsdata worden gebruikt om er een post-processingsmodel van te maken. In de drie jaar waarop ik mijn onderzoek baseer, 1 januari 2014 tot en met 31 december 2016, is één update geweest. Hierom wordt er niet verder in het verleden gekeken. In deze drie jaar was op ongeveer 7.5% procent van de dagen (ongeveer 80 van de 1096 dagen) de neerslag meer dan 5 mm in zowel De Bilt als in Schiphol.

Na het zetten van de drempelwaarde (minstens 5 mm) kan over worden gegaan op het uitvoeren van regressie. Maar, wat is regressie eigenlijk? Regressieanalyse is een methode binnen de statistiek om een waarde Y te schatten met behulp van geobserveerde onafhankelijk variabelen $X_1, X_2, \ldots X_n$, waarbij n een natuurlijk getal is.

De theorie in dit hoofdstuk is gebaseerd op Wilks (2011) [7] met als toevoeging Khuri (2009) [8] voor het deel over gegeneraliseerde lineaire modellen.

4.1 Lineaire regressie

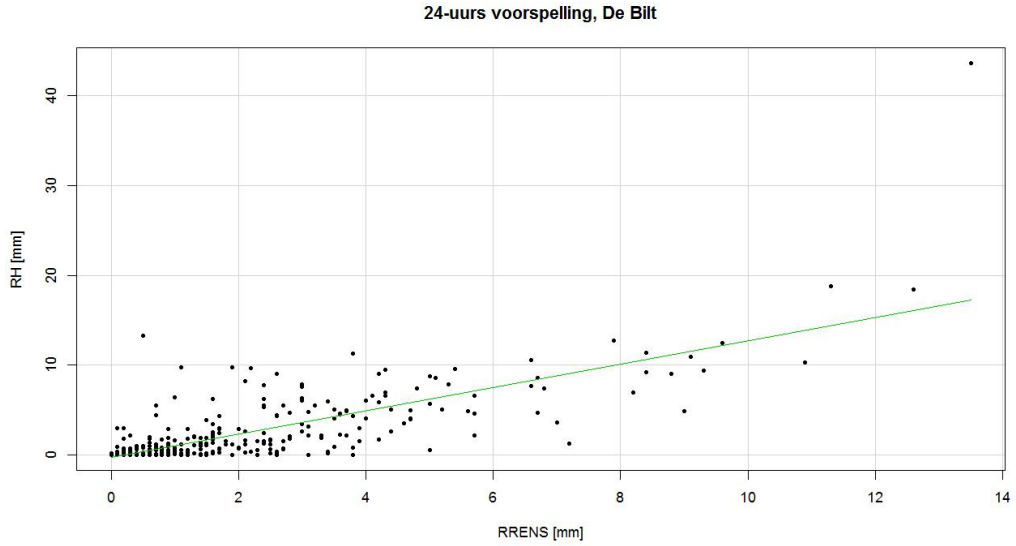
De meest eenvoudige vorm van regressie is lineaire regressie. Een constante verandering in de onafhankelijke variabelen zorgt hierbij voor een constante verandering in de afhankelijke responsvariabele Y. Hiervoor wordt gebruik gemaakt van een lineaire combinatie, vandaar de naam lineaire regressie. Dit is een lineaire combinatie van de parametervector β met bijbehorende onafhankelijke variabelen. Daarnaast is er een beginwaardevector α. De formule voor lineaire regressie wordt als volgt beschreven

$$\mathbf{Y} = \alpha + \mathbf{X}^T\beta. \tag{4.1}$$

De parameters en de beginwaarden worden met behulp van de geobserveerde data geschat. Dit wordt gedaan door gebruik te maken van de least squares methode, het minimaliseren van de som van de kwadraten van de fouten.

Regressie kan in deze vorm worden uitgevoerd als je bijvoorbeeld de hoeveelheid neerslag precies wilt voorspellen. Dit voorbeeld is te zien voor de 24-uursvoorspelling in Figuur 4.1. De echt gevallen hoeveelheid neerslag (RH) wordt voorspeld met de variabele "neerslag gemiddeld over ENS (0.1

mm) (RRENS)". Hierin is de groene lijn de regressielijn. In deze figuur is te zien dat er een duidelijk lineair verband is tussen de hoeveelheid gevallen neerslag en de variabele RRENS.



Figuur 4.1: Scatterplot van de 24 uurs voorspellingsgemiddelde (RRENS) tegen de uiteindelijk hoeveelheid gevallen neerslag op die dag. Beide zijn in millimeters. De groene lijn is de lineaire regressielijn.

4.2 Gegeneraliseerd Lineair Model

Om de kans te bepalen dat er op een bepaalde dag meer dan 5 mm neerslag valt, kan er geen gebruik gemaakt worden van lineaire regressie, maar wel van een uitbreiding hiervan. Bij deze kansbepaling hoort namelijk een discrete willekeurige variabele. De responsvariabele Y is binair, waarbij een succes (1) wordt gegeven als de neerslag op de bepaalde dag tenminste 5 mm was. Dit soort regressie kan worden uitgevoerd met behulp van een Gegeneraliseerd Lineair Model (GLM).

Een GLM bestaat uit drie componenten: [8]

i. een kansverdeling f van de exponentiële familie die wordt gevolgd door de responsvariabele,

ii. een lineaire schatter η, $\eta(X) = \mathbf{X}^T\beta$,

iii. een linkfunctie g, welke het verband tussen de lineaire schatter en het gemiddelde μ beschrijft: $E(Y) = \mu = g^{-1}(\eta)$. De linkfunctie is monotoon differentieerbaar.

Een veelgebruikt GLM waarbij de responsvariabele de waarden 1 en 0 aanneemt is de logistische link functie, de Logit-functie:

$$g(\mu) = \log\left[\frac{\mu}{1-\mu}\right], \tag{4.2}$$

met μ het gemiddelde. Vergelijking 4.2 kan gebruikt worden om de succeskans ($p = \mu$) te beschrijven. Hiervoor wordt gebruik gemaakt van de lineaire predictor op het punt $x = \begin{pmatrix} x_1 & x_2 & \dots & x_k \end{pmatrix}^T$. De succeskans ziet er dan als volgt uit

$$p(x) = \frac{\exp\left[\mathbf{X}^T\beta\right]}{1 + \exp\left[\mathbf{X}^T\beta\right]}. \tag{4.3}$$

4.3 Toepassen van het GLM

In de situatie die bij dit onderzoek beschouwd wordt, zijn meerdere observaties aanwezig en wordt er gebruik gemaakt van meerdere variabelen. Daarnaast zijn de waarden die de responsvariabele Y aanneemt alleen 1 of 0. In het geval van De Bilt:

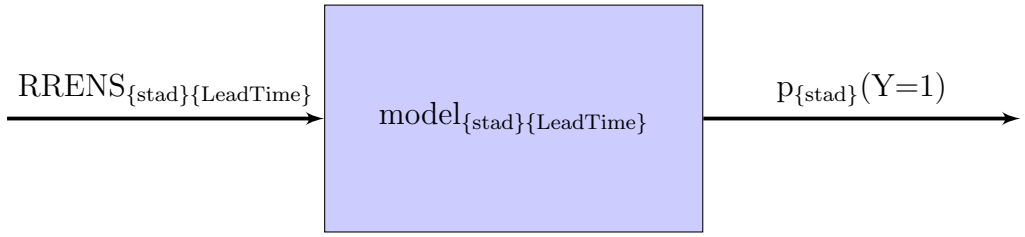
$$Y = \begin{cases} 1 & \text{RH}\left\{\text{stad}\right\} \geq 5mm \\ 0 & \text{anders} \end{cases} \tag{4.4}$$

Deze zijn Bernoulli-verdeeld met succeskans p. Er wordt gebruik gemaakt van meerdere observaties. Hierdoor kan het aantal successen worden geteld en zal de geschatte Y-waarde liggen tussen 0 en 1. Er wordt daarom gewerkt met een binomiaalverdeling, welke een uitbreiding is op de Bernoulli-verdeling. In het geval van de binomiaalverdeling, is de kansmassafunctie f van de vorm

$$f(y,p) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y = 1, 2, \dots n.$$

De binomiaalverdeling behoort tot de eerder besproken exponentiële familie (zie Subsectie 4.2).

Voor het uitvoeren van de regressie is het belangrijk om te weten welke onafhankelijke variabelen significant zijn. Alleen deze variabelen worden uiteindelijk bij de regressie gebruikt. Om te checken of een variabele significant is, wordt er uitgegaan van de nulhypothese *de coëfficiënt (β) is gelijk aan nul* tegen de alternatieve hypothese *de coëfficiënt is ongelijk aan nul*. De nulhypothese wordt verworpen wanneer de p-waarde kleiner is 0.1. Hiervoor is gekozen omdat regressie geen hele nauwkeurige parameters nodig heeft,



Figuur 4.2: Blokschema één-stad-model. Hierin kan stad viariëren tussen De Bilt en Schiphol en LeadTime tussen +12, +24, +36, +48 uur.

om toch een significant resultaat te behalen. Daarbij is dit onderzoek een check om te kijken naar de mogelijkheden van deze postprocessingsmethode na de numerieke berekening.

De significantie van de variabele kan allereerst worden bepaald door te kijken naar de onderlinge afhankelijkheid. Dit wordt gedaan middels het opstellen van een correlatiematrix. De correlatiematrix van de voorspelling 12 uur vooruit voor specifiek De Bilt is te zien in Tabel 4.1. De overige matrices zijn te vinden in Bijlage B en C. In deze matrices is te zien dat de correlatie tussen geen enkele variabele ongeveer gelijk is aan nul. Dit betekent dat alle variabelen gecorreleerd zijn en daarmee afhankelijk zijn van elkaar. Dit is ook te verwachten aangezien alle variabelenwaarden steeds uit dezelfde vijftig weersvoorspellingen worden gehaald en voor hetzelfde moment de hoeveelheid neerslag voorspellen.

	RRENS	RRSDEV	RR6_46	RRGT00	RRGE03	RRGE15	RRGE50
RRENS	1.00	0.81	0.83	0.46	0.63	0.86	0.92
RRSDEV	0.81	1.00	0.98	0.49	0.63	0.75	0.69
RR6_46	0.83	0.98	1.00	0.48	0.63	0.76	0.71
RRGT00	0.46	0.49	0.48	1.00	0.84	0.52	0.29
RRGE03	0.63	0.63	0.63	0.84	1.00	0.73	0.41
RRGE15	0.86	0.75	0.76	0.52	0.73	1.00	0.69
RRGE50	0.92	0.69	0.71	0.29	0.41	0.69	1.00

Tabel 4.1: Correlatiematrix van de 12-uursvoorspelling, De Bilt.

Er wordt besloten om variabelen met een correlatiequotiënt van hoogstens 0.4 te nemen. Het is zo dat RRENS de meeste correlatie heeft met RH , wat belangrijk is voor de voorspelling. Bij geen enkel berekeningsmoment is de correlatie tussen RRENS en een andere variabele lager dan 0.4. Voor een enkel voorspellingsmoment wordt bij het uitvoeren van regressie daarom alleen het gemiddelde van de 51 berekende voorspellingen gebruikt (RRENS).

Er is gekozen om de verschillende lead times niet te combineren. Dit omdat de correlatie tussen de bij elkaar horende lead times groter dan 0.4 is. Daarnaast neemt in het algemeen de fout in de predictie toe naarmate de

lead times toenemen. Een blokschema van hoe een gekozen model werkt is te zien in Figuur 4.2.

		00-12 UTC		12-00 UTC	
		24 uur	48 uur	12 uur	36 uur
p-waarde	Intercept	<2e-16	<2e-16	<2e-16	<2e-16
	RRENS	<2e-16	<2e-16	<2e-16	<2e-16
AIC		214.95	257.14	212.15	262.03

Tabel 4.2: Goodness of Fit tabel De Bilt.

Voor het controleren van de hypotheses en de Goodness of Fit, zijn de samenvattingen van de modellen opgevraagd. Hierin is per variabele de p-waarde gegeven. Deze zijn berekend met de Wald Inference methode. Verder wordt er ook de Akaike Informatie Criterium (AIC) gegeven. Dit criterium geeft de Goodness of Fit van een model weer en wordt bepaald volgens

$$AIC_j = -2 \log L(\hat{\theta}_j) + 2s_j.$$

Hierbij wordt het natuurlijke logaritme van de fout genomen en daar wordt twee keer het aantal parameters van het regressiemodel aan toegevoegd. In het algemeen is het model met de laagste AIC-waarde de best fit.

De resultaten van de opgevraagde samenvatting van de modellen voor De Bilt zijn te zien in Tabel 4.2. De samenvatting van Schiphol staat, net als die van De Bilt, in Bijlage D. Een model voor 00-12 UTC en een model voor 12-00 UTC wordt gekozen, aangezien daarmee een hele dag kan worden voorspeld. De p-waardes hebben vanwege hun lage waardes geen invloed op de keuze van een model. Daarom worden de modellen gekozen op hun AIC-waarde, welke wel een duidelijk onderling verschil hebben. De modellen voor 12- en 24-uursverwachtingen in zowel De Bilt als Schiphol komen hierbij het beste naar voren en worden daarom gekozen om mee verder te gaan. Deze modellen hebben de parameters gegeven in Tabel 4.3.

		De Bilt		Schiphol	
		12 uur	24 uur	12 uur	24 uur
$\hat{\beta}$	intercept	-4.69003	-4.44153	-4.46008	5.08622
	RRENS	0.76659	0.95541	0.74283	0.89133

Tabel 4.3: Parameters van de gekozen modellen.

Na het uitvoeren van de regressie, kan over worden gegaan op de bepaling van de odds ratio, of succeskans. Dit geeft de verhouding weer tussen de

waarschijnlijkheid waarop de gebeurtenis zal optreden en de waarschijnlijkheid dat deze niet zal optreden. De ratio wordt uitgevoerd met de formule gegeven in Vergelijking 4.3.

Hierna kunnen verificatiemethodes op de modellen worden uitgevoerd. Dit wordt gedaan in Subsectie 4.5 en 4.6.

4.4 Twee-Stedenmodel

Naast het uitvoeren van regressie in één stad, De Bilt of Schiphol, is het ook interessant om dit te combineren. Kan er een regressiemodel worden gemaakt dat voorspelt of er in zowel De Bilt als in Schiphol meer dan 5 mm valt?

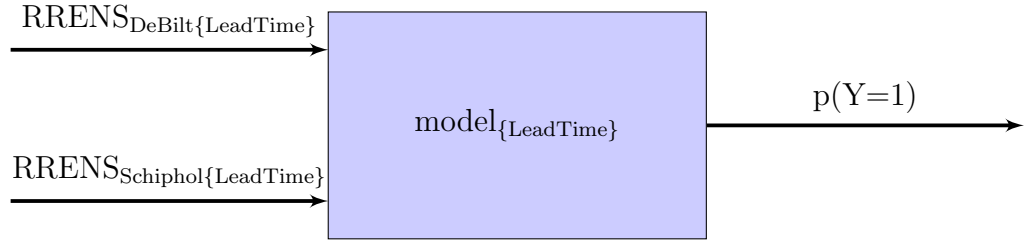
$$Y = \begin{cases} 1 & \text{RH Schiphol} \geq 5\ mm\ \&\ \text{RH De Bilt} \geq 5\ mm, \\ 0 & anders \end{cases} \tag{4.5}$$

En een regressiemodel dat voorspelt of er in De Bilt en/of in Schiphol minstens 5 mm neerslag valt?

$$Y = \begin{cases} 1 & \text{RH Schiphol} \geq 5\ mm\ \text{of RH De Bilt} \geq 5\ mm, \\ 0 & anders \end{cases} \tag{4.6}$$

Dit resulteert dan in een model van de vorm te zien in het blokschema in Figuur 4.3.

In de regressiedata komt het in meer dan vijf procent van de gevallen voor dat er in zowel De Bilt als in Schiphol meer dan 5 mm neerslag valt. Hierdoor kan regressie worden uitgevoerd. Echter, de correlatie tussen RRENS van Schiphol en De Bilt is voor alle berekeningsmomenten groter dan 0.93. Ze zijn dus zeer gecorreleerd. Regressie wordt normaalgesproken niet uitgevoerd wanneer de onafhankelijke variabele een onderlinge hoge correlatie hebben. Het effect van een van de parameters zal waarschijnlijk vrij klein zijn. Toch wordt dit geprobeerd.



Figuur 4.3: Blokschema twee-steden-model. Hierin kan LeadTime tussen +12, +24, +36, +48 uur.

Na het uitvoeren van regressie voor het eerste genoemde model blijkt steeds één van de coëfficiënten niet significant te zijn (p-waarde > 0.1). Zie hiervoor ook Tabel 4.4. Hierom wordt er gekozen deze modellen niet te gaan verifiëren.

		00-12		12-00	
		24 uur	48 uur	12 uur	36 uur
p-waarde	Intercept	<2e-16	<2e-16	<2e-16	<2e-16
	RRENS DB	0.14336	0.3240	0.000189	0.4400
	RRENS S	0.00194	0.0132	0.378593	0.0213
AIC		161.53	122.95	205.76	147.34

Tabel 4.4: Goodness of Fit tabel De Bilt samen met Schiphol.

Ook is een regressiemodel opgesteld waarbij bij minstens één van beide steden tenminste 5 mm neerslag valt. Ook hier is weinig significantie te vinden bij de coëfficiënten van het model. (Zie ook Bijlage D). Er wordt gekozen om niet met deze modellen verder te werken. Er valt wel op dat het 24-uursmodel van het laatste geval wel significante coëfficiënten heeft. Voor dit model wordt geen uitzondering gemaakt om wel mee verder te werken. Op de oorzaak van de lage significatie wordt verder ingegaan in de discussie (Sectie 6).

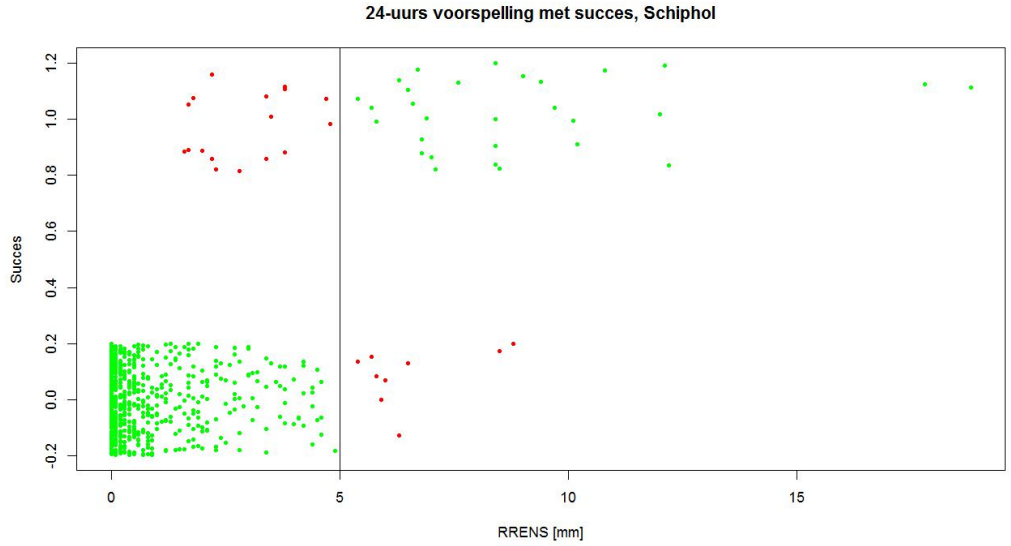
4.5 Verificatie van de modellen

In de vorige secties is de regressie uitgevoerd en vervolgens is er gekeken naar hoe goed de gemaakte modellen zijn. Uit de Goodness of Fit tabellen, die terug te zien zijn in Bijlage D, is gekozen voor het 12-uurs- en 24-uursmodel van zowel De Bilt als die van Schiphol om mee verder te werken. Daarnaast is er geen model gekozen waarbij de twee steden worden gecombineerd (Subsectie 4.4).

In deze sectie wordt gekeken naar hoe goed een model is, ten opzichte van het systeem dat het KNMI gebruikt. Bij dit onderzoek worden de regressieresultaten geverifieerd met behulp van het reliability diagram en de Brier-Score. Deze worden vervolgens vergeleken met de verificatie van het ECMWF-model. Daarvan in het bijzonder de waarden van % leden met minstens 5.0 mm neerslag (RRGE50). Deze waarden zeggen namelijk hoeveel procent van de 51 leden er meer dan 5 mm voorspelt voor die bepaalde validatiedatum. De uitkomst van de kansbepaling na regressie geeft ook weer hoeveel procent kans er is op meer dan 5 mm neerslag voor die validatiedatum. Er wordt dan ook verwacht dat deze waarden ongeveer gelijk

zijn. Voor extra informatie over de gebruikte verificatiemethoden verwijs ik u graag naar Hoofdstuk 8 van Wilks, 2011 [7].

Voordat er in wordt gegaan op de verificatie, wordt er gekeken naar het fout-positief en fout-negatief resultaat van de modellen. Hiervoor moet eerst de kans op de gebeurtenis (minstens 5 mm in een dagdeel) worden bepaald. Dit wordt berkend volgens Vergelijking 4.3. Wanneer de kans groter is dan 0.5 wordt deze gezet op 1 (er zal tenminste 5 mm neerslag vallen) en anders op 0 (er zal minder dan 5 mm vallen). In Figuur 4.4 is een scatterplot weergegeven van het 24-uursmodel in Schiphol, waarbij de groene punten correcte voorspellingen zijn en rode incorrecte. Hierin valt op dat er meer fout-negatieven (weinig neerslag voorspelt, maar veel gevallen) dan fout-positieven zijn voorspeld.



Figuur 4.4: Hierin is Succes uitgezet tegen het voorspelde gemiddelde 24 uur van te voren. De rode punten geven ofwel een succes aan terwijl er geen succes is voorspeld of geen succes terwijl er wel succes was voorspeld. De groene punten zijn de dagen waarbij de voorspelling goed was. Succes houdt in meer dan 5 mm neerslag op een dag.
Op de Succes-as is een jitter uitgevoerd, waardoor duidelijker te zien is dat er veel punten rond (0,0) liggen. Hierdoor liggen nu de punten die eigenlijk op Succes=0 liggen tussen -0.2 en 0.2 en de punten op Succes=1 tussen 0.8 en 1.2.

Dit resultaat is ook terug te vinden in Tabel 4.5. Hierin zijn ook de resultaten van De Bilt weergegeven en van de 12-uursmodellen. Daarnaast

wordt er in de tabel vergeleken met het Numerical Weather Prediction model (ECMWF). Hierbij wordt een succes gegeven als de kans die er wordt gegeven op meer dan 5 mm neerslag strikt groter is dan 0.5 bij RRGE50.

Daarnaast is te zien dat het regressiemodel een groter foutpercentage heeft dan het ECMWF -systeem. De verwachting is daarom ook dat het regressiemodel minder goed de neerslag van minstens 5 mm zal voorspellen.

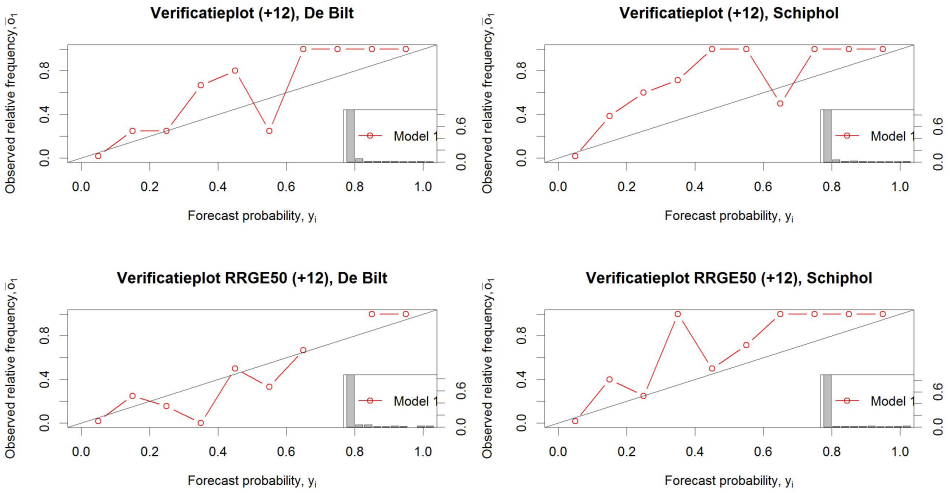
	Post-Processing				NWP			
	De Bilt		Schiphol		De Bilt		Schiphol	
	12 uur	24 uur	12 uur	24 uur	12 uur	24 uur	12 uur	24 uur
Fout-positief	0.8	1.6	0.3	0.3	1.4	0.8	0.5	0.5
Fout-negatief	4.9	5.2	6.6	5.8	3.3	5.2	4.1	5.2

Tabel 4.5

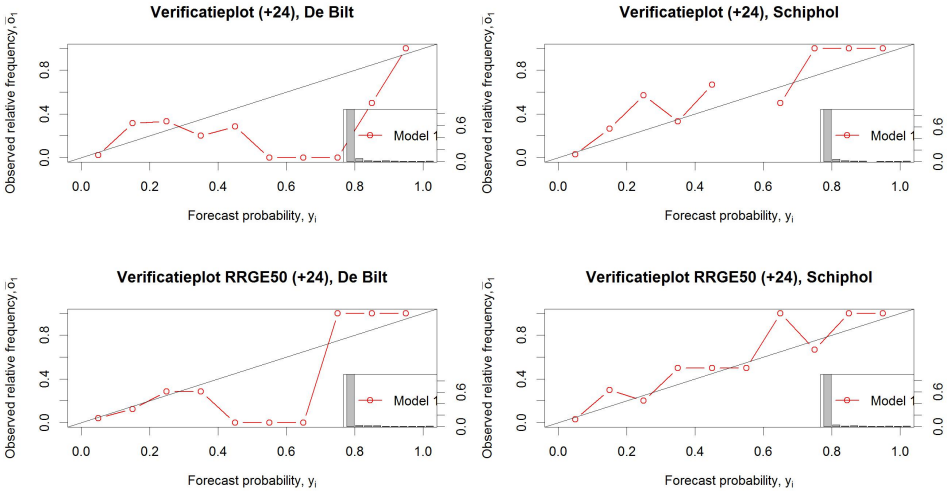
Allereerst wordt er een verificatie uitgevoerd met het reliability diagram. Hierin wordt grafisch de waarschijnlijkheid uitgezet tegen de geobserveerde frequentie. Het regressiemodel geeft na het uitvoeren van de verwachting voor de 12-uursvoorspelling de waarden gegeven in de Pred12-kolom van Tabel 4.6. De waarden van de tabel worden gegroepeerd naar Pred12 en uitgezet op de X-as. Op de verticale as staat de geobserveerde frequentie, waarvoor het percentage dat daadwerkelijk meer dan 5 mm is gevallen per groep wordt bepaald. Als het model perfect is, zullen alle punten op de de diagonaal liggen.

	YYYYMMDD	RH	Succes	Pred12
1	20160101	0.00	0	0.01
2	20160102	1.00	0	0.05
3	20160103	5.30	1	0.74
4	20160104	6.00	1	0.34
5	20160105	1.60	0	0.03
6	20160106	0.00	0	0.02
7	20160107	12.80	1	0.87
8	20160108	0.00	0	0.01
9	20160109	0.00	0	0.03
10	20160110	0.00	0	0.01
		⋮		

Tabel 4.6: De eerste tien rijen van de door het 12-uursmodel van Schiphol voorspelde data na uitvoeren van regressie en berekenen van de verwachting.



Figuur 4.5: Reliability diagrammen van de 12-uursverwachtingen in De Bilt (links) en Schiphol (rechts). De bovenste is van het regressiemodel en de onderste van de berekeningen van het systeem dat het KNMI gebruikt (RRGE50).



Figuur 4.6: Reliability diagrammen van de 24-uursverwachtingen in De Bilt (links) en Schiphol (rechts). De bovenste is van het regressiemodel en de onderste van de berekeningen van het systeem dat het KNMI gebruikt (RRGE50).

De reliability diagrammen van de 24-uursverwachting van de regressiemodellen (Figuur 4.6) lijken sterk op de diagrammen van RRGE50. Er is geen eenduidigheid over welk model meer rond de diagonaal liggen. De RRGE50 punten lijken echter wel dichter rond de lijn te liggen, vooral bij De Bilt. Dit resultaat, dat de diagrammen sterk op elkaar lijken, maar dat die vanRRGE50 net wat beter eruit ziet, wordt verwacht terug te zien in de Brier-Scores van de modellen.

Naast het reliability diagram is een veel gebruikte methode voor verificatie de Brier-Score. De score is de Mean Squared Error van de kansverwachtingen en daarmee een maat om kansen op een gebeurtenis te verifiëren. Het is het gemiddelde van de kwadratische verschillen tussen de verwachtingskans (p_f) en de binaire observatie (O),

$$BS = \frac{1}{N}\sum_{1}^{N}(p_f - O)^2. \tag{4.7}$$

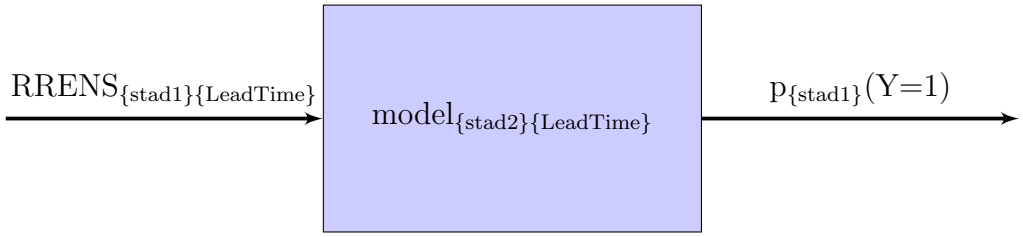
Hierbij geldt $O = 1$ als er minstens 5 mm is gevallen en $O = 0$ als er minder is gevallen dan 5 mm.

De Brier-Score is negatief georiënteerd. Dat betekent dat hoe lager de score hoe beter de voorspelling. Aangezien de verwachtingen en de observaties begrensd zijn bij nul en een, ligt de score ook tussen deze twee waarden. Hiermee is nul dus een perfecte score en hoort een score van 1 bij een niet accurate voorspelling.

De gevonden Brier-Scores horend bij de gemaakte regressiemodellen zijn weergegeven in Tabel 4.7. De Brier-Score is voor elk moment groter bij het regressiemodel, dan bij het RRGE50-model. Net als bij de reliability diagrammen wordt hier de conclusie getrokken dat voor het voorspellen van meer dan 5 mm neerslag er beter gekeken kan worden naar het model dat het KNMI gebruikt (ECMWF).

		12 uur	24 uur
De Bilt	Regressie	0.03987	0.05123
	RRGE50	0.03640	0.05056
Schiphol	Regressie	0.04384	0.04634
	RRGE50	0.03579	0.04626

Tabel 4.7: Brier-Score van de regressiemodellen en het ECMWF.



Figuur 4.7: Blokschema gegeneraliseerd één-stad-model. Hierin kan stad1 en stad2 viariëren tussen De Bilt en Schiphol en LeadTime tussen +12, +24, +36, +48 uur.

4.6 Generaliseerbaarheid van het één-stad-model

Aangezien de correlatie tussen Schiphol en De Bilt vrij hoog is qua hoeveelheid gevallen neerslag, 0.82 (RH), en dat de correlatie tussen de voorspellingsgemiddelden (RRENS) nog hoger is, meer dan 0.9, is er gekozen om te kijken of de één-stad-modellen ook de neerslag in de andere stad goed kunnen voorspellen. Het blokschema voor deze generalisatie is te zien in Figuur 4.7. Hiervoor zijn de modellen voor de 12-uursverwachting en 24-uursverwachting van De Bilt en Schiphol gebruikt, deze zijn opgesteld in Subsectie 4.3.

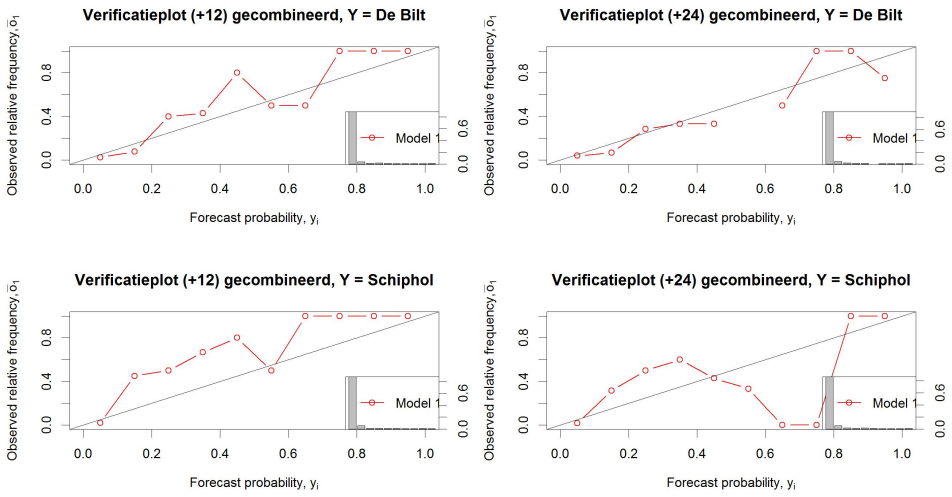
	De Bilt		Schiphol	
	12 uur	24 uur	12 uur	24 uur
Fout-positief	0.3	0.5	0.5	0.3
Fout-Negatief	5.2	4.9	6.6	6.0

Tabel 4.8: Fout negatieven en positieven voor de één-stad-modellen waarbij de andere stad als in- en output wordt gebruikt. De bovenste rij staat voor welke er wordt voorspeld.

Voor het voorspellen van de succeskans is ook hier gebruik gemaakt van Vergelijking 4.3. Vervolgens zijn de fout-positief en fout-negatief waarden hiervan genoteerd in Tabel 4.8.

Wanneer dit wordt vergeleken met Tabel 4.5, dan valt er op dat de voorspellingen voor De Bilt beter gemaakt kunnen worden door het model van Schiphol dan andersom. Zeker voor het 24-uursmodel. Verder zit er weinig positief verschil tussen. Je hebt tevens liever dat de fout-negatief score veel lager wordt en daarmee de fout-positief score omhoog gaat dan andersom.

Ook is dit te zien in de reliability diagrammen, weergegeven in Figuur 4.8. Vergeleken met de diagrammen in Figuur 4.5 en 4.6 ligt de reliability fit meer om de rechte lijn heen en dichter bij die lijn bij Figuur 4.8.



Figuur 4.8: Reliability diagrammen. Links die van de 12-uursmodellen en links van de 24-uursmodellen. Bovenste rij is met responsvariabele De Bilt en onafhankelijke variabele Schiphol. Onderste rij is met responsvariabele Schiphol en onafhankelijk variabele De Bilt.

	Responsvariabele	Brier-Score
Model12	De Bilt	0.0477
	Schiphol	0.0487
Model24	De Bilt	0.0394
	Schiphol	0.0498

Tabel 4.9: Brier-Score waarbij de één-stad-modellen zijn gebruikt met de in- en output van de andere stad.

De gevonden Brier-Scores zijn vrijwel gelijk aan de scores gevonden in Subsectie 4.5 voor de enkele modellen. Wel zijn de scores van Model12 met Schiphol als responsvariabele en De Bilt als onafhankelijke variabele en Model24 met De Bilt als responsvariabale en Schiphol als onafhankelijke variabele lager en daarmee beter dan hun eigenstad counterpart. Daarnaast is de laatste van de twee genoemden ook nog eens beter dan de Brier-Score verkregen van RRGE50.

Uit alle drie deze methoden volgt dat de 24-uursmodel met De Bilt als respons variabele en RRENS van Schiphol als onafhankelijke variabele beter werkt dan het huidige model en het regressiemodel waarbij RRENS van De Bilt ook als onafhankelijke variabele wordt meegenomen. Voor de overige

modellen geldt dit niet, maar wel dat ze vergelijkbare prestaties hebben. De modellen zijn dus generiek toepasbaar.

5 Conclusie

Het KNMI gebruikt een Europees systeem om een weersverwachting voor de langere termijn (tot 15 dagen vooruit) te maken, het ensemblesysteem (ENS) van het ECMWF . Hiervan zijn de gegevens voor de neerslag gebruikt als onafhankelijke variabelen bij de regressie. Ook historische data is gebruikt bij de regressie, als responsvariabelen. Uit de analyse van de historische data is gebleken dat Nederland geen duidelijk nat en droog seizoen heeft. Er is daarom hier ook niet mee verder gewerkt.

Er is een gegeneraliseerde vorm van lineaire regressie uitgevoerd. Hierbij is gebruik gemaakt van de logit functie en daarbij de binomiaalverdeling. Op de vraag of dit regressiemodel een goede toevoeging aan het ECMWF -model is, kan het antwoord kort zijn: nee. De Brier-Scores van de regressiemodellen gevonden in Subsectie 4.5 en Subsectie 4.6 zijn hoger dan de scores van het ECMWF -systeem. Daarnaast lijken de reliability diagrammen erg veel op elkaar en zijn de fout-positief en fout-negatief percentages ook niet beter.

Daarnaast is gebleken dat een één-stad-model kan worden gegeneraliseerd. Subsectie 4.6). Maar één model opgesteld uit de gegevens van De Bilt en Schiphol samen voor het voorspellen van minstens 5 mm in beide steden of minstens 5 mm in tenminste een van de steden kan niet worden gemaakt, want de variabelen zijn teveel gecorreleerd.

Al om al wordt het ECMWF -modeluitvoer niet verbeterd met deze regressiemethode.

6 Discussie

In deze scriptie is gekeken naar de gevallen neerslag van minstens 5 mm. Neerslag valt in verschillende vormen en elke vorm heeft zijn eigen impact. Zo heeft men minder last van regen dan van sneeuw. Met de verschillende vormen neerslag is geen rekening gehouden in deze scriptie.

De datasets die zijn gebruikt, zijn afkomstig van drie jaar. Deze zijn gesplitst omdat er een test verzameling moet zijn, om het model te kunnen testen. Hierdoor gaat er enige data verloren om het model beter te kunnen maken. Echter, het systeem van ECMWF wordt regelmatig ge-update en daarmee steeds meer geoptimaliseerd. Het kan zijn dat hierdoor de keuze van afsplitsing tussen de trainingsdataset en de verificatiedataset niet handig is gekozen. In een volgend onderzoek is het slim om hier rekening mee te houden.

Daarnaast zou het kunnen zijn dat als de dataset groter was geweest, hiermee wel een regressiemodel gemaakt zou kunnen worden gemaakt dat

beter is dan het ECMWF-model. Een grotere dataset is echter niet mogelijk door de vele updates van het ECMWF-model. Daarentegen is het misschien wel mogelijk om, door de generaliseerbaarheid van het model, de data van meerdere steden te gebruiken om de dataset te vergroten. Dit zou een mogelijkheid zijn voor vervolgonderzoek.

Tijdens dit onderzoek is er gekeken naar één soort regressie, namelijk het GLM en daarbinnen alleen naar de logit-functie. Naast deze linkfunctie is er ook de probit-functie. Deze werkt met de normaalverdeling. Aangezien deze lastiger te interpreteren en te berekenen is, is er voor deze scriptie voor de logit-functie gekozen. Er zou wel onderzoek kunnen worden gedaan naar de effectiviteit van probit in combinatie met de onderzoeksvraag behandeld in deze scriptie. Echter, in het algemeen liggen de resultaten van deze twee linkfuncties erg dicht op elkaar. Ook zou er kunnen worden gekeken naar een geheel andere regressiemethode, in de vorm van niet-parametrische regressie.

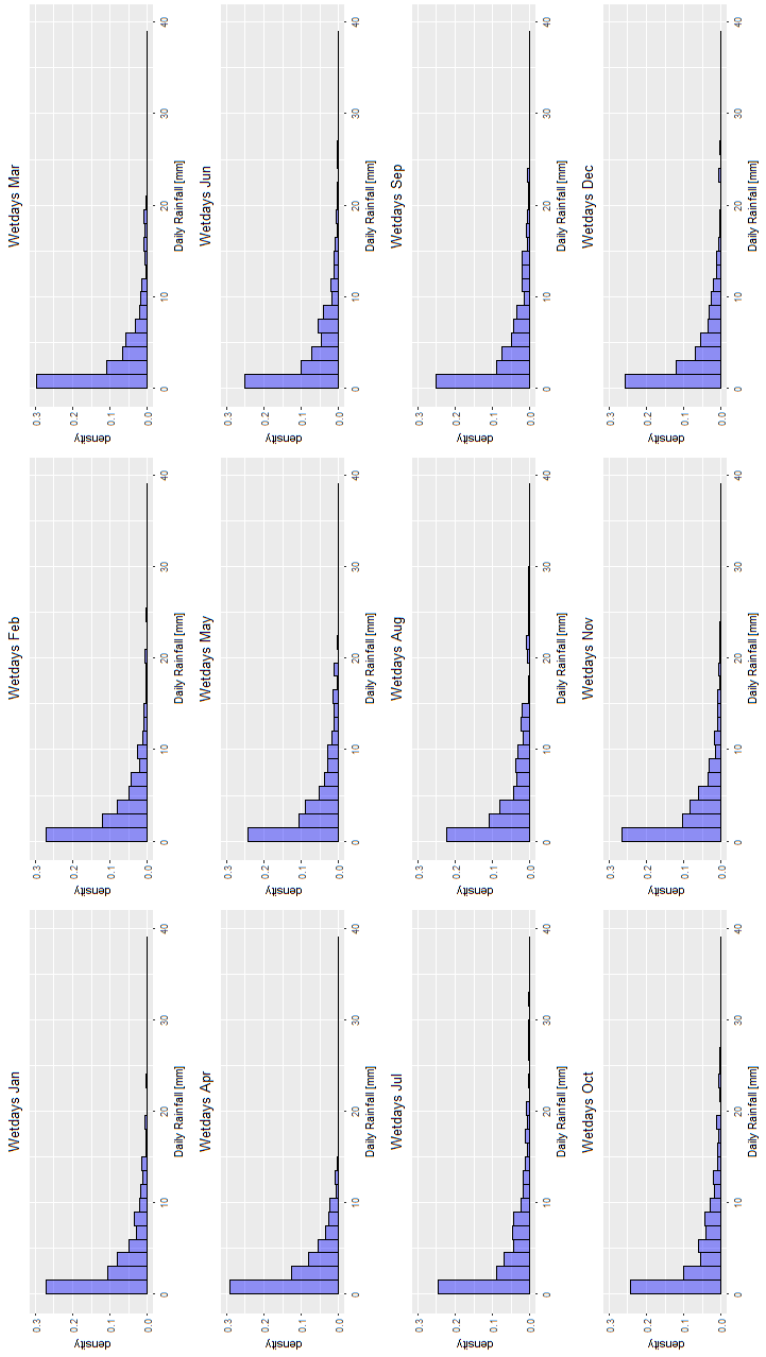
Er kan naast een andere regressiemethode ook gekeken worden naar regressie met de kolom RRGE50 in plaats van RRENS. De correlatie tussen de responsvariabelen en een van de hiervoor genoemde onafhankelijke variabelen verschilt gemiddeld 0.02. Daarnaast staat er in de kolom RRGE50 ook hoeveel procent van de ensembleleden er meer dan 5 mm neerslag verwachten. Dit is wat er geprobeerd wordt te optimaliseren in dit onderzoek door een postprocessingsmethode toe te passen. Echter, de vergelijking met de verificatie van deze variabele is dan minder bruikbaar.

Uiteindelijk is er geen regressiemodel gebruikt waarbij beide steden worden gecombineerd. Er viel op dat bij het ene model de RRENS van Schiphol voornamelijk werd gebruikt om het model samen te stellen en bij een ander model juist de RRENS van De Bilt. Het is goed mogelijk dat als hetzelfde onderzoek op een andere dataset zou zijn uitgevoerd, dit effect zal hebben op de samenstelling van de modellen en dat dan juist de RRENS van De Bilt op de voorgrond speelt in plaats van daar waar Schiphol bij een model in deze scriptie de hoofdrol speelde. Dit kan naar mijn idee verklaard worden door AIC-waarden van de verschillende modellen van de steden apart en de hoge correlatie tussen Schiphol en De Bilt. Zo is bijvoorbeeld te zien dat van de 12-uursregressiemodellen de AIC-waarden van het model van De Bilt lager is dan die van Schiphol. Vervolgens is het model, dat voorspelt of er in beide steden minstens 5 mm neerslag valt, voornamelijk gebaseerd op de RRENS van De Bilt. (Zie hiervoor Bijlage D). De waarschijnlijke oorzaak van dit fenomeen is de grote correlatie tussen de gekozen inputs.

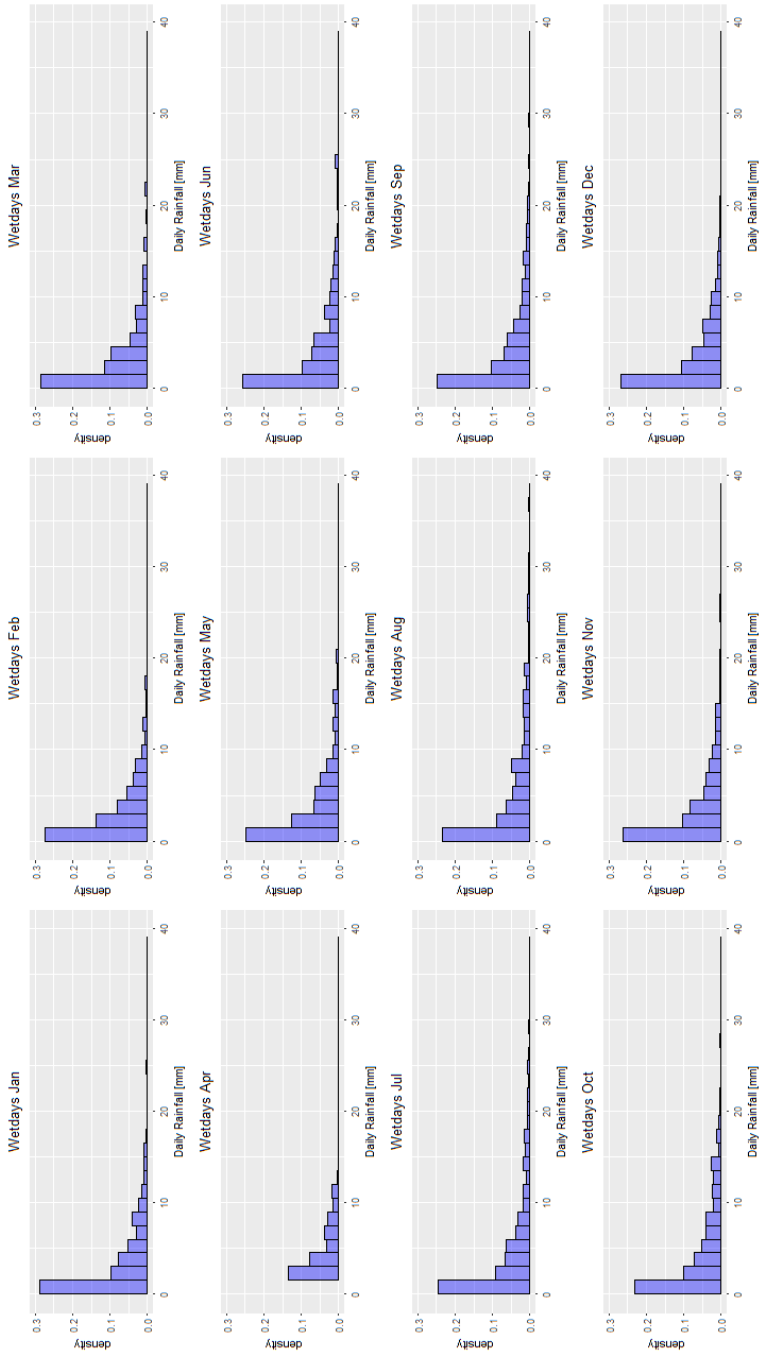
Bijlagen

A Histogram van de natte dagen

A.1 Histogram van de natte dagen in De Bilt tussen 1987-2016.



A.2 Histogram van de natte dagen in Schiphol tussen 1987-2016.



B Correlatiematrices, De Bilt

B.1 Voorspellingsvariabelen tegen Responsvariabele (Succes)

	12-uur	24-uur	36-uur	48-uur
RRENS	0.66	0.65	0.59	0.58
RRSDEV	0.49	0.48	0.46	0.45
RR6_46	0.50	0.51	0.48	0.47
RRGT00	0.21	0.25	0.26	0.27
RRGE03	0.31	0.36	0.35	0.38
RRGE15	0.49	0.58	0.51	0.56
RRGE50	0.68	0.64	0.61	0.55

B.2 Voorspelling12

	RRENS	RRSDEV	RR6_46	RRGT00	RRGE03	RRGE15	RRGE50
RRENS	1.00	0.81	0.83	0.46	0.63	0.86	0.92
RRSDEV	0.81	1.00	0.98	0.49	0.63	0.75	0.69
RR6_46	0.83	0.98	1.00	0.48	0.63	0.76	0.71
RRGT00	0.46	0.49	0.48	1.00	0.84	0.52	0.29
RRGE03	0.63	0.63	0.63	0.84	1.00	0.73	0.41
RRGE15	0.86	0.75	0.76	0.52	0.73	1.00	0.69
RRGE50	0.92	0.69	0.71	0.29	0.41	0.69	1.00

B.3 Voorspelling24

	RRENS	RRSDEV	RR6_46	RRGT00	RRGE03	RRGE15	RRGE50
RRENS	1.00	0.81	0.86	0.47	0.63	0.88	0.92
RRSDEV	0.81	1.00	0.98	0.48	0.59	0.73	0.67
RR6_46	0.86	0.98	1.00	0.48	0.61	0.77	0.72
RRGT00	0.47	0.48	0.48	1.00	0.86	0.51	0.28
RRGE03	0.63	0.59	0.61	0.86	1.00	0.71	0.39
RRGE15	0.88	0.73	0.77	0.51	0.71	1.00	0.70
RRGE50	0.92	0.67	0.72	0.28	0.39	0.70	1.00

B.4 Voorspelling36

	RRENS	RRSDEV	RR6_46	RRGT00	RRGE03	RRGE15	RRGE50
RRENS	1.00	0.84	0.88	0.53	0.69	0.88	0.93
RRSDEV	0.84	1.00	0.98	0.55	0.67	0.76	0.72
RR6_46	0.88	0.98	1.00	0.57	0.70	0.80	0.77
RRGT00	0.53	0.55	0.57	1.00	0.89	0.61	0.35
RRGE03	0.69	0.67	0.70	0.89	1.00	0.80	0.48
RRGE15	0.88	0.76	0.80	0.61	0.80	1.00	0.73
RRGE50	0.93	0.72	0.77	0.35	0.48	0.73	1.00

B.5 Voorspelling48

	RRENS	RRSDEV	RR6_46	RRGT00	RRGE03	RRGE15	RRGE50
RRENS	1.00	0.84	0.89	0.51	0.66	0.88	0.93
RRSDEV	0.84	1.00	0.96	0.52	0.63	0.74	0.73
RR6_46	0.89	0.96	1.00	0.53	0.66	0.79	0.77
RRGT00	0.51	0.52	0.53	1.00	0.89	0.57	0.31
RRGE03	0.66	0.63	0.66	0.89	1.00	0.76	0.43
RRGE15	0.88	0.74	0.79	0.57	0.76	1.00	0.72
RRGE50	0.93	0.73	0.77	0.31	0.43	0.72	1.00

C Correlatiematrices, Schiphol

C.1 Voorspellingsvariabelen tegen Responsvariabele (Succes)

	12-uur	24-uur	36-uur	48-uur
RRENS	0.66	0.68	0.59	0.65
RRSDEV	0.44	0.53	0.45	0.53
RR6_46	0.47	0.56	0.47	0.55
RRGT00	0.22	0.22	0.27	0.23
RRGE03	0.32	0.31	0.37	0.32
RRGE15	0.54	0.52	0.53	0.53
RRGE50	0.69	0.70	0.61	0.63

C.2 Voorspelling12

	RRENS	RRSDEV	RR6_46	RRGT00	RRGE03	RRGE15	RRGE50
RRENS	1.00	0.79	0.82	0.43	0.59	0.84	0.91
RRSDEV	0.79	1.00	0.98	0.45	0.57	0.71	0.66
RR6_46	0.82	0.98	1.00	0.46	0.59	0.74	0.69
RRGT00	0.43	0.45	0.46	1.00	0.84	0.50	0.26
RRGE03	0.59	0.57	0.59	0.84	1.00	0.70	0.37
RRGE15	0.84	0.71	0.74	0.50	0.70	1.00	0.66
RRGE50	0.91	0.66	0.69	0.26	0.37	0.66	1.00

C.3 Voorspelling24

	RRENS	RRSDEV	RR6_46	RRGT00	RRGE03	RRGE15	RRGE50
RRENS	1.00	0.84	0.87	0.44	0.59	0.83	0.91
RRSDEV	0.84	1.00	0.98	0.44	0.56	0.71	0.70
RR6_46	0.87	0.98	1.00	0.45	0.59	0.75	0.74
RRGT00	0.44	0.44	0.45	1.00	0.85	0.51	0.27
RRGE03	0.59	0.56	0.59	0.85	1.00	0.71	0.38
RRGE15	0.83	0.71	0.75	0.51	0.71	1.00	0.67
RRGE50	0.91	0.70	0.74	0.27	0.38	0.67	1.00

C.4 Voorspelling36

	RRENS	RRSDEV	RR6_46	RRGT00	RRGE03	RRGE15	RRGE50
RRENS	1.00	0.84	0.88	0.50	0.66	0.87	0.93
RRSDEV	0.84	1.00	0.97	0.53	0.65	0.75	0.71
RR6_46	0.88	0.97	1.00	0.53	0.66	0.79	0.76
RRGT00	0.50	0.53	0.53	1.00	0.88	0.58	0.33
RRGE03	0.66	0.65	0.66	0.88	1.00	0.78	0.46
RRGE15	0.87	0.75	0.79	0.58	0.78	1.00	0.72
RRGE50	0.93	0.71	0.76	0.33	0.46	0.72	1.00

C.5 Voorspelling48

	RRENS	RRSDEV	RR6_46	RRGT00	RRGE03	RRGE15	RRGE50
RRENS	1.00	0.86	0.90	0.49	0.64	0.85	0.93
RRSDEV	0.86	1.00	0.97	0.51	0.62	0.75	0.77
RR6_46	0.90	0.97	1.00	0.51	0.64	0.78	0.81
RRGT00	0.49	0.51	0.51	1.00	0.88	0.57	0.30
RRGE03	0.64	0.62	0.64	0.88	1.00	0.76	0.42
RRGE15	0.85	0.75	0.78	0.57	0.76	1.00	0.70
RRGE50	0.93	0.77	0.81	0.30	0.42	0.70	1.00

D Goodness of Fit

		00-12 UTC		12-00 UTC	
		24 uur	48 uur	12 uur	36 uur
p-waarde	Intercept	<2e-16	<2e-16	<2e-16	<2e-16
	RRENS	<2e-16	<2e-16	<2e-16	<2e-16
AIC		214.95	257.14	212.15	262.03

Tabel D.1: Goodness of Fit tabel De Bilt.

		00-12 UTC		12-00 UTC	
		24 uur	48 uur	12 uur	36 uur
p-waarde	Intercept	<2e-16	<2e-16	<2e-16	<2e-16
	RRENS	<2e-16	<2e-16	<2e-16	<2e-16
AIC		157.26	179.11	227.96	278.72

Tabel D.2: Goodness of Fit tabel Schiphol.

D.1 Twee-steden-modellen

		00-12 UTC		12-00 UTC	
		24 uur	48 uur	12 uur	36 uur
p-waarde	Intercept	<2e-16	<2e-16	<2e-16	<2e-16
	RRENS DB	0.14336	0.3240	0.000189	0.4400
	RRENS S	0.00194	0.0132	0.378593	0.0213
AIC		122.95	147.34	161.53	205.76

Tabel D.3: Goodness of Fit tabel, minstens 5 mm in zowel De Bilt als Schiphol.

		00-12 UTC		12-00 UTC	
		24 uur	48 uur	12 uur	36 uur
p-waarde	Intercept	<2e-16	<2e-16	<2e-16	<2e-16
	RRENS DB	0.00508	0.244710	0.145	0.457
	RRENS S	0.00192	0.000358	2.05e-06	6.61e-07
AIC		225.41	264.71	246.04	296.32

Tabel D.4: Goodness of Fit tabel, minstens 5 mm in De Bilt of Schiphol.

E R-Code

E.1 Historische data-analyse

De code die hieronder is beschreven, maakt gebruik van de historische gegevens van de De Bilt tussen 1 januari 1987 en 31 december 2016. Van deze gegevens is alleen de uursom van de neerslag gebruikt (RH). Onderstaande code is ook gebruikt voor het analyseren van de historische data gemeten in Schiphol. Dit werd gedaan door het geïmporteerde bestand te vervangen en overal "De Bilt"(of soortgelijk geschreven) te vervangen door "Schiphol".

Monthly_Precipitation_DeBilt.R

Pascalle

Wed Jun 21 19:14:38 2017

```
#Clear workspace
rm(list = ls());

#Load Packages
library(ggplot2);
library(gridExtra);
library(plyr);
library(xtable);

#Load historical data, change the unrelevant -1 into 0's
# and make the unit mm instead of 0.1mm
KNMI_Hourly<-read.table("./Historisch/KNMI_87-16_hourly_DeBilt.txt",
                        header = TRUE,sep=',');
attach(KNMI_Hourly);
KNMI_Hourly$RH[RH==-1]<-0;
KNMI_Hourly$RH<-(KNMI_Hourly$RH)/10;

#Aggregate without column HH
KNMI_Daily<-aggregate( cbind(RH) ~ YYYYMMDD , data = KNMI_Hourly ,
                       FUN = sum );
attach(KNMI_Daily);

#Percentage >5mm
count(KNMI_Daily$RH>5)[2,2]/count(KNMI_Daily$RH>5)[1,2];

#Sorting per month
#Create monthly-vectors
Jan<-vector(); Feb<-vector(); Mar<-vector();
Apr<-vector(); May<-vector(); Jun<-vector();
Jul<-vector(); Aug<-vector(); Sep<-vector();
Oct<-vector(); Nov<-vector(); Dec<-vector();

for (i in 1:length(YYYYMMDD)){
  if (floor(YYYYMMDD[i]/100)%%100 == 1){
    Jan<-append(Jan, RH[i]);
  }
  if (floor(YYYYMMDD[i]/100)%%100 == 2){
    Feb<-append(Feb, RH[i]);
  }
  if (floor(YYYYMMDD[i]/100)%%100 == 3){
    Mar<-append(Mar, RH[i]);
  }
```

```
  if (floor(YYYYMMDD[i]/100)%%100 == 4){
    Apr<-append(Apr, RH[i]);
  }
  if (floor(YYYYMMDD[i]/100)%%100 == 5){
    May<-append(May, RH[i]);
  }
  if (floor(YYYYMMDD[i]/100)%%100 == 6){
    Jun<-append(Jun, RH[i]);
  }
  if (floor(YYYYMMDD[i]/100)%%100 == 7){
    Jul<-append(Jul, RH[i]);
  }
  if (floor(YYYYMMDD[i]/100)%%100 == 8){
    Aug<-append(Aug, RH[i]);
  }
  if (floor(YYYYMMDD[i]/100)%%100 == 9){
    Sep<-append(Sep, RH[i]);
  }
  if (floor(YYYYMMDD[i]/100)%%100 == 10){
    Oct<-append(Oct, RH[i]);
  }
  if (floor(YYYYMMDD[i]/100)%%100 == 11){
    Nov<-append(Nov, RH[i]);
  }
  if (floor(YYYYMMDD[i]/100)%%100 == 12){
    Dec<-append(Dec, RH[i]);
  }
}

#making datafram of months
#note: not all months have same length
MonthsLong <- data.frame(Jan,Mar,May,Jul,Aug,Oct,Dec);
MonthsShort <- data.frame(Apr,Jun,Sep,Nov);
MonthsFeb <- data.frame(Feb);
df.list <- list(MonthsLong,MonthsShort,MonthsFeb);

#Plotting histograms of precipitation >0mm; package:ggplot2
for (dat in df.list){
  for (Month in colnames(dat)){
    assign(paste0("Hist", Month), ggplot(data = dat, aes_string(Month))
           + geom_histogram(aes(y = ..density..),
                            breaks=seq(0.01, 40, by=1.5),
                            col = 'black',
                            fill = 'blue',
                            alpha = 0.4) +
           labs(title = paste0("Wetdays ", Month),
                x = "Daily Rainfall [mm]") +
           theme(plot.title = element_text(hjust = 0.5)) +
           xlim(c(0,40)) +
           ylim(c(0,.3)));
```

```
  }
}

#making one figure. Package: GridExtra
jpeg("WetdaysHistYear_DeBilt.jpg",
     width = 1280, height = 720, pointsize = 18);
grid.arrange(HistJan, HistFeb, HistMar, HistApr, HistMay, HistJun,
             HistJul, HistAug, HistSep, HistOct, HistNov, HistDec);
dev.off();

#making table (matrix) with data, Package: plyr
#Wetdays: Total amount of wetdays over 30 years per month
#Non Wetdays: Total amount of dry days over 30 years per month
#Wetdays Av.: Percentage of wetdays over 30 years per month [%]
#Precipitation: The average precipitation over 30 years per month [mm]
#Prec. Av.: The average precipitation if it rains on a day over 30 year per
month [mm]
M<-matrix(ncol=13,nrow=6);
M[1,1]<-"";
M[2,1]<-"Wetdays";
M[3,1]<-"Non Wetdays";
M[4,1]<-"Wetdays Av.";
M[5,1]<-"Precipitation";
M[6,1]<-"Prec. Av.";
for (i in 1:length(MonthsLong)){
  M[1,i+1]<-colnames(MonthsLong)[[i]];
  M[2,i+1]<-count(MonthsLong[[i]]>0)[2,2];
  M[3,i+1]<-count(MonthsLong[[i]]>0)[1,2];
  M[4,i+1]<-round(count(MonthsLong[[i]]>0)[2,2]/
                            length(MonthsLong[[i]]),2);
  M[5,i+1]<-round(mean(MonthsLong[[i]]),2);
  M[6,i+1]<-round(sum(MonthsLong[[i]])/sum(MonthsLong[[i]]>0),2);
}
for (i in 1:length(MonthsShort)){
  M[1,i+1+length(MonthsLong)]<-colnames(MonthsShort)[[i]];
  M[2,i+1+length(MonthsLong)]<-count(MonthsShort[[i]]>0)[2,2];
  M[3,i+1+length(MonthsLong)]<-count(MonthsShort[[i]]>0)[1,2];
  M[4,i+1+length(MonthsLong)]<-round(count(MonthsShort[[i]]>0)[2,2]/
                                          length(MonthsShort[[i]]),2);
  M[5,i+1+length(MonthsLong)]<-round(mean(MonthsShort[[i]]),2);
  M[6,i+1+length(MonthsLong)]<-round(sum(MonthsShort[[i]])/
                                          sum(MonthsShort[[i]]>0),2);
}
M[1,2+length(MonthsLong)+length(MonthsShort)]<-colnames(MonthsFeb)[[1]];
M[2,2+length(MonthsLong)+length(MonthsShort)]<-
          count(MonthsFeb[[1]]>0)[2,2];
M[3,2+length(MonthsLong)+length(MonthsShort)]<-
          count(MonthsFeb[[1]]>0)[1,2];
M[4,2+length(MonthsLong)+length(MonthsShort)]<-
          round(count(MonthsFeb[[1]]>0)[2,2]/length(MonthsFeb[[1]]),2);
```

```
M[5,2+length(MonthsLong)+length(MonthsShort)]<-
          round(mean(MonthsFeb[[1]]),2);
M[6,2+length(MonthsLong)+length(MonthsShort)]<-
          round(sum(MonthsFeb[[1]])/sum(MonthsFeb[[1]]>0),2);

#Swap collumns to get right order
M[,c(3,4,5,6,7,8,9,10,11,12,13)]<-M[,c(13,3,9,4,10,5,6,11,7,12,8)];
#print table, package: xtable
print(xtable(M));
```

E.2 één-stad-model 12- en 36-uursverwachting

De code die in deze bijlage is gezet is toepasbaar op de 12- en 36-uursverwachtingen verkregen van het model van het Europees Centrum voor Weersverwachtingen op Middellange Termijn en wanneer "De Bilt""(of soortgelijk geschreven) en "260""wordt vervangen door respectievelijk "Schiphol'ën "240'ïs de code ook geschikt voor het opstellen en uitvoeren van regressie in Schiphol.

RegressionGLM_12_DB.R

Pascalle

Wed Jun 21 19:31:16 2017

```
setwd("C:/Users/Pascalle/Google Drive/Pascalle/Bachelor/BEP")

# clear workspace
rm(list = ls());

# Load packages
library("car");
library("plyr");
library("verification");
library("xtable");

# Open and attaching historical data file
Hourly <- read.table("./Historisch/KNMI_14-16_hourly_DeBilt.txt",
                     header = TRUE, sep = ',');

# Delete Unrelevant rows (12-00 UTC)
Hourly <- Hourly[!c(Hourly$HH == 1 | Hourly$HH == 2 | Hourly$HH == 3 |
                  Hourly$HH == 4 | Hourly$HH == 5 |Hourly$HH == 6 |
                  Hourly$HH == 7 | Hourly$HH == 8 |Hourly$HH == 9 |
                  Hourly$HH == 10 | Hourly$HH == 11 |Hourly$HH == 12),];

# Set all the -1 to 0's and 0.1mm to 1mm.
Hourly$RH[Hourly$RH == -1] <- 0;
Hourly$RH <- (Hourly$RH)/10;
attach(Hourly);

# Aggregate, to get daily details
Daily <- aggregate( cbind(RH) ~ YYYYMMDD , data = Hourly , FUN = sum );
Daily <- Daily[!c(Daily$YYYYMMDD == 20161231),];
detach(Hourly);

# Open prediction data
Voorspelling12 <-
    read.table("./Voorspellingen/ECMO12_FP012_06260_1416.dat",
               header = TRUE);

# Validation date is at 00UTC, 1 day ahead
Voorspelling12 <- Voorspelling12[!c(Voorspelling12$datum == 20140101),];

# Set 0.1mm to 1mm in RRENS, RRSDEV, RR6_46
Voorspelling12[, c(2,3,4)] <- Voorspelling12[, c(2,3,4)]/10;
```

```r
# Deleting non-valuable rows and get even sets of data
k <- 0; #counter
for (i in 1:length(Voorspelling12$RRENS)){
  if (
    Voorspelling12$RRENS[i-k]>=9999
  ){
    Voorspelling12 <- Voorspelling12[-c(i-k), ];
    Daily <- Daily[-c(i-k), ];
    k <- k+1;
  }
}

# Regression and test subset
Daily_reg <- subset(Daily, YYYYMMDD < 20160101);
Daily_test <- subset(Daily, YYYYMMDD >= 20160101);

V12_reg <- subset(Voorspelling12, datum < 20160102);
V12_test <- subset(Voorspelling12, datum >= 20160102);

#Percentage Precipitation >= 5mm
count(Daily_reg$RH >= 5);

# Regression
# Scatterplots. Package: car
jpeg("ScatterplotLM12.jpg", width = 1280, height = 720, pointsize = 18);
scatterplot(V12_reg$RRENS, Daily_reg$RH,
            main = "12-uurs voorspelling, De Bilt",
            ylab = "RH [mm]", xlab = "RRENS [mm]",
            boxplots = FALSE, smoother = FALSE, pch = 20);
dev.off();

# Dependency check of usefull variables. Package: xtable
corMat12 <- cor(Voorspelling12[-c(1,9)]);
print(xtable(corMat12)); #table to LaTeX

# Add succes column. 1 if succes, 0 otherwise
Daily_reg$Succes <- ifelse(Daily_reg$RH >= 5, 1, 0);

# Correlation of variables with respons variable
cor(V12_reg[-c(1,9)], Daily_reg$Succes);

# Succes Scatterplot
jpeg("ScatterSucces12.jpg", width = 1280, height = 720, pointsize = 18);
plot(V12_reg$RRENS, jitter(Daily_reg$Succes),
     xlab = "RRENS [mm]", ylab = "Succes",
     main = "12-uurs voorspelling with succes, De Bilt",
     pch = 20, col = ifelse((V12_reg$RRENS >= 5 & Daily_reg$Succes == 1)
                            | (V12_reg$RRENS < 5 & Daily_reg$Succes == 0),
                              "green", "red"));
```

```
abline(v = 5);
dev.off();

# Perform Logistic regressions
#  one explanatory variable, RRENS
model12 <- glm(Daily_reg$Succes ~ V12_reg$RRENS
               , family = binomial(link = "logit"));

# Summary's to see which model is the "best"
summary(model12);

# Open the Prediction function
source("Prediction.R");

# Perform prediction(). package: Plyr
beta12 <- matrix(c(coef(model12)[1], coef(model12)[2]), ncol = 1);
X12 <- matrix(cbind(rep(1, length(V12_test$RRENS)),V12_test$RRENS[]),
              ncol = 2);
Daily_test$Pred12 <- Prediction(beta12, X12);

# Verification. Package; verification
ver <- verify(ifelse(Daily_test$RH>=5,1,0),Daily_test$Pred12);
jpeg("Verification12.jpg", width = 1280, height = 720, pointsize = 36);
reliability.plot(ver, titl = "Verificatieplot (+12), De Bilt");
dev.off();
BS <- brier(ifelse(Daily_test$RH>=5,1,0),Daily_test$Pred12,
            bins = FALSE);

# Comparison to RRGE50
ver50 <- verify(ifelse(Daily_test$RH>=5,1,0), V12_test$RRGE50[]/100);
jpeg("RRGE50_12.jpg", width = 1280, height = 720, pointsize = 36);
reliability.plot(ver50, titl = "Verificatieplot RRGE50 (+12), De Bilt");
dev.off();
BS50 <- brier(ifelse(Daily_test$RH>=5,1,0),V12_test$RRGE50[]/100,
              bins = FALSE);

# False Positive/False Negative. Package; Plyr
# Boundary: >50% == 1
# 1 : TRUE Positive, prediction > 5mm & real > 5mm
# -1: TRUE Negative, prediction < 5mm & real < 5mm
# 2 : FALSE Positive, prediction > 5mm & real < 5mm
# -2: FALSE Negative, prediction < 5mm & real > 5mm
# Regression model
Bound <- 0.5;
bound <- 50;
Daily_test$Pred <- ifelse(Daily_test$Pred12 > Bound &
                            Daily_test$RH >= 5, 1,
                          ifelse(Daily_test$Pred12 <= Bound &
                                   Daily_test$RH < 5, -1,
                          ifelse(Daily_test$Pred12 > Bound &
```

```
                                    Daily_test$RH <= 5, 2,-2)));
count(Daily_test$Pred);

V12_test$Pred <- ifelse(V12_test$RRGE50 > bound &
                          Daily_test$RH >= 5, 1,
                        ifelse(V12_test$RRGE50 <= bound &
                                 Daily_test$RH < 5, -1,
                        ifelse(V12_test$RRGE50 > bound &
                                 Daily_test$RH <= 5, 2, -2)));
count(V12_test$Pred);

# Save the chosen model for later use
saveRDS(c(Daily_reg[c("YYYYMMDD", "RH", "Succes")],
          V12_reg["RRENS"]), file="DB_1200UTC_regression12.Rdata");
saveRDS(c(Daily_test[c("YYYYMMDD", "RH")],
          V12_test["RRENS"]), file="DB_1200UTC_Test12.Rdata");
saveRDS(model12, file = "DB_1200UTC_model12.rdata");
```

E.3 één-stad-model 24- en 48-uursverwachting

De code die in deze bijlage is gezet is toepasbaar op de 24- en 48-uursverwachtingen verkregen van het model van het Europees Centrum voor Weersverwachtingen op Middellange Termijn en wanneer "De Bilt""(of soortgelijk geschreven) en "260""wordt vervangen door respectievelijk "Schiphol'ën "240'ïs de code ook geschikt voor het opstellen en uitvoeren van regressie in Schiphol.

RegressionGLM_24_DB.R

Pascalle

Wed Jun 21 19:42:03 2017

```
#clears workspace
rm(list = ls());

#Load packages
library("car");
library("plyr");
library("verification");
library("xtable");

#Open and attaching historical data file
Hourly <- read.table("./Historisch/KNMI_14-16_hourly_DeBilt.txt",
                     header = TRUE, sep = ',');

#Delete Unrelevant rows (12-00 UTC)
Hourly <- Hourly[!c(Hourly$HH == 13 | Hourly$HH == 14 | Hourly$HH == 15 |
                Hourly$HH == 16 | Hourly$HH == 17 | Hourly$HH == 18 |
                Hourly$HH == 19 | Hourly$HH == 20 | Hourly$HH == 21 |
                Hourly$HH == 22 | Hourly$HH == 23 | Hourly$HH == 24),];

#Set all the -1 to 0's and 0.1mm to 1mm.
Hourly$RH[Hourly$RH==-1] <- 0;
Hourly$RH <- (Hourly$RH)/10;
attach(Hourly);

#Aggregate, to get daily details
Daily <- aggregate(cbind(RH) ~ YYYYMMDD , data = Hourly , FUN = sum );
detach(Hourly);

#Open prediction data
Voorspelling24 <-
      read.table("./Voorspellingen/ECMO12_FP024_06260_1416.dat",
                 header = TRUE);

#Set 0.1mm to 1mm in RRENS, RRSDEV, RR6_46
Voorspelling24[, c(2,3,4)] <- Voorspelling24[, c(2,3,4)]/10;

#Deleting non-valuable rows and get even sets of data and same moment
k <- 0; #counter
for (i in 1:length(Voorspelling24$RRENS)){
  if (
    Voorspelling24$RRENS[i-k]>=9999
```

```
      ){
    Voorspelling24 <- Voorspelling24[-c(i-k), ];
    Daily <- Daily[-c(i-k), ];
    k <- k+1;
  }
}

# Regression and test subset
Daily_reg <- subset(Daily, YYYYMMDD < 20160101);
Daily_test <- subset(Daily, YYYYMMDD >= 20160101);

V24_reg <- subset(Voorspelling24, datum < 20160101);
V24_test <- subset(Voorspelling24, datum >= 20160101);



# Regression
# Scatterplots. Package: car
jpeg("ScatterplotLM24.jpg", width = 1280, height = 720, pointsize = 18);
scatterplot(V24_reg$RRENS, Daily_reg$RH,
            main="24-uurs voorspelling, De Bilt",
            ylab = "RH [mm]", xlab = "RRENS [mm]",
            boxplots = FALSE, smoother = FALSE, pch = 20);
dev.off();

# Dependency check of usefull variables. Package: xtable
corMat24 <- cor(V24_reg[-c(1,9)]);
print(xtable(corMat24)); #table to LaTeX

#Add succes column. 1 if succes, 0 otherwise
Daily_reg$Succes <- ifelse(Daily_reg$RH >= 5, 1, 0);

# Correlation of variables with respons variable
cor(V24_reg[-c(1,9)],Daily_reg$Succes);

#Succes Scatterplot
jpeg("ScatterSucces24.jpg", width = 1280, height = 720, pointsize = 18);
plot(V24_reg$RRENS, jitter(Daily_reg$Succes),
     xlab = "RRENS [mm]", ylab = "Succes",
     main = "24-uurs voorspelling met succes, De Bilt",
     pch = 20, col = ifelse((V24_reg$RRENS >= 5 & Daily_reg$Succes == 1)
                            | (V24_reg$RRENS < 5 & Daily_reg$Succes == 0),
                            "green", "red"));
abline(v = 5);
dev.off();

# Perform Logistic regressions
# one explanatory variable, RRENS
model24 <- glm(Daily_reg$Succes ~ V24_reg$RRENS
, family = binomial(link = "logit"));
```

```
# Summary to see in the end which model is the "best"
summary(model24);

# Open the Prediction function
source("Prediction.R");

# Perform prediction(). package: Plyr
beta24 <- matrix(c(coef(model24)[1], coef(model24)[2]), ncol = 1);
X24 <- matrix(cbind(rep(1, length(V24_test$RRENS)),V24_test$RRENS[]),
              ncol = 2);
Daily_test$Pred24 <- Prediction(beta24, X24);

# Verification. Package; verification
ver <- verify(ifelse(Daily_test$RH >= 5,1,0),Daily_test$Pred24);
jpeg("Verification24.jpg", width = 1280, height = 720, pointsize = 36);
reliability.plot(ver, titl = "Verificatieplot (+24), De Bilt");
dev.off();
BS <- brier(ifelse(Daily_test$RH >= 5,1,0),Daily_test$Pred24,
            bins = FALSE);

# Comparison to RRGE50
ver50 <- verify(ifelse(Daily_test$RH >= 5,1,0), V24_test$RRGE50[]/100);
jpeg("RRGE50_24.jpg", width = 1280, height = 720, pointsize = 36);
reliability.plot(ver50, titl = "Verificatieplot RRGE50 (+24), De Bilt");
dev.off()
BS50 <- brier(ifelse(Daily_test$RH >= 5,1,0),V24_test$RRGE50[]/100,
              bins = FALSE);

# False Positive/False Negative. Package; Plyr
# Boundary: >50% == 1
# 1 : TRUE Positive, prediction > 5mm & real > 5mm
# -1: TRUE Negative, prediction < 5mm & real < 5mm
# 2 : FALSE Positive, prediction > 5mm & real < 5mm
# -2: FALSE Negative, prediction < 5mm & real > 5mm
# Regression model
Bound <- 0.5;
bound <- 50;
Daily_test$Pred <- ifelse(Daily_test$Pred24 > Bound &
                            Daily_test$RH >= 5, 1,
                          ifelse(Daily_test$Pred24 <= Bound &
                                   Daily_test$RH < 5, -1,
                          ifelse(Daily_test$Pred24 > Bound &
                                   Daily_test$RH <= 5, 2, -2)));
count(Daily_test$Pred);

V24_test$Pred <- ifelse(V24_test$RRGE50 > bound &
                          Daily_test$RH >= 5, 1,
                        ifelse(V24_test$RRGE50 <= bound
                               & Daily_test$RH < 5, -1,
                        ifelse(V24_test$RRGE50 > bound &
```

```
                                    Daily_test$RH <= 5, 2, -2)));
count(V24_test$Pred);

# Save the chosen model for later use
saveRDS(c(Daily_reg[c("YYYYMMDD", "RH", "Succes")],
          V24_reg["RRENS"]), file="DB_0012UTC_regression24.Rdata");
saveRDS(c(Daily_test[c("YYYYMMDD", "RH")],
          V24_test["RRENS"]), file = "DB_0012UTC_Test24.Rdata");
saveRDS(model24, file = "DB_0012UTC_model24.rdata");
```

E.4 Twee steden combineren

In de code hier beschreven staat de code voor het implementeren van de twee-steden-modellen, waarbij in zowel De Bilt als Schiphol minstens 5 mm neerslag valt en waarbij in minstens een van die twee steden tenminste 5 mm neerslag zal vallen. Daarnaast is er ook de implementatie van de generaliseerbaarheidscheck van de één-stad-modellen beschreven. De code hieronder is voor de 12-uursgegevens voor de overige momenten kan gelijke code worden gebruikt.

TwoCities_12.R

Pascalle

Wed Jun 21 20:17:21 2017

```
# Clear workspace
rm(list = ls());

# Load packages
library("car");
library("plyr");
library("verification");

# Open regression data
data_DB <- data.frame(readRDS("DB_1200UTC_regression12.Rdata"));
data_S <- data.frame(readRDS("S_1200UTC_regression12.Rdata"));

# Open test data
test_DB <- data.frame(readRDS("DB_1200UTC_Test12.Rdata"));
test_S <- data.frame(readRDS("S_1200UTC_Test12.Rdata"));

# Get even rows from same dates, change number for a new try
data_DB <- subset(data_DB, (data_DB$YYYYMMDD %in% data_S$YYYYMMDD));
data_S  <- subset(data_S, (data_S$YYYYMMDD %in% data_DB$YYYYMMDD));
test_DB <- subset(test_DB, (test_DB$YYYYMMDD %in% test_S$YYYYMMDD));
test_S <- subset(test_S, (test_S$YYYYMMDD %in% test_DB$YYYYMMDD));

# Plot of precipitation De Bilt against Schiphol. Package; car
jpeg("DB_tegen_S.jpg", width = 720, height = 720, pointsize = 18);
scatterplot(data_DB$RH, data_S$RH,
            main = "Gevallen Neerslag, De Bilt tegen Schiphol",
            xlab = "De Bilt", ylab = "Schiphol",
            boxplots = FALSE, smoother = FALSE, pch = 20);
dev.off();

# Make one dataframe of usefull information
df_TwoCities <- data.frame(data_DB$YYYYMMDD, data_DB$RRENS,
                           data_S$RRENS);
colnames(df_TwoCities) <- c("YYYYMMDD", "DB_RRENS", "S_RRENS");

# Correlationcoefficients
cor(df_TwoCities$DB_RRENS, df_TwoCities$S_RRENS);

cor(data_DB$RH, data_S$RH);

#### Regression, both cities >= 5mm
# Set new response variable Y_both
```

```r
df_TwoCities$Y_both <- ifelse(data_DB$Succes == 1 & data_S$Succes == 1,
                              1, 0);

# Percentage Precipitation >=5mm
count(df_TwoCities$Y_both);

# Perform regression
model_both <- glm(df_TwoCities$Y_both ~ df_TwoCities$DB + df_TwoCities$S
                  , family = binomial(link = "logit"));
summary(model_both);

#### Regression, at least one city >= 5mm
# Set new response variable Y_both
df_TwoCities$Y_one <- ifelse(data_DB$Succes == 1 | data_S$Succes == 1,
                             1, 0);

# Percentage Precipitation >=5mm
count(df_TwoCities$Y_one);

# Perform regression
model_one <- glm(df_TwoCities$Y_one ~ df_TwoCities$DB + df_TwoCities$S
                 , family = binomial(link = "logit"));
summary(model_one);

################################################################################
# Predict precipitation of De Bilt with model of Schiphol
# Open model of Schiphol
model_YDB <- readRDS("S_1200UTC_model12.Rdata");

# Open the Prediction function
source("Prediction.R");

# Perform prediction() for model_YDB. package: Plyr
beta_YDB <- matrix(c(coef(model_YDB)[1], coef(model_YDB)[2]),
                   ncol = 1);
X_YDB <- matrix(cbind(rep(1, length(test_S$RRENS)),test_S$RRENS[]),
                ncol = 2);
test_DB$Pred <- Prediction(beta_YDB, X_YDB);

# False Positive/False Negative. Package; Plyr
# Boundary: >50% == 1
# 1 : TRUE Positive, prediction > 5mm & real > 5mm
# -1: TRUE Negative, prediction < 5mm & real < 5mm
# 2 : FALSE Positive, prediction > 5mm & real < 5mm
# -2: FALSE Negative, prediction < 5mm & real > 5mm
# Regression model
Bound <- 0.5;
bound <- 50;
test_DB$PredSucces <- ifelse(test_DB$Pred > Bound &
                               test_DB$RH >= 5, 1,
```

```
                              ifelse(test_DB$Pred <= Bound &
                                       test_DB$RH < 5, -1,
                              ifelse(test_DB$Pred > Bound &
                                       test_DB$RH <= 5, 2, -2)));
count(test_DB$PredSucces);

# Verification model_YDB
ver <- verify(ifelse(test_DB$RH>=5,1,0),test_DB$Pred);
jpeg("Verification_12YDB.jpg", width = 1280, height = 720,
     pointsize = 36);
reliability.plot(ver,
              titl = "Verificatieplot (+12) gecombineerd, Y = De Bilt");
dev.off();
BS <- brier(ifelse(test_DB$RH >= 5, 1, 0),test_DB$Pred, bins = FALSE);

#### Predict precipitatio of Schiphol with model of De Bilt
#Open Model of De Bilt
model_YS <- readRDS("DB_1200UTC_model12.Rdata");

# Perform prediction() for model_YDB. package: Plyr
beta_YS <- matrix(c(coef(model_YS)[1], coef(model_YS)[2]), ncol = 1);
X_YS <- matrix(cbind(rep(1, length(test_DB$RRENS)),test_DB$RRENS[]),
               ncol = 2);
test_S$Pred <- Prediction(beta_YS, X_YS);

# False Positive/False Negative. Package; Plyr
# Boundary: >50% == 1
# 1 : TRUE Positive, prediction > 5mm & real > 5mm
# -1: TRUE Negative, prediction < 5mm & real < 5mm
# 2 : FALSE Positive, prediction > 5mm & real < 5mm
# -2: FALSE Negative, prediction < 5mm & real > 5mm
# Regression model
Bound <- 0.5
test_S$PredSucces <- ifelse(test_S$Pred > Bound &
                             test_S$RH >= 5, 1,
                            ifelse(test_S$Pred <= Bound &
                                     test_S$RH < 5, -1,
                            ifelse(test_S$Pred > Bound &
                                     test_S$RH <= 5, 2, -2)));
count(test_S$PredSucces);

# Verification model_YS
ver <- verify(ifelse(test_S$RH >= 5,1,0),test_S$Pred);
jpeg("Verification_12YS.jpg", width = 1280, height = 720,
     pointsize = 36);
reliability.plot(ver,
                 titl = "Verificatieplot (+12) gecombineerd, Y = Schiphol");
dev.off();
BS <- brier(ifelse(test_S$RH >= 5,1,0),test_S$Pred, bins = FALSE);
```

F Glossary

AIC	Akaike Informatie Criterium
datum	Validatiedatum
ECMWF	Europees Centrum voor Weersverwachtingen op Middellange Termijn
ENS	ensemble, de 51 weerverwachtingsberekeningen samen
GLM	Gegeneraliseerd Lineair Model
HH	Meettijdstip in uren
KNMI	Koninklijk Nederlands Meteorologisch Instituut
RH	uursom van de neerslag (in 0.1 mm) (-1 voor <0.05 mm)
RR6-46	waarde lid 6 afgetrokken van lid 46
RRENS	neerslag gemiddeld over ENS (0.1 mm)
RRGE03	% leden met minstens 0.3 mm neerslag
RRGE15	% leden met minstens 1.5 mm neerslag
RRGE50	% leden met minstens 5.0 mm neerslag
RRGE100	% leden met minstens 10.0 mm neerslag
RRGT00	% leden met meer dan 0 mm neerslag
RRSDEV	standaarddeviatie neerslag ENS (0.1 mm)
STN	meetstationnummer (240 = Schiphol, 260 = De Bilt)
UTC	gecoördineerde wereldtijd
YYYYMMDD	meetdatum (YYYY = jaar, MM = maand, DD = dag)

Referenties

[1] KNMI. https://www.knmi.nl/kennis-en-datacentrum/uitleg/regenintensiteit, 2017.

[2] KNMI. http://www.sciamachy-validation.org/research/statistical_postprocessing/, 2017.

[3] ECMWF. https://www.ecmwf.int, 2017.

[4] KNMI. http://www.knmi.nl/nederland-nu/weer/waarschuwingen-en-verwachtingen/weer-en-klimaatpluimvoordeactueleverwachtingen, 2017.

[5] ECMWF, "User guide to ecmwf forecast products," 2015. https://www.ecmwf.int/sites/default/files/elibrary/2015/16559-user-guide-ecmwf-forecast-products.pdf.

[6] KNMI. http://projects.knmi.nl/klimatologie/uurgegevens/selectie.cgi, 2017.

[7] D. S. Wilks, *Statistical methods in the atmospheric sciences*, vol. 100. Academic press, 2011. pp. 215-394.

[8] A. Khuri, *Linear Model Methodology*. CRC Press, 2009. pp. 473-509.