Non-Parametric Bayesian Networks (NPBNs) versus Ensemble Kalman Filter (EnKF) in Reservoir Simulation with non-Gaussian Measurement Noise

Aurelius A. Zilko

Department of Applied Mathematics
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology
Delft University of Technology
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft Institute of Applied Mathematics

Non-Parametric Bayesian Networks (NPBNs) versus Ensemble Kalman Filter (EnKF) in Reservoir Simulation with non-Gaussian Measurement Noise

A thesis submitted to the Delft Institute of Applied Mathematics in partial fulfillment of the requirements for the degree MASTER OF SCIENCE in APPLIED MATHEMATICS by

Aurelius Armando Zilko

Delft, The Netherlands August 2012

Copyright © 2012 by Aurelius Armando Zilko. All rights reserved.
MSc THESIS APPLIED MATHEMATICS

"Non-Parametric Bayesian Networks (NPBNs) versus Ensemble Kalman Filter (EnKF) in Reservoir Simulation with non-Gaussian Measurement Noise"

Aurelius Armando Zilko

Delft University of Technology

Daily Supervisor
Dr. Anca Hanea

Responsible Professor
Prof. Dr. Ir. Arnold Heemink

Other thesis committee members:
Dr. Remus Hanea
Mariya Krymskaya, M.Sc

August 2012
Delft, the Netherlands
ABSTRACT

Non-Parametric Bayesian Networks (NPBNs) versus Ensemble Kalman Filter (EnKF) in Reservoir Simulation with non-Gaussian Measurement Noise

by AURELIUS A. ZILKO (4122135)

Lately, the objective of reservoir engineering is to optimize hydrocarbon recovery from a reservoir. To achieve that goal, a good knowledge of the subsurface properties is crucial. The author is concerned with estimating one of the properties of the field: the permeability of a reservoir. To characterize the fluid flow, a two phase (oil-water) 2D model represented as a system of coupled nonlinear partial differential equations which is unsolvable analytically is used. Ensemble Kalman Filter (EnKF) is the most common tool used to deal with this situation. However, it is not the only way. Recently, a research, as presented in [8], on a more general approach based on a dynamic Bayesian network using the Non Parametric Bayesian Networks (NPBN) has been initiated. This research, which uses twin experiment, indicates that the NPBN approach appears to be a promising alternative to EnKF. However, a number of open questions emerge. The first one is the normality assumption for the noise used in the measurements generation in the twin experiment. Even though Gaussian noise for measurements is sensible in the sense that the knowledge about the noise is unavailable, it does not mean that other noise from different distributions cannot be applied. The second one is the exclusion of saturation in the NPBN approach performed in the previous research. This may result in the loss of valuable information. The previous research discovers that NPBN approach seems to work well in recovering only part of the reservoir. The entire permeability field may be approximated by means of interpolation between several approximated parts of the field. Hence, the third question relates to an interpolation method that may be used in recovering the permeability of the entire reservoir. This project aims to experiment on these three key points of interest. A fourth objective is surfaced during the analysis, which is to use an alternative measure of performance to the well-known Root Mean Square Error (RMSE). Along the way, the performance of both EnKF and NPBN are going to be observed and compared one more time.

Keywords: Ensemble Kalman Filter, Non-Parametric Bayesian Networks, Parameter Estimation, Reservoir Engineering, Reservoir Simulation.
Acknowledgements

All the work presented in this thesis report was accomplished during the second year of Applied Mathematics Master programme at Delft University of Technology (TU Delft). It was part of the two year Master programme at TU Delft to obtain the degree of Master of Science. Here, I would like to acknowledge all the people who, at any level, have contributed to the completion of this thesis work and the two year Master programme at TU Delft.

First of all, I would like to thank my daily supervisor, Dr. Anca Hanea, whom without her, it would be impossible to finish the project. Very sincere gratitude is also expressed to Dr. Remus Hanea and Prof. Arnold Heemink who gave me the incredible opportunity to attend the wonderful "7th International EnKF Workshop" in Bergen, Norway in June 2012 to present the main results of this thesis project. It was such a priceless experience for me to be able to take part in a conference with that extraordinary calibre. I would like to thank Mariya Krymskaya for taking some of her precious time to answer my questions and gave me some insight about some problems I encountered along the way. I would also like to acknowledge Prof. Jan Dirk Jansen who has given us some insight from reservoir-engineering point of view to help us building the model. Abundant amount of gratitude goes to all my colleagues as well, for the (lengthy) discussions and ideas about the scientific research.

Of course I must also thank all the professors whose courses I attended during the two year period of my Master study at TU Delft. Not only them, all the non-academic staffs are also thanked for all the kind help and support. Looking back even further, I would like to acknowledge all the mathematics professors and staffs at Universitas Katolik Parahyangan in Bandung, Indonesia, where I spent three and a half year of Bachelor programme; especially Dr. Dharma Lesmono and Dr. Ferry Jaya Permana who believed in
me and were more than happy to write recommendation letters for me to apply for this Master programme at TU Delft.

Regarding the non-academic life in Delft for the past two years, I would like to thank all the friends for the friendship, support, insight, fun, amazing time, awesome experience, and all the traveling across Europe that we have spent together during this amazing two year of life in Delft. To avoid misunderstanding and confusion, they all remain nameless; but I am sure they know who they are.

Last but not least, I would like to express a lot of "thank you" to Dr. Dorota Kurowicka and Prof. Roger Cooke who, in the first place, opened the door for me to take this Master degree programme in TU Delft through the Risk and Environmental Modeling Scholarship. Without them, everything else would not even have been a reality.

_Delft, July 2012_

_Aurelius A. Zilko_
# Table of Contents

Abstract

Acknowledgements

Table of Contents

List of Figures

List of Tables

Chapter I: INTRODUCTION 1

1.1 Background .................................................. 1

1.2 Aim of the Project .......................................... 4

1.3 Organization of Report ................................. 5

Chapter II: THEORETICAL KNOWLEDGE 7

2.1 Bayesian Network, Copula, and Sklar’s Theorem .......... 8

2.1.1 Bayesian Network ...................................... 8

2.1.2 Copula and Sklar’s Theorem .......................... 9

2.2 Ensemble Kalman-Filter (EnKF) ........................ 9
2.3 Kalman Filter in Bayesian Formulation ........................................ 12

Chapter III: EXPERIMENTAL SETUP .............................................. 19

3.1 Problem Description ................................................................ 19

3.1.1 Two Phase Flow ................................................................. 19

3.2 The Twin Experiment ............................................................... 22

3.2.1 Arcs Directionality in the NPBN ......................................... 24

3.3 Measure of Performance .......................................................... 26

Chapter IV: CASE STUDY ................................................................. 27

4.1 Measurements with non-Gaussian Noise .................................... 27

4.1.1 The Generated Measurements ............................................ 30

4.2 The Performance of the NPBN Method ....................................... 32

4.2.1 Permeabilities Estimation for A $7 \times 7$ Grid Block ............... 34

4.2.2 Permeabilities Estimation for A $5 \times 5$ Grid Block ............... 40

4.2.3 Overlapping Grid Block and Consistency of NPBN ............... 44

4.3 Inclusion of Saturation .............................................................. 51

4.3.1 (Updated) Arcs Directionality in the NPBN ......................... 51

4.3.2 Permeabilities Estimation with Saturation ......................... 53

4.4 Choices for Further Experiment .............................................. 57

Chapter V: INTRODUCTION OF INTERPOLATION ......................... 59

5.1 Possible Methods of Interpolation ............................................ 60

5.1.1 Experimental Setup ........................................................... 60

5.1.2 Method 1 (Simplest Linear Interpolation) ............................ 61
A.1.2 The Log-normal Distribution .......................... 113
A.1.3 Logistic Distribution ................................. 114
A.1.4 Mixture Distribution ................................. 115
A.2 Relative Information ................................. 116

Appendix B: MEASUREMENTS GENERATION 119

B.1 Data Spread of Observable Variables .................. 119
B.2 Distribution Fitting ................................... 122
B.3 Goodness of Fit ..................................... 126
  B.3.1 Q-Q Plot Based Approach ......................... 127
  B.3.2 Relative Information Based Approach .......... 131
B.4 Concerns and Possible Further Questions ............. 134

Appendix C: REUSING THE GAUSSIAN NOISE 141

C.1 The Measurements .................................. 141
C.2 Applying the NPBN Method ........................... 142
C.3 Analysis and Comparison with EnKF .................. 143
C.4 Evaluation of the Cubic Interpolation Method ......... 147
List of Figures

2.1 A Simple Bayesian Network ........................................... 8
3.1 Reservoir Field .......................................................... 20
3.2 The chosen "true" permeability field. ............................... 22
3.3 Mean of 900 log permeability chosen as the initial guess. ...... 23
3.4 Choices of the Arcs Directionality. ................................. 25
4.1 The generated measurements with mixture of Gaussian noise ... 30
4.2 The chosen $7 \times 7$ grid block ...................................... 33
4.3 A saturated NPBN for the $7 \times 7$ grid block ...................... 34
4.4 The ENKF and NPBN estimations of the permeability field of the $7 \times 7$ grid block .................................................. 36
4.5 RMSE of EnKF versus NPBN estimation of permeability of the $7 \times 7$ grid block .................................................. 36
4.6 Comparison of adding non-Gaussian noise instead of Gaussian noise .......................................................... 39
4.7 The chosen $5 \times 5$ grid block within bhp proximity ............... 41
4.8 The ENKF and NPBN estimations of the permeability field of the
$5 \times 5$ grid block close to bhp ...................................... 42
4.9 RMSE of EnKF versus NPBN estimation of permeability of the
$5 \times 5$ grid block close to bhp ...................................... 42
4.10 The chosen $5 \times 5$ grid block with overlapping cells ............ 44
4.11 The ENKF and NPBN estimations of the permeability field of the
overlapping $5 \times 5$ grid block ...................................... 46
4.12 RMSE of EnKF versus NPBN estimation of permeability of the
overlapping $5 \times 5$ grid block ...................................... 46
4.13 Comparison of the overlapping cells ............................... 49
4.14 Choices of the Arcs Directionality (Updated) ...................... 52
4.15 The ENKF and NPBN estimations without and with Saturation of
the $5 \times 5$ grid block close to bhp .................................. 54
4.16 RMSE of EnKF versus NPBN estimation with and without
Saturation of permeability of the $5 \times 5$ grid block close to bhp ... 54
4.17 The ENKF and NPBN estimations without and with Saturation of
the overlapping $5 \times 5$ grid block .................................. 55
4.18 RMSE of EnKF versus NPBN estimation with and without
Saturation of permeability of the overlapping $5 \times 5$ grid block ... 55
5.1 Illustration of the simplest case .................................. 62
5.2 Interpolated permeability field with the simplest method ........ 63
5.3 Illustration of method 2 ............................................. 64
C.6 The absolute error of each cell with the NPBN method with Gaussian noise .......................... 147

C.7 The RMSE of the interpolation cells, blocks cells, and relevant block cells .......................... 148
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Parameters of the mixture of Gaussian distributions.</td>
<td>28</td>
</tr>
<tr>
<td>4.2</td>
<td>Some numbers to help the comparison of the overlapping cells.</td>
<td>50</td>
</tr>
<tr>
<td>5.1</td>
<td>RMSE of the first method</td>
<td>62</td>
</tr>
<tr>
<td>5.2</td>
<td>RMSE of the second method</td>
<td>65</td>
</tr>
<tr>
<td>B.1</td>
<td>p-values from Relative Information based hypothesis testing</td>
<td>133</td>
</tr>
<tr>
<td>B.2</td>
<td>p-values of first assumption testing</td>
<td>135</td>
</tr>
<tr>
<td>B.3</td>
<td>p-values of first assumption testing on further time steps</td>
<td>136</td>
</tr>
</tbody>
</table>
1.1 Background

In reservoir engineering, it is of importance to study and understand the properties of physical and chemical processes taking place in the subsurface. One way to achieve this goal is by means of reservoir simulation. Reservoir simulation is a part of reservoir engineering which incorporates computer models to predict the flow of fluids (i.e. water, oil, or gas) through a porous media. Reservoir simulation is used by oil and gas companies in the development of new fields.

One of the rock's properties that is of interest it the permeability. Permeability is a measure of the ease of fluid’s flow through a porous media. In reservoir engineering, the problem of estimating models’ parameters, like permeability, is referred to as a history matching problem.

History matching matches the history of observable variables at the wells with the prediction of these variables’ values from a numerical model by adjusting the parameters. Basically, there are two ways of "correcting" the parameters of the model. The traditional way is the manual history matching.
this way, one runs the reservoir model for some time with a set of parameters and then modifies the parameters such that the observation and the prediction match. The more advance approach is the computer assisted history matching which, just like its name says, incorporates the help of a computer to approximate the parameters.

Data assimilation is one methodology used in computer assisted history matching. Data assimilation combines a mathematical-physical model with available measurements to estimate and predict different environmental process. There are two kinds of methods in data assimilation: the variational approach and the sequential approach. In the variational approach, one determines an objective function, usually defined as the distance between the measured value of data and the forecasted value of that particular data from the model, and optimizes this function (in the case of distance, one wants to minimize the objective function).

The second method is the sequential approach. One of the most famous and powerful tools often used to solve linear systems in this approach is the Kalman filtering [12]. There have been developments of the Kalman filter to handle non linear systems. One of them is the Ensemble Kalman Filter (EnKF) which was introduced in [6]. Currently, EnKF is one of the most used data assimilation methods for the computer assisted history matching in reservoir simulation [1, 5, 15]. There are, however, several limitations in EnKF. One of them is the limited number of ensembles ones may apply in EnKF for it to be affordable computationally. In general, between 50 and 100 ensembles are used
in reservoir engineering applications. This becomes a problem when the number of variables to be estimated in the system is much larger than the ensemble size as this may lead to inaccuracies in the estimation of the outcome. One way to deal with this problem is by means of localization [2, 10]. However, sometimes localizations introduces other inconsistencies in the system.

The author of [8] introduces another sequential approach. This approach utilizes the non parametric Bayesian Networks (NPBN). A BN is a probabilistic graphical model based on acyclic graphs. Graphical models have been used in high dimensional probabilistic modeling in several scientific fields. The popularity of such models has been increasing over time due to its flexibility; hence the theory behind them has been constantly developed and extended. A graphical approach combines probabilistic theory and graph theory to provide a general setting for a model where a number of variables interact. The graphical structure itself is represented by a collection of nodes and arcs where each node represents a random variable and the absence of arc (between two nodes) represents independence between the two unconnected variables.

This thesis project is aimed as a continuation of the initial research presented in [8] over this new NPBN based graphical method. Three key points will be of interest to test or improve the performance of this approach further. Those key points are the distributions of the measured variables, the inclusion of saturation to the NPBN model, and an interpolation method to combine the NPBN approximations of several parts of the reservoir\textsuperscript{1} to obtain the NPBN

\textsuperscript{1}As mentioned in [8], NPBN method works well in recovering only part of the reservoir instead of the entire reservoir at once.
approximation of permeability of the whole reservoir field. Also, in this thesis project, the performance of the NPBN graphical method will be compared with the performance of the EnKF approach over the same subject by means of measure of performance, mainly the Root Mean Square Error (RMSE).

This NPBN based graphical method is still a very new approach used in the history matching problem in reservoir simulation. Many more studies are still needed to investigate different aspects of its performance.

1.2 Aim of the Project

As mentioned in the previous section, there are three key points of interest aimed by this thesis project. The first one is the distributions of the measurements noise, the second one is the inclusion of saturation to the NPBN approach, and the third one is a possible interpolation method to combine the approximations of different parts of the field where the NPBN is performed upon.

In reality, measurements taken from the field are "noisy"; hence, they do not represent the actual exact value of the measured variables at that specific time. This noise may come from technical error, human error, or other possibilities. One of the basic assumptions used in [8] is that the noise of the measurements is a white Gaussian noise, hence it is normally distributed with zero mean and a predetermined variance (for a production well, normally the variance is between 5% and 20%). One of the reasons behind this assumption is the unavailability of knowledge regarding the actual noise of the
measurements. Hence, it is taken as Gaussian. It, however, implies that actually one has the freedom to take any other noise from different distributions. This project will investigate the behavior of the system under non-Gaussian noise of the measurements.

In the NPBN approach performed in [8], one of the variables, saturation, is excluded from the model because theoretically it has an almost constant value at a given time (and a constant is independent of any other variables). This is good news because less variables are now involved in the system. However, this exclusion may cause lost of valuable information. This thesis project also tries to find out whether inclusion of saturation to the model increases the performance of the approach or not.

The author of [8] finds that the NPBN approach seems to work well in recovering only part of the whole reservoir field with limited size. Hence, one idea to estimate the entire permeability field is by using the NPBN method in several parts of the reservoir and combining the results via interpolation. This constitutes one of this thesis’ aims.

1.3 Organization of Report

This report is organized as follows:

Chapter 1 introduces and explains the background and aim of the project. Chapter 2 contains the necessary theoretical background that underlies the simulation and analysis presented in the latter stage of the report. In chapter 3, the experimental setup and twin experiment are briefly introduced.
and explained. The case study is presented in chapter 4. This chapter consists of two parts. The first part is about the comparison between EnKF method and the NPBN method using non-Gaussian measurements noise in several parts of the field. The consistency on the NPBN method will also be investigated in this section. The second part is about the inclusion of saturation. The NPBN method with and without saturation will be performed in several parts of the field. With these results, whether saturation should be included to the model or not will be analyzed. Whichever model performs better (the one with or without saturation) will be used in the next chapter, chapter 5, where interpolation is performed in order to recover the entire permeability field. Before actually performing the interpolation, several methods of interpolation will be investigated. Using the best-performing method of interpolation, the entire field will be recovered. Analysis of the result (including comparison to the obtained result with EnKF) is going to be performed thoroughly as well in this chapter. In chapter 6, another approach of measure of performance, the forecast or prediction, is introduced and implemented. Finally, this report is closed in chapter 7 with conclusions and recommendations.
CHAPTER II
THEORETICAL KNOWLEDGE

The most important theoretical details underlying the project carried out in this report have already been presented in Chapter 2 and Chapter 3 of [8]. In this chapter, theoretical knowledge that has not been provided in [8] but is important for the project will be introduced.

This chapter will start with brief descriptions about the Bayesian Network, copula, and the Sklar’s Theorem. These theories are the ones that underlie the Non-Parametric Bayesian Networks (NPBN) Method, the key interest of this thesis project. Next, the Ensemble Kalman Filter (EnKF) is going to be described. EnKF is the method whose performance is going to be used as a comparison to the performance of the NPBN Method. Then, in the next section, Kalman Filter is going to be explained with a Bayesian approach. This section can be seen as the "connection" between Kalman Filter (which also underlies the EnKF) and the NPBN Method.
2.1 Bayesian Network, Copula, and Sklar’s Theorem

2.1.1 Bayesian Network

Bayesian Network is directed acyclic graphs. The nodes of the graph represent univariate random variables (continuous or discrete) and the arcs represent influences between two variables. The absence of an arc between two nodes represents (conditional) independence between the two variables; but the presence of an arc between two nodes does not automatically imply dependence between them. The dependence should be quantified. After quantification, calculations may reveal independence between variables connected with arcs.

A simple example of a Bayesian Network is presented in Figure 2.1.

**Figure 2.1:** A Simple Bayesian Network

In Figure 2.1, there are five variables (named $V_1$, $V_2$, $V_3$, $V_4$, and $V_5$) which are represented by five nodes. There are arcs directing from $V_1$ to $V_2$, from $V_3$ to $V_2$, from $V_3$ to $V_4$, and from $V_4$ to $V_5$. The directionality of the arcs means that $V_1$ is a cause of $V_2$, and so is $V_3$. It also means that $V_3$ is the *common cause* of both $V_2$ and $V_4$. Because a Bayesian Network has to be acyclic, one cannot have an arc directing from $V_5$ to $V_3$ because this arc yields a cycle ($V_3$ to $V_4$, $V_4$ to $V_5$, and $V_5$ to $V_3$).
2.1.2 Copula and Sklar’s Theorem

The copula of two continuous random variables $X$ and $Y$ is the joint distribution of $F_X(X)$ and $F_Y(Y)$, where $F_X$ and $F_Y$ are the cumulative distribution functions of $X$ and $Y$, respectively [13].

If $\Phi_\rho$ is the bivariate normal cumulative distribution function with product moment correlation $\rho$ and $\Phi^{-1}$ is the inverse of standard univariate normal distribution, then

$$C_\rho(u, v) = \Phi_\rho(\Phi^{-1}(u), \Phi^{-1}(v)), \quad (u, v) \in I^2$$

is called the normal or Gaussian copula [13].

One important result in the copulas theory is the Sklar’s Theorem. The theorem states that a multivariate cumulative distribution function

$$H(x_1, \ldots, x_d) = \Pr(X_1 \leq x_1, \ldots, X_d \leq x_d)$$

of a random vector $(X_1, \ldots, X_d)$ with margins $F_i(x) = \Pr(X_i \leq x)$ can be written as

$$H(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)),$$  \hspace{1cm} (2.2)

where $C$ is a copula [13, 16].

2.2 Ensemble Kalman-Filter (EnKF)

Let $X_k$ denote the state vector and at $k$-th time step. When adding model errors, the functional description of the model reads:

$$X_{k+1} = f(X_k) + G_k W(k)$$  \hspace{1cm} (2.3)
where \( f(X_k) \) represents the model update and \( G_k W(k) \) denotes the model noise. \( G_k \) is the noise input matrix and \( W(k) \) is the white Gaussian model noise process. \( W(k) \) has the properties of zero mean and covariance matrix \( Q(k) \), i.e.

\[
\mathbb{E}\{W(k)\} = 0, \quad \mathbb{E}\{W(k)W(k)^T\} = Q(k).
\]

Moreover, it is assumed that \( W(k) \) is uncorrelated in time (white).

The initial state, \( \hat{X}_0 \), is assumed to be Gaussian with mean \( X_0 \) and covariance matrix \( P_0 \). Hence, it can be written that

\[
\hat{X}(0|0) = X_0, \quad P(0|0) = P_0.
\]

In other words, the simulation usually starts from one point that is quite at a distance from the truth.

When adding measurement errors, the measurement function or observation operator is defined as:

\[
Z_k = M_k Y_k + V(k), \tag{2.4}
\]

where \( Y_k \) denotes the true state, \( M_k \) is a matrix determining the position of available measurements corresponding to the entries of \( Y_k \), and \( V(k) \) is the measurement noise. This measurement noise is assumed to follow a white Gaussian process. In other words, \( V(k) \) is assumed to have zero mean and covariance matrix \( R(k) \), i.e:

\[
\mathbb{E}\{V(k)\} = 0, \quad \mathbb{E}\{V(k)V(k)^T\} = R(k).
\]

An important assumption is that the model, initial state, and measurement noise processes are all independent of each other.
The Ensemble Kalman Filter (EnKF) is a Monte Carlo type method that starts from a large number of realizations, also called ensemble members. Let $N$ denote the number of ensemble members used in an EnKF process. In practice, this number $N$ usually lays around 50 and 100 for EnKF to be computationally affordable. The first initial ensemble of state vectors, $\xi_i(0|0)$, $i=1,\ldots,N$ is generated with mean $X_0$ and covariance matrix $P_0$. Then, to update the ensemble members, realizations $w^{i}_k$ and $v^{i}_k$, the model noise and measurement noise process respectively for each time step and ensemble member, are generated. Next, the algorithm continues recursively using the following recurrence relations:

**Time Update or Model Update or Forecast of State Prediction**

\begin{align*}
\xi_i(k|k-1) &= f(\xi_i(k-1|k-1)) + G_{k-1}w^{i}_{k-1}, \\
\hat{X}(k|k-1) &= \frac{1}{N}\sum_{i=1}^{N}\xi_i(k|k-1), \\
L(k|k-1) &= \begin{bmatrix} \xi_1(k|k-1) - \hat{X}(k|k-1), \ldots, \xi_N(k|k-1) - \hat{X}(k|k-1) \end{bmatrix}, \\
P(k|k-1) &= \frac{1}{N-1}L(k|k-1)L(k|k-1)^T.
\end{align*}

**Measurement Update**

\begin{align*}
\xi_i(k|k) &= \xi_i(k|k-1) + K(k)\left(Z_k - M_k\xi_i(k|k-1) + v^{i}_k\right), \\
\hat{X}(k|k) &= \frac{1}{N}\sum_{i=1}^{N}\xi_i(k|k), \\
L(k|k) &= \begin{bmatrix} \xi_1(k|k) - \hat{X}(k|k), \ldots, \xi_N(k|k) - \hat{X}(k|k) \end{bmatrix}, \\
P(k|k) &= \frac{1}{N-1}L(k|k)L(k|k)^T.
\end{align*}
where $K(k)$ is known as the Kalman Gain matrix which is defined as:

$$
K(k) = P(k|k-1)M_k^T \left( M_k P(k|k-1) M_k^T + R(k) \right)^{-1}.
$$

(2.13)

The ENKF approximation for the state vector at timestep $k$ is hence presented in equation (2.10).

### 2.3 Kalman Filter in Bayesian Formulation

This section is a summary of [14] where Kalman Filter is explained using the Bayesian formalism. This can be seen as a proof of Kalman Filter formulae using the Bayesian approach of conditioning a joint distribution on known values. Under the joint Gaussianity assumption, the well-known Kalman Filter formulae can be derived using the conditional mean and variance of a multivariate normal.

The Bayesian Network (BN), as explained in Chapter 3 of [8], is also based on the Bayesian formulation. The NPBN approach uses the assumption of a joint normal dependence structure. When the dependence is that of a multivariate normal, a simple transformation of the margins alone transforms the joint distribution into a multivariate normal distribution. The NPBN takes advantages of this transformation by performing all calculations in a joint normal "world".

The difference between the NPBN approach and the regular EnKF approach is that NPBN approach works with the transformed variables while EnKF works with the variables as they are. The transformation procedure is described in detail in page 34 - 35 of [8].
Using the same notation as the previous section, let the model update \( f(X_k) \) in equation (2.3) be linear. Hence, it can be written as \( F_{k+1}X_k \) with \( F_{k+1} \) is a matrix. Therefore, equation (2.3) becomes:

\[
X_k = F_k X_{k-1} + G_k W(k) = F_k X_{k-1} + w_k
\]

with \( w_k = G_k W(k) \) for simplicity.

Kalman Filter is a recursive procedure for inference about the state of nature \( X_k \). The idea is that given the data (or measurements) \( Z_k = (Z_k, \ldots, Z_1) \), inference about \( X_k \) can be performed through a direct application of the Bayes’ theorem:

\[
Pr(X_k|Z_k) \propto Pr(Z_k|X_k, Z_{k-1}) \times Pr(X_k|Z_{k-1}).
\]

(2.15)

The left side expression of the equation above is the posterior distribution for \( X \) at time \( k \), whereas the first and second terms on the right side are the likelihood and prior distribution for \( X \), respectively.

At time step \( k-1 \), the following statement is what is known about \( X_{k-1} \):

\[
(X_{k-1}|Z_{k-1}) \sim N(\hat{X}_{k-1}, P_{k-1}),
\]

(2.16)

where \( \hat{X}_{k-1} \) and \( P_{k-1} \) are the expectation and the variance of \( (X_{k-1}|Z_{k-1}) \). Hence, the above equation is the posterior distribution of \( X_{k-1} \). It is worth mentioning that the recursive procedure is started off at time 0 by choosing \( \hat{X}_0 \) and \( P_0 \).

At time \( k \), prior to observing \( Z_k \), \( X_k \) is governed by equation (2.14), where \( X_k = F_k X_{k-1} + w_k \). The expectation of \( X_k \) is therefore

\[
\mathbb{E}(X_k) = \mathbb{E}(F_k X_{k-1} + w_k) = \mathbb{E}(F_k \hat{X}_{k-1}) + \mathbb{E}(w_k) = F_k \hat{X}_{k-1} \quad \text{(by (2.16))}
\]
and the variance is

\[ P(k|k-1) = \text{Var}(X_k) = \text{Var}(F_k X_{k-1} + w_k) = \text{Var}(F_k X_{k-1}) + \text{Var}(w_t) \quad \text{(by independence)} \]

\[ = F_k \text{Var}(X_{k-1}) F_k^T + Q(k) = F_k P_{k-1} F_k^T + Q(k). \quad \text{(by (2.16))} \]

Hence, the prior distribution can be written as:

\[ (X_k|Z_{k-1}) \sim N(F_k \hat{X}_{k-1}, P(k|k-1) = F_k P_{k-1} F_k^T + Q(k)). \quad (2.17) \]

On observing \( Z_k \), the next step is to compute the posterior of \( X_k \) using (2.15). The prior is already known in the form of (2.17). The following is introduced as a tool to help finding the posterior:

\[ e_k = Z_k - \hat{Z}_k = Z_k - M_k F_k \hat{X}_{k-1}. \]

Because \( M_k, F_k, \) and \( \hat{X}_{k-1} \) are all known, observing \( Z_k \) is equivalent to observing \( e_k \). Now, (2.15) can be rewritten as

\[ \Pr(X_k|Z_k, Z_{k-1}) = \Pr(X_k|e_k, Z_{k-1}) \propto \Pr(e_k|X_k, Z_{k-1}) \times \Pr(X_k|Z_{k-1}), \]

with \( \Pr(e_k|X_k t, Z_{k-1}) \) is the likelihood.

Using (2.4), \( e_k \) can be written as\( e_k = M_k (X_k - F_k \hat{X}_{k-1}) + V(k) \) which means that \( \mathbb{E}(e_k|X_k, Z_{k-1}) = M_k (X_k - F_k \hat{X}_{k-1}) \). Because it is known that \( V(k) \sim N(0, R(k)) \), the likelihood is in the form of:

\[ (e_k|X_k, Z_{k-1}) \sim N(M_k (X_k - F_k \hat{X}_{k-1}), R(k)). \quad (2.18) \]

A well-known result in multivariate statistics as provided in [3] and some standard properties of the normal distribution are helpful in obtaining the posterior. They are the following:
Let \( \Omega_1 \) and \( \Omega_2 \) have a bivariate normal distribution with means \( \mu_1 \) and \( \mu_2 \), respectively, and a covariance matrix
\[
\begin{pmatrix}
\Sigma_{11} & \Sigma_{12} \\
\Sigma_{21} & \Sigma_{22}
\end{pmatrix}
\]
This can be denoted as:
\[
\begin{pmatrix}
\Omega_1 \\
\Omega_2
\end{pmatrix}
\sim
\mathcal{N}
\begin{pmatrix}
\begin{pmatrix}
\mu_1 \\
\mu_2
\end{pmatrix}
, \\
\begin{pmatrix}
\Sigma_{11} & \Sigma_{12} \\
\Sigma_{21} & \Sigma_{22}
\end{pmatrix}
\end{pmatrix}
\tag{2.19}
\]
When (2.19) holds, the conditional distribution of \( \Omega_1 \) given \( \Omega_2 = \omega_2 \) is described by \[3, 14\]:
\[
(\Omega_1 | \Omega_2 = \omega_2) \sim \mathcal{N}\left(\mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (\omega_2 - \mu_2), \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}\right)
\tag{2.20}
\]
The converse of (2.20) will be used. Whenever (2.20) holds and when \( \Omega_2 \sim \mathcal{N}(\mu_2, \Sigma_{22}) \), then (2.19) holds \[3, 14\].

Now, let \( \Omega_1 \) correspond to \( e_k \), \( \Omega_2 \) correspond to \( X_k \), and the conditioning variable \( Z_{k-1} \) be suppressed. Because \( (X_k | Z_{k-1}) \sim \mathcal{N}(F_k \hat{X}_{k-1}, P(k|k-1)) \) by (2.17), it implies that \( \mu_2 = F_k \hat{X}_{k-1} \) and \( \Sigma_{22} = P(k|k-1) \). Inputting all this information back to (2.20) and using (2.18), the following is obtained:
\[
\mu_1 + \Sigma_{12} P(k|k-1)^{-1}(X_k - F_k \hat{X}_{k-1}) = M_k (X_k - F_k \hat{X}_{k-1})
\]
which implies \( \mu_1 = 0 \) and \( \Sigma_{12} = M_k P(k|k-1) \). Similarly,
\[
\Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} = \Sigma_{11} - M_k P(k|k-1) M_k^T = R(k)
\]
which implies \( \Sigma_{11} = R(k) + M_k P(k|k-1) M_k^T \).

By (2.17), \( \Omega_2 \) is normal with mean \( F_k \hat{X}_{k-1} \) and variance \( P(k|k-1) \). Hence, the aforementioned converse relation can be applied. Therefore, the joint
distribution of $X_k$ and $e_k$ given $Z_{k-1}$ can be described as (the order of $\Omega_1 = e_k$ and $\Omega_2 = X_k$ is intentionally flipped for the further step):

$$
\begin{bmatrix}
X_k \\
e_k
\end{bmatrix} | Z_{k-1} \sim N \left( \begin{bmatrix} F_k \hat{X}_{k-1} & P(k|k-1)M_k^T \\ M_kP(k|k-1) & R(k) + M_kP(k|k-1)M_k^T \end{bmatrix} \right) .
$$

Using (2.21), (2.20) is applied one more time with $e_k$ as the conditioning variable. Hence, the following is obtained:

$$
\begin{bmatrix} X_k | e_k, Z_{k-1} \end{bmatrix} \sim N \left[ F_k \hat{X}_{k-1} + P(k|k-1)M_k^T (R(k) + M_kP(k|k-1)M_k^T)^{-1} e_k, P(k|k-1) - P(k|k-1)M_k^T (R(k) + M_kP(k|k-1)M_k^T)^{-1} M_kP(k|k-1) \right].
$$

(2.22)

The formulae (2.22) above is the desired posterior distribution.

Now, recall the Kalman Gain Formulae as presented in equation (2.13), where $K(k) = P(k|k-1)M_k^T (R(k) + M_kP(k|k-1)M_k^T)^{-1}$. Using (2.22), the definition of $e_k$, and the definition of the aforementioned Kalman Gain, the mean of $X_k$, denoted as $\hat{X}_k$, is just:

$$
\hat{X}_t = F_k \hat{X}_{k-1} + P(k|k-1)M_k^T (R(k) + M_kP(k|k-1)M_k^T)^{-1} e_k
$$

(by definition of Kalman Gain)

$$
= F_k \hat{X}_{k-1} + K(k)e_k
$$

(by definition of $e_k$)  \hspace{1cm} (2.23)

and the variance is:

$$
P(k|k) = P(k|k-1) - P(k|k-1)M_k^T (R(k) + M_kP(k|k-1)M_k^T)^{-1} M_kP(k|k-1)
$$

(by definition of Kalman Gain)

$$
= (I - K(k)M_k)P(k|k-1).
$$

(2.24)
One may notice that both (2.23) and (2.24) are none other than the well-known Kalman Filter Formulae.
CHAPTER III
EXPERIMENTAL SETUP

The experimental setup that is used in this project is the same as the one used in [8]. It is important to keep in mind that the goal of the project is to estimate the permeability field.

3.1 Problem Description

The experimental field from this project is a (synthetic) two-dimensional square petroleum reservoir with a size of 700(m)×700(m). This area is divided into uniform 21×21 grid cells as shown in Figure 3.1. The reservoir is equipped with an injector in the middle of the field and four producers in each of the corners. A fluid (typically water) is injected into the reservoir through the injector well, and oil (and maybe also water after some time) is pumped out from the four producers. This is known as the two phase flow in which two fluids are distinguished: water and oil.

3.1.1 Two Phase Flow

One basic assumption of this experimental field is that the reservoir is considered as a closed space; where liquid gets in and out only through the
3.1 : PROBLEM DESCRIPTION

Figure 3.1: Reservoir Field

The whole reservoir is divided into $21 \times 21$ grid cells and hence there are a total of 441 grid cells. Each grid cell has its own grid cell pressure $p$ and
saturation $S$ which are considered to be time-dependent. The state vector of the system is defined as the vector containing grid block pressure and saturation of all the grid cells at some time $t$, hence:

$$X(t) = \begin{pmatrix} p(t) \\ S(t) \end{pmatrix}. \quad (3.1)$$

In this experimental setup, observations can only be performed on the five drilled wells. From the injector well, bottom hole pressure, denoted as $bhp$, is measured and from the four producers, total flow rates, denoted as $q$, are measured. The measured total flow rates are the summation of water and oil flow rates. In reservoir simulation, $bhp$ and $q$ are often referred as production data. The vector of measurable variables for each time $t$ is defined as:

$$Z(t) = \begin{pmatrix} bhp(t) \\ q(t) \end{pmatrix}. \quad (3.2)$$

As mentioned earlier, the purpose of the project is to estimate the permeability field of the reservoir. The permeability parameter is denoted by $k$. Generally, the value of permeability is very small. One often finds it to be around $10^{-13}$ for example. For this reason, permeability is presented in the form of its natural logarithm in this experimental setup.

Adding the measurable variables and the log permeability to the state vector in (3.1), the state vector is renewed to be in the following form:

$$X(t) = \begin{pmatrix} \log k(t) \\ p(t) \\ S(t) \\ bhp(t) \\ q(t) \end{pmatrix}. \quad (3.3)$$
The variable $\log k(t)$ corresponds to the log permeability of each grid cell. Hence, the size of the state vector is $1328 \times 1$.

### 3.2 The Twin Experiment

The simulation performed in this project is based on a twin experiment. A twin experiment is a synthetic experiment where the truth is synthesized, hence it is already "known". With this synthesized truth, the performance of the method(s) may be tested and observed. For the upcoming simulation, the true permeability field is chosen to be the following, presented in logarithmic scale in Figure 3.2:

**Figure 3.2:** The chosen "true" permeability field.
In this project, measurements are conducted in five of the wells once every 60 days for a period of 480 days. This frequency of measurements is reasonable to be performed in real life application. Hence, after 480 days, eight sets of measurements are already produced.

Both the Ensemble Kalman Filter (EnKF) and the Non Parametric Bayesian Networks (NPBN) require an initial guess for the permeability field. This project works with 900 samples (ensemble members); thus, both methods also require 900 ensemble members as an initial guess. For the comparison between the two methods to be fair, of course both methods shall start with the same initial guess. The mean permeability value of the 900 ensembles for the initial guess is presented in Figure 3.3 below.

**Figure 3.3:** Mean of 900 log permeability chosen as the initial guess.
Typically, a reservoir is in an equilibrium state at the starting time. Therefore, the initial pressure and water saturation are assumed to be known and equal in all grid cells. For this simulation, the starting point for pressure and the grid block saturation are, respectively, taken to be $3 \times 10^7 \text{[Pa]}$ and 0.2.

One may question the decision to take 900 ensemble members as it is unusually high for EnKF, where one normally works with only 50 - 100 ensemble members. The reason behind this decision is that in the simulation performed in this project, a large number of variables (55 or 103 variables, depending on the case) in the system will be encountered. For the NPBN method, taking (only) 50 - 100 ensemble members means that the system is not well-represented, which may lead into inaccuracy in the outcome. For this reason, 900 ensemble members are taken into account.

Another requirement of both methods are the measurements. However, measurements generation will be discussed later in the upcoming chapter.

### 3.2.1 Arcs Directionality in the NPBN

To build the directed acyclic graph of the NPBN, the flow of influence between variables of the state vector at different locations must be determined. The choice for the directionality of arcs are explained and discussed in detail in [8]. In short, they are determined as the following:

One variable that influences all other variables is the permeability, where the directionality between permeability and grid block pressure is given.

\[^3\text{For the sake of simplicity, the term "permeability" will be used when the "log-permeability" is actually meant.}\]
by the physical interpretation of the Darcy’s Law. Saturation is excluded from the NPBN model because theoretically saturation has an almost constant value at a given time. It is valued around 0.2 before the water breaks through and 0.8 after the water breakthrough. Water breakthrough is the moment where water starts to also sip out from the production wells. Any constant is independent of any other variables; thus saturation is excluded. This is good news as it means that fewer variables are involved in the system. Nevertheless, valuable information might be missed. Therefore, inclusion of saturation will also be researched and discussed in more details during the case study.

**Figure 3.4**: Choices of the Arcs Directionality.

A modification of grid block pressure has a direct impact on both the bottom hole pressure and the total flow rate. Therefore, the arcs in NPBN are oriented from the grid block pressures to the bottom hole pressure and total flow rates. Regarding the arcs between bottom hole pressure and total flow rates, experiments performed in [8] suggested that there is no major difference between estimated permeability between the two possibilities (of arcs direction). Hence, it is arbitrarily chosen that the arc is directed from bottom hole pressure to total flow rates.

The NPBN representing all these choices is provided in Figure 3.4.
3.3 Measure of Performance

There are various ways to measure the performance of an experiment. The one chosen in this project is the root mean square error (RMSE). The RMSE is computed in every time step as follows:

\[
RMSE(t) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (k_{i}^{true} - k_{i}(t))^{2}}, \quad t = 1, \ldots, T
\]  

(3.4)

where \( k^{true} \) denotes the true permeability (time-invariant), \( k \) denotes the estimated permeability obtained from either the EnKF method or the NPBN method, \( t \) denotes the time step, and \( n \) denotes the number of grid blocks considered in the simulation. The RMSE shall be calculated in every time step and hence the performance of the corresponding method over time may be observed. In fact, the result obtained from calculating RMSE of both methods will be used when comparing the performance of the two methods.
As mentioned in the beginning of this thesis report, the project carried out in this thesis will use measurements that are generated with non-Gaussian noise. In the previous study as presented in [8], the measurements’ noise is assumed to be Gaussian with zero mean and 5% variance. One of the reasons behind this assumption is the unavailability of knowledge regarding the actual noise of the measurements. It, however, implies that actually one has the freedom to take any other noise from different distributions.

4.1 Measurements with non-Gaussian Noise

As presented in Chapter III, there are five variables that are measured in the underlying experiment of this project. They are the bottom hole pressure in the middle of the reservoir field and four total flow rates in the four corners of the field. Measurements of these variables are performed once every sixty days.

For all case studies performed in this project, the distribution of all those five measured variables is taken to be mixture of two Gaussian distributions with the following parameters:
Table 4.1: Parameters of the mixture of Gaussian distributions.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Mean</th>
<th>Variance</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bhp</td>
<td>5.5492e+05</td>
<td>5.5696e+11</td>
<td>0.3936</td>
</tr>
<tr>
<td>Q1</td>
<td>3.7235e-05</td>
<td>2.2492e-08</td>
<td>0.4105</td>
</tr>
<tr>
<td>Q2</td>
<td>-1.5162e-05</td>
<td>2.2345e-08</td>
<td>0.6069</td>
</tr>
<tr>
<td>Q3</td>
<td>2.2740e-08</td>
<td>1.5587e-08</td>
<td>0.6408</td>
</tr>
<tr>
<td>Q4</td>
<td>2.8644e-05</td>
<td>2.1042e-08</td>
<td>0.5899</td>
</tr>
<tr>
<td>Gaussian 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bhp</td>
<td>-3.6018e+05</td>
<td>9.3507e+10</td>
<td>0.6064</td>
</tr>
<tr>
<td>Q1</td>
<td>-2.5930e-05</td>
<td>2.3588e-09</td>
<td>0.5895</td>
</tr>
<tr>
<td>Q2</td>
<td>2.3415e-05</td>
<td>2.6365e-09</td>
<td>0.3931</td>
</tr>
<tr>
<td>Q3</td>
<td>-4.0567e-08</td>
<td>1.7812e-09</td>
<td>0.3592</td>
</tr>
<tr>
<td>Q4</td>
<td>-4.1206e-05</td>
<td>2.3183e-09</td>
<td>0.4101</td>
</tr>
</tbody>
</table>

An introduction to mixture of Gaussian distribution is available in Appendix A. The reasoning and process of obtaining those five distributions as presented in Table 4.1 above are presented in more details in Appendix B of this thesis. This makes the subject of the preliminary internship project of this thesis.

With these new noises introduced to the measurements, it turns out that the variance of the bottom hole pressure is 2.13%, the variance of total flow rate at location 1 is 20.21%, the variance of total flow rate at location 2 is 24.23%, the variance of total flow rate at location 3 is 21.44%, and the variance of total flow rate at location 4 is 25.19%. Hence, it means that taking the noise to be normally
distributed with mean zero and variance 5%, like previously performed in [8], is overestimating the noise of bottom hole pressure and underestimating the noise of all four total flow rates.

This higher variance of the error of total flow rates, however, triggers higher chance to obtain bad measurements. Here, bad measurements are unreliable or unrealistic measurements that lay (really) far from the truth. These measurements may "mislead" both the EnKF and the NPBN by giving unreliable information. Hence, these measurements may affect the performance of both methods.

EnKF has a factor called Kalman Gain (given in equation (2.13)) which shall damp directly the effect of bad measurements in this context. However, this factor does not distinguish the occurrence of bad or good measurements. If the observed variable has a higher variance, the Kalman Gain tells EnKF to trust the measurement values of that particular variable less; regardless of the actual accuracy of each particular measurement values. Hence, when bad measurements occur, this is good news for EnKF. However, when good measurements are available, EnKF does not make use of that good information much in its algorithm. This may slow down the speed of the recovery of permeability value by EnKF. A Kalman-Gain-like factor is actually also implied in the NPBN method. However, instead of working with the actual values of the state vectors' variables, this factor works in the transformed space\(^4\) of these values. Hence, the magnitude of the effect of this factor is likely to be different.

\(^{4}\text{The transformation is explained in detail in Chapter 3, section 3.1.2 (p. 34 - 35) of [8]}\)
than the Kalman Gain in EnKF. By performing the conditioning operations with the NPBN’s networks, one actually implies the use of this factor, even though it may not be explicit. More insight into this implied factor can be found in section 2.3 of this thesis or in [14].

This phenomenon and a comparison between both the EnKF and the NPBN will be observed and discussed further in this chapter.

### 4.1.1 The Generated Measurements

Figure 4.1 below shows the measurements generated with the mixture of Gaussians noise with parameters as in Table 4.1 above in comparison to the true values. The blue stars represent the measurements and the connected black circles represent the truths.

**Figure 4.1:** The generated measurements with mixture of Gaussian noise from the true observation used for the experiment in this project
Note: the actual total flow rates that are used in the experiment are negatively valued. The negativity corresponds to the fact that water and oil are flowing out from the reservoir (hence the flow is "negative") in the production wells. Figure 4.1 above shows the absolute value of these total flow rates.

As mentioned earlier, the chance of bad measurements to occur (with this noise) is higher than in [8]. One bad measurement (of one of the five variables) might occur in one time step but its effect might be damped by the other measurements at the same time step if these other measurements are good enough (closer to their respective true values).

A bad situation would be if all (or almost all) of the measurements at one time step are located relatively far away from the truth. Sometimes in this case when the measurement is located too far away from the truth, truncation is needed. In the software UNINET, the software where the conditioning step of the NPBN Method is performed in, one cannot condition on one value that is located outside of the sample range of the particular variable. When facing this situation, the author manually truncates the value of the measurement to the closer end of the sample range.

Unfortunately, as visible in Figure 4.1 above, this kind of situation occurs with this set of measurements. At time step 2 (day 120), even though the measurement of \( Q_1 \) seems to be really good, the other four are pretty far off. In fact, the measurement of \( Q_4 \) had to be manually truncated because it originally fell outside of the sample range of total flow rate at this location. Another time step that might be of concern is time step 8 (day 480) where, even though
manual truncation is not needed in any of the five variables, all five measurements lie relatively far from the true values. The other time steps appear to be seemingly fine.

Hence, these bad measurements at time step 2 and time step 8 might affect the performance of both the EnKF and the NPBN method. In fact, this will be explored further in this chapter with the case studies.

One might guess that the emergence of these bad measurements is just a bad luck. However, it is seemingly not. In many trials that the author performed, bad measurements always emerge in each of the generated measurements set. Section B.4 in Appendix B of this thesis report presents more insight into this.

4.2 The Performance of the NPBN Method

Case study is going to be performed with a new set of measurements that has been built in the previous section. The purpose of this case study is to compare the performance of EnKF versus the NPBN based approach. Results presented in [8] indicate that the NPBN method generally works well in a medium-sized grid block \((7 \times 7)\). For a larger grid block \((13 \times 13)\), even though visual analysis looks promising, RMSE analysis contradicts that. For this reason, the case study performed in this thesis will consider only medium-sized grid blocks \((7 \times 7)\) or smaller.

Even though the NPBN method only updates part of the reservoir, it does not affect the accuracy of the estimation. To estimate the entire reservoir, as
proposed in [8], one may perform NPBN method in several parts of the reservoir and combine the results together with an interpolation. This will be a topic discussed further in Chapter V of this thesis.

Three experiments investigating three different grid blocks across the reservoir are presented in this section. Based on the results, comparisons between EnKF and the NPBN based approach will be performed. The NPBN approach that is performed in this project is the saturated NPBN approach. For convenience, "saturated NPBN method" and "NPBN method" will be used interchangeably in this thesis.

Figure 4.2: The location of the chosen medium-sized (7 × 7) grid block in the reservoir
4.2.1 Permeabilities Estimation for A $7 \times 7$ Grid Block

This first experiment uses the same $7 \times 7$ grid block as in [8]. The location of the chosen grid in the reservoir is presented in Figure 4.2.

To estimate the 49 permeabilities of this $7 \times 7$ grid block, 103 variables are involved. Hence, for NPBN approach, a saturated graph with 103 nodes will be built for every time step for a total period of 480 days. A saturated NPBN with 103 nodes has 5,253 arcs, therefore it is practically impossible to visualize all the nodes and arcs. An example of an NPBN with these characteristics that is used in this experiment is presented in Figure 4.3.

**Figure 4.3:** A saturated NPBN with 103 variables for the $7 \times 7$ grid block

The experiment is performed following these steps: First, the simulator is run for one time step. Then, a saturated NPBN, which looks like the one in Figure 4.3 above, is built using the information regarding the permeability, pressure, bottom hole pressure, and total flow rate that is...
provides. The arcs directionality in the NPBN follows the description as in section 3.2.1 in Chapter 3. Then, the assimilation step is performed. The assimilation step translates in conditionalizing on five of the measurable variables (bottom hole pressure and four total flow rates) using the available measurements at day 60. Post-conditionalizing, the NPBN is sampled and this set of samples will be used as the starting point for the next time step. Using this new set of samples as the initial, \texttt{sim} is run for another 60 days (so day 120 is reached) where other measurements are available; and the same procedure is repeated all over again. This procedure is repeated until the entire 480 days period is completed, therefore there are eight time steps in the experiment.

The samples sets of conditionalized NPBN contain the NPBN method’s approximation of the permeability. Because there are 900 samples (of permeability) for each grid cell, taking the mean value of these 900 samples yields the approximation of permeability value of that particular grid cell obtained from executing the NPBN method. Then, these approximations are compared with the true value by means of Root Mean Square Error (RMSE).

Following the described procedure above, the obtained result for the particular $7 \times 7$ grid block is presented in Figure 4.4. Figure 4.4 presents the true value of permeability of this $7 \times 7$ grid, the mean of initial ensemble members, and the estimated permeability coming from both EnKF and Saturated NPBN method at time step 7 (day 420) and time step 8 (day 480). The RMSE of the two approaches are presented in Figure 4.5.
Figure 4.4: The truth (top left), mean of initial ensembles (top right), EnKF estimation (bottom left), and NPBN estimation of permeability (bottom right) of the $7 \times 7$ grid block.

Figure 4.5: RMSE of the estimated permeability field of the $7 \times 7$ grid block using EnKF and saturated NPBN.
It might be worth reminding that the variances of the measurements used in this experiment are approximately 2% for bottom hole pressure and above 20% for the four total flow rates.

As visible in Figure 4.5, the RMSE of EnKF decreases steadily and slowly with almost a constant behavior. This slowness and steadiness of EnKF is likely to be caused by the less trust EnKF has on the measurements due to the higher variances in measurements. This is good for EnKF to damp the effect of bad measurements; but at the same time it also causes EnKF to be slower in recovering the actual permeability field due to "less" trusted extra information. The RMSE of Saturated NPBN has a very good decrease between time step 2 (day 120) and time step 7 (day 420), but increases at time step 2 and time step 8. The cause of this increasing behavior is suspected to be the bad measurements, that have been introduced earlier in this chapter, which appear at these two time steps. EnKF actually also picks up these bad measurements at these two time steps as can be seen by the increasing behavior of the RMSE at these two occasions. However, both increases are not as dramatic as in NPBN probably due to the less trust EnKF has on the measurements; hence damping the "bad effect" (increasing RMSE) because of these bad measurements.

The NPBN method seems to be more sensitive to poor measurements than EnKF is. However, it also compensates this sensitivity with a much faster recovery rate than EnKF. As can be seen between time step 2 (day 120) and time step 7 (day 420) on Figure 4.5, NPBN outperforms EnKF until other bad measurements occur at time step 8 (day 480). As mentioned in the previous
chapter, NPBN based approach also implies the use of Kalman Gain-like factor in it. However, not like in EnKF, this factor works in the transformed space of the state vectors’ values instead of working on the actual values [14]. Hence, the effect’s magnitude of this factor is different between the two methods. The fact that the Kalman-Gain-like factor of NPBN method works in the transformed space appears to cause the method to trust the measurements more than EnKF; as indicated by the more sensitive behavior observed in the RMSE of NPBN method than in EnKF.

Visual comparison using Figure 4.4 indicates that the last two steps of EnKF have very similar estimation of permeability, explaining the almost constant value of the RMSE of EnKF at these two steps. The NPBN estimation of permeability at time step 7 (day 420) seems to approximate the truth better than EnKF. This is also supported by the RMSE. At time step 8 (day 480), more apparent changes are visible in the NPBN estimation of permeability than in EnKF. However, visual inspection might conclude that this NPBN estimation is still slightly better than EnKF, even though the RMSE values disagree with this visual conclusion.

It is, however, worth stressing that RMSE shows only an average behavior of an estimated field. Hence, it can often be misleading.

At this point, one may wonder about the effect of adding non-Gaussian noise to the measurements rather than adding Gaussian noise. To observe the effect, a comparison is made between the RMSE values presented in Figure 4.5 and the RMSE value of the same $7 \times 7$ grid block case where Gaussian noise is
used as the measurements noise, as presented in [8]. The two RMSE figures are presented next to each other in Figure 4.6.

Comparison based on Figure 4.6 below implies that NPBN based approach benefits from this generated set of noise. When bad measurements do not occur, NPBN with non-Gaussian measurements noise recovers the truth faster than NPBN with Gaussian noise. This can be observed by comparing the slopes of the RMSE curves in time. The slope of the RMSE of the non-Gaussian case between time step 2 and time step 5 is steeper compared to any drops in the Gaussian case.

**Figure 4.6:** Comparison of the effect of adding non-Gaussian noise (figure (a)) instead of Gaussian noise (figure (b)) to the measurements in term of RMSE of each method.

(a) RMSE of EnKF versus NPBN with non-Gaussian noise

(b) RMSE of EnKF versus NPBN with Gaussian noise

As for EnKF, the observable difference between the two cases is that EnKF is more steady and stable in the case of non-Gaussian noise. The cause of this phenomenon is the Kalman-Gain factor. In the non-Gaussian case, the variance of the noise is (generally) higher than in the Gaussian case. Hence, as
mentioned before, EnKF trusts the measurements of the non-Gaussian case much less, causing the steadier behavior of the RMSE.

On the other hand, EnKF does not really seem to benefit from the change of distributions of the measurements noise. The values of the RMSE of both cases are also similar to each other. This makes sense because EnKF does not recognize the non-Gaussian distributions of the noise. One basic assumption of EnKF is that the measurements noise is Gaussian with mean zero and variance $R(k)$. Hence, when the new set of measurements is introduced to EnKF, EnKF treats it as normally-distributed; just like it does in the Gaussian case presented in Figure 4.6.b. Therefore, there is no difference in how EnKF treats the two sets of measurements aside from the fact that they have different variances. As mentioned earlier, the different choice of variances only causes the change in the steadiness of the truth recovery. This explains why EnKF does not seem to benefit from the change of measurements noise’s distributions.

### 4.2.2 Permeabilities Estimation for A $5 \times 5$ Grid Block

This second experiment focuses on another grid block. Now, a smaller grid block of size $5 \times 5$ is considered. The chosen $5 \times 5$ grid block is located much closer to measurement with less variation, which is the bottom hole pressure (bhp). Its location implies that this grid block is also located further away from the locations with more volatile measurements. The location of the chosen grid in the reservoir is presented in Figure 4.7.

Intuitively, performing assimilation (with either methods) on this grid
**Figure 4.7:** The location of the chosen $5 \times 5$ grid block, which is within proximity of measurement with less variation (bhp), in the reservoir.

This will be investigated.

For this $5 \times 5$ case, there are 25 cells whose permeabilities are to be estimated. In this case, 55 variables are involved. Therefore, a saturated non-parametric Bayesian Networks with 55 nodes shall be built for every time step. A saturated NPBN with 55 nodes has 1,485 arcs; hence it is still impossible to visualize all the nodes and arcs.

Following the same procedure as in the $7 \times 7$ case, the results for this particular $5 \times 5$ grid block closer to bhp are presented in Figure 4.8 and 4.9.
**Figure 4.8:** The truth (top left), mean of initial ensembles (top right), EnKF estimation (bottom left), and NPBN estimation of permeability (bottom right) of the $5 \times 5$ grid block close to bhp.

**Figure 4.9:** RMSE of the estimated permeability field of the $5 \times 5$ grid block close to bhp using EnKF and saturated NPBN.
Intuitively, both methods should benefit from the position of the grid block; however, Figure 4.9 suggests otherwise. In both methods, the RMSE is above 0.6 at all time steps; and this is no better than the RMSE of the previous case as presented in Figure 4.5. The RMSE of the saturated NPBN method decreases steadily over time with exception at time step 2 (day 120) and time step 8 (day 480) where bad measurements are present. However, both increases are negligible compared to the one in Figure 4.5. The RMSE of EnKF, on the other hand, increases at (almost) every time step. Not only EnKF does not seem to appreciate the position of the grid block which is within proximity of good measurement, it also performs worse than in the previous case.

However, visual observation is somewhat disappointing. The recovered permeability field from EnKF is very smooth, unlike the truth. It also fails to recover both the very permeable and very impermeable areas. However, in general, EnKF still performs better than NPBN in approximating the permeable grid cells. NPBN, on the other hand, performs much better in recovering impermeable areas and it estimates the very impermeable grid cells more accurately. In fact, NPBN overestimates the impermeability of the field.

It is quite puzzling to observe that for a similar value of RMSE between the two grid blocks (7 × 7 and 5 × 5), the estimated field could look so different to each other and to the truth. There is no visual correspondence between a 0.7 RMSE for the NPBN approximation in the 7 × 7 grid at time step 8 (day 480) with a 0.7 RMSE for the NPBN approximation in this 5 × 5 grid at time step 7 (day 420). The estimated field by NPBN in the 7 × 7 case at time step 8
approximates the truth quite well while the estimated field by NPBN in the 5 × 5 case at time step 7 does not. This raises concern about the use of RMSE as a measure of performance.

4.2.3 Overlapping Grid Block and Consistency of NPBN

Now, a third grid block is chosen for the third experiment. It is another 5 × 5 grid block; but this grid block has a nice feature that some of the cells are overlapping with the previous two blocks. The chosen grid block is presented in Figure 4.10 below.

Figure 4.10: The location of the chosen 5 × 5 grid block with overlapping cells in the reservoir
This choice of grid block makes the comparison of the NPBN estimates of permeability for the overlapping grid cells between the two experiments possible. It has been mentioned several times that based on the findings thus far, the NPBN method works well in a medium-sized grid block or smaller (though the size might depend on the number of available samples in a simulation). In reality, one might encounter a large field with a much larger-sized grid block. One recommendation from [8] is for one to work with separate several smaller grid blocks across the entire field instead. Then, the result for the entire field is just the results of these smaller partial grid blocks combined together. To combine these smaller partial grid blocks, one may employ an interpolation method to estimate the permeability values at the cells between the two grid blocks.

However, before proceeding that far, "consistency" of NPBN method should be investigated first. Here, "consistency" refers to the consistency of results between one experiment and another. One way to test the consistency is by observing the results of two (or more) experiments with overlapping grid cell(s). The NPBN method is consistent if the results for the overlapping grid cell(s) between the experiments agree with each other. Of course the results between different experiments are not expected to be exactly the same; but if they exhibit similar behavior, the method can be seen as being consistent. For an example from the project undergone in this report, an inconsistent behavior would be that if for one particular overlapping cell, one experiment produces a (very) impermeable result while the other produces (very) permeable result.
**Figure 4.11**: The truth (top left), mean of initial ensembles (top right), EnKF estimation (bottom left), and NPBN estimation of permeability (bottom right) of the overlapping $5 \times 5$ grid block.

**Figure 4.12**: RMSE of the estimated permeability field of the overlapping $5 \times 5$ grid block using EnKF and saturated NPBN.
As can be seen in Figure 4.10, eight of the cells in the third grid block are overlapping with either one of the previous two grid blocks: two of them with the first $7 \times 7$ case and six of them with the previous $5 \times 5$ case. Consistency of the NPBN method in this case of reservoir simulation may be examined by observing these eight cells and comparing the results between this third experiment and the corresponding previous experiment.

Following the same procedure as before, the results for this particular $5 \times 5$ grid block are obtained and presented in Figure 4.11 and Figure 4.12.

The RMSE seems to have the same characteristics as the previous $5 \times 5$ case. The RMSE of EnKF seems to have a tendency to increase over time, even though the value itself is much better than the previous case as in Figure 4.9. The RMSE of the NPBN method decreases steadily between time step 2 (day 120) and time step 5 (day 300) before going steadily up. This increasing behavior of the RMSE of NPBN at time step 6 and 7, which is a different behavior from the previous two cases, might be caused by the proximity of this grid block to the total flow rate 4 ($Q_4$) measurement. As can be seen in Figure 4.1, the measurement value of $Q_4$ at time steps 6 and 7 are relatively off (in term of percentage of the true value) from the truth. The cause of increases at time step 2 (day 120) and time step 8 (day 480) is suspected to be the same as before, bad measurements that are present at these two time steps. These increases of RMSE at these two time steps are more apparent compared to the previous $5 \times 5$ case in Figure 4.9 but definitely less apparent compared to the first $7 \times 7$ case in Figure 4.5. This may be caused by the fact that the location of this third grid block
is closer to bottom hole pressure measurement, which is a good measurement, compared to the first $7 \times 7$ grid block; but its proximity to bhpp is still further than the previous $5 \times 5$ grid block. Therefore, it is likely that the performance of NPBN method is influenced by the location of the grid block in regard to its proximity to good or bad measurements. It may be also worth to mention that the RMSE of both methods at this third grid block is generally lower than the previous two experiments (except for the initial value where nothing has been done); indicating that all in all both methods perform better than previously.

The top left plot of Figure 4.11 suggests that the cells of this grid block typically have relatively more impermeable property. Visual observation over the recovered permeability of EnKF is, again, very smooth; unlike the truth. As a whole, it underestimates the impermeability of the cells in this grid block. As indicated by the RMSE, there seems to be not much difference between the EnKF approximation at time step 7 and time step 8. Even though not "superb", the NPBN method seems to approximate the truth better than EnKF. Part of the field approximation at time step 7 seems to be okay, with the other part underestimates the impermeability. Moving to time step 8, the general approximation becomes worse (as indicated also by the RMSE) and the underestimation becomes even stronger.

Considering the eight overlapping cells, it seems that NPBN method is quite consistent in recovering the permeability of each of the cells. The analysis performed in this thesis is based on simple visual observation with the help of Figure 4.13 and Table 4.2. With this approach, the result looks quite
Figure 4.13: Comparison of the eight overlapping cells at time steps 7 and 8

promising. The two cells that overlap with two cells from the first $7 \times 7$ grid block seem to have similar results from both experiments (at the same time step), stating that these cells are both impermeable. This is also supported by the values in Table 4.2 where the average and maximum absolute difference between the same corresponding cell are relatively not that high. Even though the permeability values are not exactly the same, the similarity of conclusion regarding the permeableness is a good sign. As for the six cells that overlap with six cells from the previous $5 \times 5$ case, they also appear alright, especially
Table 4.2: Some numbers to help the comparison of the overlapping cells.

<table>
<thead>
<tr>
<th>Case</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two Overlapping Cells</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Time Step 7</td>
</tr>
<tr>
<td>Average Difference</td>
<td>0.2769</td>
</tr>
<tr>
<td>Maximum Difference</td>
<td>0.3915</td>
</tr>
<tr>
<td>Six Overlapping Cells</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Time Step 7</td>
</tr>
<tr>
<td>Average Difference</td>
<td>0.4870</td>
</tr>
<tr>
<td>Maximum Difference</td>
<td>0.6485</td>
</tr>
</tbody>
</table>

the result at time step 7 (day 420). Even though the estimated permeability values between them might not be as similar as the other two overlapping cells, the difference does not seem to be that much either, as also indicated by Table 4.2. For time step 8 (day 480), however, the difference is much more apparent. This is very visible in Table 4.2 as well where the average difference is 1.0499 with maximum difference 1.6282. However, the consistency regarding which cell is more permeable than the other seems to be present. Hence, even though for this eighth time step the difference in permeability values might be more evident, a correspondence between the two experiments still looks to be existing.
4.3 Inclusion of Saturation

All experiments with NPBN approach that have been performed thus far, including those in [8], are under assumption that saturation can be excluded from the model because, theoretically, it has an almost constant value at a given time. However, as pointed in section 3.2.1, this may cause in loss of valuable information. Therefore, in this section, it will be investigated whether inclusion of saturation will improve the performance of the NPBN approach or not.

4.3.1 (Updated) Arcs Directionality in the NPBN

By including saturation to the model, one adds extra node to the acyclic graph of the NPBN. Before being able to include saturation to the model, one needs to find out the flow of influence between saturation and all other variables that are present in the system. The flow is represented by the direction of the arc in the acyclic graph. In other words, Figure 3.4 needs to be updated with an addition of a saturation node.

One basic property that permeability influences all other variables (including saturation), as described in Section 3.2.1, is still going to be used here. The important relation that needs to be investigated is the flow of influence between pressure \( p \) and saturation \( s \).

To characterize the flow of influence between pressure and saturation, knowledge about the nature of pressure and saturation equations is very helpful. The nature of the equations is presented in detail in [11] in page 24 – 26. According to this, the pressure equation is diffusive (elliptic) with a
4.3: INCLUSION OF SATURATION

coefficient that is mildly non-linear function of the saturation. On the other hand, the saturation equation is nearly hyperbolic with a coefficient that is a function of the spatial gradient of the pressure. Moreover, the hyperbolic character of saturation leads to the formation of fronts, such that the effects of a pressure gradient in a certain point may influence the saturation at a much later moment in that point.

Therefore, there is a two-way influential flow between the two. However, because the changes in saturations usually occur much slower than the pressure changes [11], the direct influence of saturation on pressure is weak. On the other hand, the direct influence of pressure on saturation is much stronger but delayed in time (due to the fronts formation). Therefore, the arc in the NPBN is taken to be oriented from the pressure node to the saturation node.

Figure 4.14: Choices of the Arcs Directionality (Updated).

Regarding the arc directionality between saturation and both the bottom hole pressure and total flow rates, the well model (also described in details in [11]) provides the desired relation. According to the model, both bottom hole pressure and total flow rates are non-linear functions of pressure and saturation ([11] page 61). Therefore, the arcs in NPBN are taken to be oriented from the saturation node to the bottom hole pressure and total flow rate.
CHAPTER IV : CASE STUDY

The updated NPBN with inclusion of saturation node is provided in Figure 4.14.

4.3.2 Permeabilities Estimation with Saturation

Using the updated arc directionality as described above, permeabilities estimation can now be performed. Firstly, it is worth mentioning that addition of saturation to the model introduces more dimension to be involved in the system. In a $7 \times 7$ grid block, without saturation, 103 variables are already involved. With saturation, another 49 variables are introduced to the system, making a total of 153 variables to work with. As already mentioned in [8], the NPBN method seems to be able to handle, in an accurate manner, only around 100 variables at a time. Therefore, the size of the grid block should be narrowed down if saturation is going to be included.

For this reason, the investigation performed in this section will be based only on smaller $5 \times 5$ grid blocks. With a $5 \times 5$ grid block, only 80 variables are involved in total; thus, still within NPBN reach based on the number of variables. Also, analysis will be based on the same two $5 \times 5$ grid blocks as used in Section 4.2.2 and 4.2.3, but with saturation included to the system. The upcoming results with saturation included in the model can, therefore, be compared to the previous results presented in Section 4.2 to see whether inclusion of saturation improves the performance of the model or not.

Following the same procedure and including saturation to the NPBN model, the results are the obtained as presented in Figure 4.15 - 4.18 below.
4.3: INCLUSION OF SATURATION

**Figure 4.15:** The truth (top left), mean of initial ensembles (top middle), EnKF estimation (bottom left), NPBN estimation without Saturation (bottom middle), NPBN estimation with Saturation (bottom right) of the $5 \times 5$ grid block close to bhp

**Figure 4.16:** RMSE of the estimated permeability field of the $5 \times 5$ grid block close to bhp using EnKF and saturated NPBN with and without saturation
**Figure 4.17:** The truth (top left), mean of initial ensembles (top middle), EnKF estimation (bottom left), NPBN estimation without Saturation (bottom middle), NPBN estimation with Saturation (bottom right) of the overlapping $5 \times 5$ grid block.

**Figure 4.18:** RMSE of the estimated permeability field of the overlapping $5 \times 5$ grid block using EnKF and saturated NPBN with and without saturation.
Looking at the RMSE values of both cases, there appears to be an agreement that introduction of saturation to the model does not really improve the performance of the NPBN method. This is visible from the almost coincident RMSE of the NPBN with saturation and without saturation (the blue lines in the RMSE figures) at both cases. In the end (day 480) for both cases, the NPBN method with saturation seems to perform marginally better than the NPBN method without saturation, but not that significant. Based on the RMSE, inclusion of saturation seems to damp the effect of bad measurements at time step 8 as can be visible by the not so steep increase of the RMSE (compared to the RMSE of the NPBN without saturation).

Visual observation upon the recovered permeability fields also exposes similar trait. The recovered permeability fields from NPBN with introduction of saturation do not appear to be so much different from the corresponding recovered permeability fields from NPBN without inclusion of saturation. In fact, they are very similar to each other. Again, one may notice that the NPBN with inclusion of saturation seems to recover the truth a little bit better than the NPBN without inclusion of saturation (especially at the more impermeable cells); but the difference does not seem to be that substantial.

Based on the two grid blocks above, an improvement is, to some degree, present after the inclusion of saturation to the NPBN method. However, the improvement does not appear to be prominent. This conclusion actually legitimates the earlier decision to exclude saturation from the model.
4.4 Choices for Further Experiment

Although inclusion of saturation does not really improve the performance of the NPBN method substantially (as presented in in Section 4.3), further experiment in this thesis project will still incorporate saturation in the NPBN method. Even though not eminent, improvement still does exist. The drawback of this decision is, of course, smaller usable grid blocks (smaller than $7 \times 7$) due to the increasing number of variables involved in the NPBN method. As a consequence, the experiment presented in Chapter V will work with $6 \times 6$ grid blocks. With this size, 113 variables are involved in the system.

However, it means that in the future one can actually "trade" saturation for something else to be included in the model (to maintain the involved dimension to not be too high; otherwise one might need to pick even smaller grid blocks, which may be inconvenient), if that something else contributes more (than saturation) to the improvement of the performance of the NPBN method.

One other idea for inclusion of saturation is that instead of taking the saturation at all cells of the entire block, alternatively one may just pick some cells in that block where saturation is going to be taken into account. So, if for a $7 \times 7$ block one only takes saturation at three of its cells, instead of adding another 49 variables to the system, one only introduces three extra variables. The premise behind this idea is the same one as why saturation is excluded in the preceding experiments. Saturation has almost a constant value, hence it may be sufficient to only pick some cells of the block from where saturation is going
to be represented. The first advantage of this (possible) action is that there is no need to "shrink" the size of the grid block(s) because the dimension expansion in the system is not that much. Secondly, one may introduce another node/factor to the system without having to completely "trade off" saturation nor shrinking the size of the grid block(s) massively.

However, which cells to pick and how to pick those cells definitely need further rigorous research.
CHAPTER V
INTRODUCTION OF INTERPOLATION

As has been mentioned several times previously, the NPBN method only updates part of the reservoir. To recover the permeability field of the entire reservoir, it has been proposed to perform the NPBN method several times at several non-overlapping grid blocks of the reservoir and then interpolate the outcomes.

This chapter is meant as an initial attempt to research this interpolation method. There are a lot of possible interpolation methods, including one that performs on a random field (kriging) [17]. However, as a starter, the methods that are going to be considered in this chapter are the simplest ones first: methods based on known locations and not taking into the account the ‘randomness’ of the field. Then, the best performing one is going to be applied to the underlying case of permeability estimation. Several grid blocks of the reservoir are going to be updated with NPBN method with incorporation of saturation (as in Chapter IV). Then, the outcomes will be combined with this interpolation method. The end result will be the permeability estimation of the entire field of the NPBN method. Along the way, analysis over the method is
5.1 Possible Methods of Interpolation

5.1.1 Experimental Setup

Before actually throwing in ideas and checking possible interpolation methods, the first question that has to be answered is "How do we perform the investigation of the performance of the possible interpolation methods?". In other words, the first thing to do is to define an experimental setup for the interpolation method. The following setup has been chosen:

Investigation over the performance of the possible interpolation methods is going to be performed on the "truth" which has been chosen in Chapter III (presented as Figure 3.2). With some "known" grid cells, the value of permeability in between them is going to be approximated. The interpolation method is going to be performed a number of times until all "interpolable" grid cells in the field are approximated. Further, because the underlying reservoir field is a square-shaped one, the investigation of one method will be performed twice: once when the interpolation is performed vertically and the other when it is performed horizontally. Also, each method will be tested in the case of any distance or width between the known grid cells. To analyze the performance of the methods, the usual Root Mean Square Error (RMSE) is going to be used, as well as visual comparison of the "recovered" permeability field and the truth.

To get more technical, the value of permeability at one cell will be viewed as a "function" of the location of the corresponding cell. This is so
because the only known information about the interpolated cell is that its position is somewhere in between the known grid cells. Hence, let $x_i$ be the position of the grid cell, with $i$ corresponds to the index of the cell, it is assumed that a function $f$ exists such that

$$k_i = f(x_i),$$

with $k_i$ corresponds to the value of permeability at grid cell $x_i$. The objective now is to find the function $f$ such that equation (5.1) holds based on the known data (or the deviation is minimized).

Now that the experimental setup has been defined, investigation of the possible interpolation methods can be performed.

5.1.2 Method 1 (Simplest Linear Interpolation)

The first idea in consideration is the simplest one: linear interpolation. Using two separated grid cells in the same column (row) with known permeability for vertical (horizontal) approximation, the permeability value of the cells between the two cells can be approximated by taking linear interpolation. An illustration for the vertical case with only one cell to be approximated is provided in Figure 5.1 below.

In the case where the distance between the two known cells is just one cell (hence, only one cell to be interpolated), like in Figure 5.1 below, the interpolated permeability value is simply just the average of the permeability of the two cells. But of course this method can be extended when wider distance between two known grid cells is present.
Figure 5.1: Illustration of the simplest case with one cell to be interpolated.

Table 5.1: RMSE values of the interpolated permeability using the first method with all possible distances (in cell) between "known" cells

<table>
<thead>
<tr>
<th>Distance</th>
<th>Vertical</th>
<th>Horizontal</th>
<th>Distance</th>
<th>Vertical</th>
<th>Horizontal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 cell</td>
<td>0.321855</td>
<td>0.220378</td>
<td>11 cell</td>
<td>0.823282</td>
<td>0.568084</td>
</tr>
<tr>
<td>2 cell</td>
<td>0.372335</td>
<td>0.264078</td>
<td>12 cell</td>
<td>1.163666</td>
<td>0.640079</td>
</tr>
<tr>
<td>3 cell</td>
<td>0.497231</td>
<td>0.304998</td>
<td>13 cell</td>
<td>1.210377</td>
<td>0.674130</td>
</tr>
<tr>
<td>4 cell</td>
<td>0.572559</td>
<td>0.358104</td>
<td>14 cell</td>
<td>1.144409</td>
<td>0.669463</td>
</tr>
<tr>
<td>5 cell</td>
<td>0.501070</td>
<td>0.381400</td>
<td>15 cell</td>
<td>1.048866</td>
<td>0.672340</td>
</tr>
<tr>
<td>6 cell</td>
<td>0.677740</td>
<td>0.470465</td>
<td>16 cell</td>
<td>0.955716</td>
<td>0.661221</td>
</tr>
<tr>
<td>7 cell</td>
<td>0.504603</td>
<td>0.504149</td>
<td>17 cell</td>
<td>0.911317</td>
<td>0.657777</td>
</tr>
<tr>
<td>8 cell</td>
<td>0.606678</td>
<td>0.474550</td>
<td>18 cell</td>
<td>0.922017</td>
<td>0.691398</td>
</tr>
<tr>
<td>9 cell</td>
<td>0.726254</td>
<td>0.504357</td>
<td>19 cell</td>
<td>0.989682</td>
<td>0.730876</td>
</tr>
<tr>
<td>10 cell</td>
<td>0.569667</td>
<td>0.571540</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Running this method for the "true" permeability field, the result in RMSE is summarized in Table 5.1 above. A few of the recovered permeability (with 1 to 3 cell distance) in comparison with the truth is presented in Figure 5.2 below.

It is not surprising to observe that the wider the distance is, the less
accurate the interpolated permeability field tends to be, as indicated in Table 5.1. The method seems to perform fine for the shorter distances (as supported by Figure 5.2), but this will, of course, be compared with the performance of the other methods. Back to Table 5.1, for this particular truth, there seems to be a tendency that for the same distance, the method performs better on horizontal approximation.
5.1.3 Method 2

When applying method 1, one only takes into consideration the boundary of the two grid blocks to interpolate what is (are) in between. One may argue that if more known cells are involved, the performance of the interpolation method may increase. Plus, with addition of known cells, higher degree interpolation can be performed, not limited as linear (one-degree polynomial) interpolation only. For this reason, method 1 can be "generalized" by taking the second row (column) of the two grid blocks for vertical (horizontal) interpolation; thus providing four known grid cells for one interpolation. Then, the procedure goes exactly the same as in method 1. Figure 5.3 below illustrates the vertical interpolation with only one cell to be approximated when method 2 is performed.

Figure 5.3: Illustration of method 2 with one cell to be interpolated.
Table 5.2: RMSE values of the interpolated permeability using the second method with all possible distances (in cell) and with all three possible degrees of interpolation

<table>
<thead>
<tr>
<th>Distance</th>
<th>Vertical</th>
<th>Horizontal</th>
<th>Distance</th>
<th>Vertical</th>
<th>Horizontal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Linear Interpolation (one-degree polynomial)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.383412</td>
<td>0.259907</td>
<td>10</td>
<td>0.995309</td>
<td>0.596614</td>
</tr>
<tr>
<td>2</td>
<td>0.456798</td>
<td>0.310937</td>
<td>11</td>
<td>1.194147</td>
<td>0.639883</td>
</tr>
<tr>
<td>3</td>
<td>0.461458</td>
<td>0.379361</td>
<td>12</td>
<td>1.172779</td>
<td>0.667459</td>
</tr>
<tr>
<td>4</td>
<td>0.506908</td>
<td>0.392643</td>
<td>13</td>
<td>1.077733</td>
<td>0.668481</td>
</tr>
<tr>
<td>5</td>
<td>0.701089</td>
<td>0.485499</td>
<td>14</td>
<td>0.988289</td>
<td>0.665366</td>
</tr>
<tr>
<td>6</td>
<td>0.560873</td>
<td>0.540201</td>
<td>15</td>
<td>0.923349</td>
<td>0.659367</td>
</tr>
<tr>
<td>7</td>
<td>0.561576</td>
<td>0.487003</td>
<td>16</td>
<td>0.882855</td>
<td>0.674423</td>
</tr>
<tr>
<td>8</td>
<td>0.684456</td>
<td>0.489109</td>
<td>17</td>
<td>0.906858</td>
<td>0.714158</td>
</tr>
<tr>
<td>9</td>
<td>0.665399</td>
<td>0.557920</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Quadratic Interpolation (two-degree polynomial)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.320976</td>
<td>0.224281</td>
<td>10</td>
<td>0.888118</td>
<td>0.755613</td>
</tr>
<tr>
<td>2</td>
<td>0.414849</td>
<td>0.289697</td>
<td>11</td>
<td>1.146219</td>
<td>0.871737</td>
</tr>
<tr>
<td>3</td>
<td>0.437615</td>
<td>0.356764</td>
<td>12</td>
<td>1.234178</td>
<td>0.754569</td>
</tr>
<tr>
<td>4</td>
<td>0.596100</td>
<td>0.372990</td>
<td>13</td>
<td>1.405404</td>
<td>0.710299</td>
</tr>
<tr>
<td>5</td>
<td>0.632423</td>
<td>0.413086</td>
<td>14</td>
<td>1.342767</td>
<td>0.815405</td>
</tr>
<tr>
<td>6</td>
<td>0.664909</td>
<td>0.496386</td>
<td>15</td>
<td>1.279747</td>
<td>0.881930</td>
</tr>
<tr>
<td>7</td>
<td>0.642061</td>
<td>0.524304</td>
<td>16</td>
<td>1.721740</td>
<td>1.007697</td>
</tr>
<tr>
<td>8</td>
<td>0.741764</td>
<td>0.649501</td>
<td>17</td>
<td>1.783753</td>
<td>0.930270</td>
</tr>
<tr>
<td>9</td>
<td>0.642238</td>
<td>0.610432</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Cubic Interpolation (three-degree polynomial)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.320976</td>
<td>0.224281</td>
<td>6</td>
<td>0.671810</td>
<td>0.499302</td>
</tr>
<tr>
<td>2</td>
<td>0.410509</td>
<td>0.285831</td>
<td>7</td>
<td>0.604892</td>
<td>0.512646</td>
</tr>
<tr>
<td>3</td>
<td>0.414070</td>
<td>0.359698</td>
<td>8</td>
<td>0.756019</td>
<td>0.644424</td>
</tr>
<tr>
<td>4</td>
<td>0.579886</td>
<td>0.371632</td>
<td>9</td>
<td>0.658574</td>
<td>0.596984</td>
</tr>
<tr>
<td>5</td>
<td>0.635801</td>
<td>0.410309</td>
<td>10</td>
<td>0.860408</td>
<td>0.722653</td>
</tr>
</tbody>
</table>
5.1: POSSIBLE METHODS OF INTERPOLATION

<table>
<thead>
<tr>
<th>Distance</th>
<th>Vertical</th>
<th>Horizontal</th>
<th>Distance</th>
<th>Vertical</th>
<th>Horizontal</th>
</tr>
</thead>
<tbody>
<tr>
<td>... continued</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>1.156109</td>
<td>0.887024</td>
<td>15</td>
<td>1.244411</td>
<td>0.887609</td>
</tr>
<tr>
<td>12</td>
<td>1.286378</td>
<td>0.757757</td>
<td>16</td>
<td>1.709685</td>
<td>1.010981</td>
</tr>
<tr>
<td>13</td>
<td>1.473586</td>
<td>0.724525</td>
<td>17</td>
<td>1.771581</td>
<td>0.938670</td>
</tr>
<tr>
<td>14</td>
<td>1.344339</td>
<td>0.836019</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The fact that four known grid cells are available opens up the possibility to perform up to three-degree polynomial interpolation. In fact, all three polynomial possibilities are going to be investigated. The coefficients of the first and second degree polynomials that are used in the method are chosen so that the polynomials fit the permeability value in the known grid cells in a least square sense. The author chooses to make use the commands `polyfit` and `polyval` in the software MATLAB to choose these coefficients and perform the interpolation\(^5\). The result in term of RMSE value is presented in Table 5.2.

Table 5.2 provides mixed-results regarding the performance of the three methods in general. For longer distance between known cells, quadratic and cubic interpolations seem to perform worse than linear interpolation. On the other hand, for shorter distance between known cells, it seems to be the other way around. However, for further research in this project, large distance between known cells is not going to be encountered. In fact, the research will focus on a case where distance between known cells is not more than two cells. For this reason, from the second method, linear interpolation is sidelined. Comparing quadratic and cubic interpolation, there seems to be little

---

\(^5\)More insight into these commands is available in MATLAB help
difference between the two in the case of shorter distance. However, cubic interpolation seems to perform just slightly better than quadratic interpolation. Therefore, from the second method, cubic interpolation is going to be used.

**Figure 5.4:** Interpolated permeability field with the second method’s cubic interpolation with 1 - 3 cell distance.
Comparing cubic interpolation from the second method and simplest linear interpolation in the first method for shorter distance between known grids, again, yields mixed result. Simplest linear interpolation in the first
method seems to perform better in RMSE, but not that much. To help the investigation, the recovered permeability field (with one to three cell distance) using second method’s cubic interpolation in comparison with the truth is presented in Figure 5.4.

Figure 5.5 shows the interpolated permeability field for both methods for the vertical interpolation case (top to bottom) side by side. By means of visual comparison between Figure 5.2 and 5.4 (with the help of Figure 5.5), it seems that cubic interpolation from the second method recovers the truth better than simplest linear interpolation. Simplest linear interpolation appears to smooth the permeability approximation while cubic interpolation does not really (to some degree). The truth is not really smooth, and hence cubic interpolation seems to suit better. For this reason, cubic interpolation with the second method is chosen over simplest linear interpolation.

5.2 Recovering the Entire Permeability Field

With the chosen interpolation method in the previous section, the next objective is to recover the permeability field of the entire reservoir using the NPBN Method. To execute this, the following experimental setup and procedure are going to be followed.

5.2.1 Experimental Setup and Procedure

Interpolation is performed after NPBN approximation of several non-overlapping grid blocks across the reservoir have been recovered.
Therefore, the first step would be to pick these non-overlapping grid blocks. As mentioned in the previous chapter, saturation is included in the model. For this reason, the grid blocks that are selected are all in the size of $6 \times 6$. For this research, nine grid blocks are chosen across the reservoir field. The location of the selected nine grid blocks in the reservoir is shown in Figure 5.6 below.

**Figure 5.6:** The location of the chosen nine medium-sized ($6 \times 6$) grid blocks in the reservoir

After the NPBN approximations of all grid blocks are obtained, interpolation to approximate the permeability field of the rest of the reservoir will be performed. The nature of the location of the selected nine grid blocks requires interpolation to be performed in two stages. First, interpolation is performed upon all grid cells that are located in between any two grid blocks (vertical interpolation for the purple cells in Figure 5.6 and horizontal
interpolation for the green cells). A problem appears when attempting to recover cells that are located not directly between two grid blocks (the white cells in Figure 5.6). This is where the second stage takes place. Using the interpolated cells obtained in the first stage, it is now possible to interpolate these cells. Two choices are available: to interpolate these cells vertically (using the green cells as the "known" cells) or horizontally (using the purple cells). RMSE analysis presented in Table 5.2 shows that horizontal interpolation yields in smaller error than vertical interpolation. For this reason, horizontal interpolation using the purple cells is chosen to recover these white cells. After performing these two steps, the permeability value of all cells in the entire reservoir field is recovered.

The experiment is, thus, performed following a similar step as described in Section 4.2.1 but with a little bit of improvisation. After the simulator is run for one time step, all nine grid blocks are updated with NPBN. Then, is run for another time step and the same procedure is repeated. Updates of the nine grid blocks with NPBN at the same time step (with the same measurements) are performed separately.

This way, the obtained result should be better because most of the (or more) cells in the entire field (in this experiment, 73.47% of all cells) are now updated "simultaneously". It implies that "only" 26.53% of the entire field that do not "assimilate" the measurements when running for the succeeding time step (ideally they should all be assimilated). Before, only one grid block of the entire reservoir that is updated in the experiment; that contributes to only
around 5 – 12% of the entire reservoir. This means that about 88 – 95% of all cells are not updated in previous experiments.

Following this procedure, the permeability of the entire reservoir field can be obtained with the NPBN method. The outcomes for this particular experiment are presented in the following section.

5.2.2 The Obtained Permeability Field by the NPBN Method

The permeability fields obtained with the NPBN method for all time steps are presented in Figure 5.7.

**Figure 5.7:** The estimated permeability field of the entire reservoir by NPBN at all time steps
5.3 Analysis of the Outcomes

Comparison between the obtained permeability field by the NPBN method and EnKF for time step 4, 7, and 8 (day 240, 420, and 480 respectively) is presented in Figure 5.8 below. The RMSE of the two approaches are presented in Figure 5.9.

Figure 5.8: The truth (top left), mean of initial ensembles (top right), NPBN estimation (bottom left), and EnKF estimation of permeability (bottom right) of the entire reservoir
Visual analysis based on Figure 5.8 yields quite a promising conclusion. The recovered permeability field by the NPBN Method seems to resemble the truth well (even though not superb). Further, comparing with the smooth-field produced by the EnKF, it seems that the NPBN delivers a relatively better approximation of the truth than EnKF does.

**Figure 5.9:** RMSE of the estimated permeability field of the entire reservoir EnKF and saturated NPBN

![Figure 5.9: RMSE of the estimated permeability field of the entire reservoir EnKF and saturated NPBN](image)

However, RMSE analysis states the otherwise. As visible in Figure 5.9, not only the RMSE value of the NPBN method is higher than EnKF, but also it seems to increase over time (after the first time step). The cause of this is suspected to be bad measurements. Yet, if so, the EnKF Method also uses the same set of measurements, but why does it seem to perform "well" (based on the RMSE) and does not seem to pick up these bad measurements? The answer to this question has been explained in the previous chapter. First of all, it seems that the NPBN method is much more sensitive to poor measurements than EnKF is. Back to back bad measurements (this will be explained in a few
paragraphs through Figure 5.10 below), for sure, does not help the case. "Fortunately" for EnKF, it has the Kalman Gain factor which tells the algorithm to trust the measurements less (because of the higher variance in this set of measurements, as already described in the beginning of the previous chapter). The NPBN method actually also implies the use of a Kalman Gain-like factor in its algorithm but this factor works in the transformed space of the state vector instead of at the state vector itself. Hence, the effect's magnitude is difference from the one EnKF experiences. Therefore, the effect of the presence of these bad measurements is much more negligible in EnKF than in the NPBN Method.

**Figure 5.10:** The five measurements, the truth, and boundaries over all time steps

![Graphs showing measurements for different variables over time](image)

Taking the analysis a little bit further, the nature of the procedure
presented in the previous section to obtain this result actually "tightens" the definition of "not-bad" measurements. Because more of the cells in the field are now assimilated during the assimilation process, "more" information is introduced to the model which results in more "certain" predictions or updates of the five measurements. This means that the "spread" of these measured variables becomes narrower and hence it is more likely for a measurement to be located outside of the "spread" (therefore it has to be truncated) than in previous experiments. Indeed, this is the case. Figure 5.10 illustrates this.

In Figure 5.10, the black line and circle represent the truth, the blue stars represent the measurements, the green dashed lines represent the "boundaries" of each measurement from previous experiments where only one part of the reservoir that is updated, and the red dashed lines represent the "boundaries" of each measurement from the current experiment. The area between the two same-color "boundaries" is what is meant by the "spread" (for each corresponding experiment) above.

First of all, Figure 5.10 supports the hypothesis that by following the procedure as described in the previous section, the obtained result should be better (than in previous experiments where only one block in the entire reservoir that is updated). This is visible by the fact that the "spread" becomes narrower (hence, the amount of uncertainty becomes less) but it still captures the truth. In other words, with this procedure, the NPBN method still calibrates well (capturing all truths) but becomes more informative (narrower spread).
The four measurements that are circled with pink circles in Figure 5.10 above are "bad" measurements which are considered as "not-bad" measurements in previous experiments. These measurements, of course, have to be truncated in the current experiment. Interestingly enough, these four measurements occur at time step 5, 7, and 8; with two of them occur at time step 5. Looking at Figure 5.9, the RMSE of NPBN method starts to increase considerably at time step 5 and it never recovers thereafter. Therefore, it seems reasonable to suspect these bad measurements as the cause of the increasing RMSE.

Further, this suspicion is strengthened by the fact that three of these four measurements are measurement of total flow rate 1 (Q1). From Figure 5.7 or 5.8, it seems that part of the NPBN approximation fields that does not resemble the truth quite as well as the other parts is the top left section of the reservoir field. Indeed, total flow rate 1 is a measurement located exactly at this area of the reservoir.

To investigate this suspicion, analysis based on the absolute error (the absolute difference between the truth and the NPBN approximation) of each cell in the reservoir field is going to be carried out. A similar analysis on EnKF approximation is going to be performed as well as a comparison.

Figure 5.11 below shall be the basis of this analysis. This figure depicts the absolute error of each cell for both methods at time step 4 (day 240), where none of the four aforementioned "bad" measurements has occurred yet, time step 5, time step 7, and time step 8. The plots in this figure are quite surprising because they reveal that cells with larger error (yellow cells) are quite spread
Figure 5.11: The absolute difference between the truth and the approximation of each cell in the reservoir with both methods (NPBN and EnKF) throughout the entire field; not just clustered in one part of the reservoir. Nonetheless, indeed, some of the yellow cells are located somewhat close to the
top left part of the field.

It is noticeable that actually the NPBN approximation behaves quite similarly with the EnKF approximation in regard to the location of "relatively big" errors. Both methods have bigger errors (mostly) at the same locations in the field. It is just that in the location where large error occurs, NPBN overestimates (or underestimates) the truth much stronger than EnKF. This can be seen by the stronger yellow in the NPBN figures. As a result, this brings the RMSE, which is the average of the absolute error of all cells in the field, of the NPBN Method up.

5.3.1 Bringing Back the Gaussian Noise

As mentioned a little bit earlier, some of the cells with big error are located quite close to the top left area of the field. This area of the reservoir is where the total flow rate 1, the measurement with several bad measurements, is measured. A further investigation to test whether the presence of bad measurements affects the performance of the NPBN Method (in term of RMSE) strongly or not is going to be performed.

The main idea is to, somehow, generate another set of measurements with no bad measurement existing in it, use this set of measurements instead, and compare the result. If the outcomes are improved when this new set of measurements is used, it is likely that bad measurements indeed cause the observed phenomenon where the RMSE of the NPBN Method increases as seen in Figure 5.9.
The set of measurements used in this experiment uses mixture of Gaussian noises with parameters as in Table 4.1. It has been mentioned in the beginning of the previous chapter that with this set of parameters, the variance of the bottom hole pressure is 2.13%, the variance of total flow rate at location 1 is 20.21%, the variance of total flow rate at location 2 is 24.23%, the variance of total flow rate at location 3 is 21.44%, and the variance of total flow rate at location 4 is 25.19%.

These values of variance are actually also well-represented in Figure 5.10. None of the bottom hole pressure measurements, whose variance is only 2.13%, is located quite far from the truth. As a result, it is located well within the spread of the measured variables. Whereas for the other four measurements, whose variances are in the 20s percent, the generated measurements are visibly located at quite some distance from the truth. Some still lay within the spread; but some of them do not.

With this particular set of measurements, it happens that there appear back to back bad measurements at total flow rate 1. However, observation upon Figure 5.10 above implies that this phenomenon is very likely to happen to any of the four total flow rate measurements when attempting to generate another set of measurements with the same noise. Therefore, generating another set of measurements with this noise is out of the table.

Based on the above observation, it seems reasonable and efficient to reuse the set of measurements used in [8] as the comparison where the noise in all measurements is Gaussian with variance 5%. With this noise, it is more likely
for the four total flow rate measurements to be located within the spread; while for bottom hole pressure, even though more noise is present, 5% does not seem to be a too big number. It is efficient because some results presented in [8] are reusable so that there is no need to run extra simulations.

To recover the permeability field of the entire reservoir with this set of measurements, the NPBN method is going to be reapplied. The same nine grid blocks as presented in Figure 5.6 will be used, and the exact same procedure is going to be followed. The only difference is the use of a different set of measurements with smaller noise. More details in the process of obtaining the result with this set of measurements, as well as the analysis, are available in Appendix C of this thesis report. The approximated permeability fields with the NPBN method and the EnKF method with this set of measurement are presented in Figure C.2 and Figure C.3 in Appendix C as well.

**Figure 5.12:** Comparison of the effect of adding non-Gaussian noise (figure (a)) instead of Gaussian noise (figure (b)) to the measurements in term of RMSE of each method.

(a) RMSE of EnKF versus NPBN with non-Gaussian noise

(b) RMSE of EnKF versus NPBN with Gaussian noise
The result in term of RMSE is presented in Figure 5.12. The previous RMSE with non-Gaussian measurements noise is presented as well a comparison.

Figure 5.12 indicates that indeed the presence of bad measurements partly causes the increasing behavior of the RMSE of the NPBN Method. With Gaussian Noise with small(er) variance, the RMSE of the NPBN Method is still higher than the RMSE of EnKF. However, it is worth stating that the difference between the two methods is not as big as in the non-Gaussian case (only about at the magnitude of around 0.1 or less); and it also behaves quite constantly throughout the time.

It is very important, however, to stress out once more that the main factor that makes the methods seem to perform better here is not the distribution (of the measurement noise). It is the variance. The non-Gaussian distributions that are used in the experiment have variance around 20% (except for the bottom hole pressure whose variance is only around 2%); while the Gaussian noise that is used in [8] only has variance 5%. The larger variance increases the probability of the occurrence of bad measurements, whose presence seems to affect the accuracy of the obtained results.

Again, RMSE is a measure of the average of the absolute difference (between the approximation and the truth) in the entire field and hence, it can be misleading. Therefore even though in Figure 5.12 above the RMSE of the NPBN Method is higher than the EnKF Method, there still exist reasons to
believe that this NPBN Method is a promising approach. Further research and investigation may result in better understanding of the method and probably, to some degree, in better performance.

5.4 Evaluation of the Chosen Interpolation Method

As stated in the beginning of this chapter, the interpolation methods that are considered in this thesis project are based on known locations (of known permeability from the nine grid blocks) only. It is also mentioned there that more advanced interpolation methods, some even consider the randomness of the field, are available. At this point, one question emerges: "Do we need these more advanced methods in our case?". To answer this question, evaluation of the method that is eventually used in this project is performed.

In Section 5.1, some possible interpolation methods are investigated and in the end, cubic interpolation with four known grids is chosen. In Section 5.2, the cubic interpolation method is applied and the NPBN approximation of the entire field is obtained. In this section, the performance of the cubic interpolation is going to be evaluated by using the approximated permeability field of the NPBN method.

To evaluate the performance, consistency between the permeability of the interpolation cells and the cells that are part of the blocks (as shown in Figure 5.6) is going to be observed. The observation is done by looking at the deviation between the approximated permeability and the corresponding truth of one cell to another. In principal, the error of the interpolation cells and the "blocks cells"
should behave similarly because no extra information, other than the location of the interpolation cells, is introduced to the system during the permeability approximation of the interpolation cells. For example, it would be strange to encounter a case where bad approximation at the block cells produces much more accurate approximation at the interpolation cells.

Based on the above description, Figure 5.11 is perfect for this purpose. The left column of the figure presents the absolute difference of the NPBN approximation and the truth at time step 4, 5, 7, and 8 in every cell. These plots provide exactly what is needed for the evaluation. Figure 5.13 below presents these four plots with black lines to help distinguishing the blocks and the interpolation cells.

**Figure 5.13:** The absolute difference between the truth and the approximation of each cell in the reservoir with the NPBN method
Visual observation upon Figure 5.13 suggests that there seems to be no problem with the cubic interpolation method. The absolute error of the interpolation cells and the blocks cells appears to behave quite similarly with no distinguishable pattern between the two nor observed "strange" behavior. Some interpolation cells even look "worse" than the blocks cells; but this is, of course, not that surprising to encounter.

However, one might still be interested in the quantification of this evaluation procedure. One possible way to do this is by calculating the RMSE of the blocks cells and the interpolation cells separately and compare them. The result is presented in Figure 5.14.

**Figure 5.14:** The RMSE of the interpolation cells, blocks cells, and "relevant" block cells

![Figure 5.14](image)

Figure 5.14 shows how the RMSE of the blocks cells (blue line) and the interpolation cells (red line) progress through the eight time steps in the simulation. Directionality-wise, the two seem to agree with each other (albeit
at time step 8 where the the blocks cells’ RMSE goes up while the interpolation
cells’ goes down).

However, one may notice that at all time, the RMSE of the interpolation
cells is always below the RMSE of the blocks cells with difference magnitude of
around 0.1. This phenomenon, however, does not mean that the "strange"
behavior –where bad approximation at the blocks cells somehow produces
much better approximation at the interpolation cells– is encountered. This
phenomenon is explained by the nature of the cubic interpolation that is used
in this thesis project and also the obtained approximation of the permeability
field. As implied in the description of the cubic interpolation in Section 5.1, not
all cells in the grid blocks are used in recovering the interpolation cells. Hence,
these "non-contributing" cells "have nothing to do" with the interpolation cells.
However, as visible in Figure 5.13, some of these non-contributing cells have
larger absolute error than the others. These cells bring the RMSE of the blocks
cells up; and this is likely to be the cause of this phenomenon.

For this reason, a third RMSE line (the black line) is introduced in Figure
5.14 above. This black line represents the RMSE of only the "relevant" cells in
the grid blocks, i.e., cells that are involved in the cubic interpolation performed
in the project. By analyzing this black line (instead of the blue line), the concern
about non-contributing cells bringing up the RMSE of the blocks cells becomes
irrelevant.

Directionality-wise and value-wise, the black RMSE line appears to
behave much more similarly to the red RMSE line. This is a good news because
it means that the interpolation cells behave hand in hand with the relevant blocks cells. A similar analysis has also been performed in the case where the Gaussian noise is used; and it yields a similar conclusion (more insight into the analysis is presented in Section C.4 of Appendix C in this thesis report).

Therefore, it appears that for this particular case study, a simple cubic interpolation performs reasonably well and there may be no urgent need to research a more advanced interpolation technique. However, for the interest of further development, a more advanced interpolation method is, of course, still not ruled out and might, in fact, also be of interest.
CHAPTER VI
FORECAST (PREDICTION)

As mentioned several times in the previous chapters, the performance of Root Mean Square Error (RMSE) as the measure of performance is questionable because it shows only an average behavior of the estimated field. Therefore, it can be misleading. As a consequence, another measure of performance should be used. This chapter tries to solve this problem by introducing one other way of measuring the performance: forecast, which is also commonly known as prediction in reservoir engineering. For this reason, the terms forecast and prediction are going to be used interchangeably from hereon.

6.1 Forecast (Prediction)

Forecast is a measure of performance that "assesses" (or tests) the quality of the obtained parameter approximation; which in this thesis project is the approximated permeability field (from either methods: EnKF or NPBN). The quality is defined in terms of how "well" (or how "accurate") the approximated parameter is able to forecast or predict the value of predetermined criterion(s) in the model. The criterion(s) can be anything that one thinks is (are) important.
Because the truth is known, the actual value of the criterion(s) is (are) known and hence one is able to observe how accurate the approximation is in regard to the corresponding criterion(s). If the approximation is able to predict the value of the criterion(s) well, the approximation is said to have good quality.

There are two ways of performing forecast. The first way is to run the simulator further in time from the last time step of estimation; and the second one is to run the simulator from the starting time (time zero) with the estimated permeability field and check how much it resembles the true forecast of the criterion(s).

### 6.2 Implementing Forecast as A Measure of Performance

#### 6.2.1 Choice of Criterion: Water Breakthrough

This new measure of performance is going to be implemented in this thesis project to assess the quality of the recovered permeability field from a different point of view. The second way of performing forecast (running the simulation from the beginning) is the one that is going to be used in this project. The next question to answer is an important one: "What criterion(s) should be used in assessing the quality of the estimated permeability field?"

The answer to that question is the time of *water breakthrough* in the four production wells. Water breakthrough, as already mentioned in Chapter 3, is the moment water starts to also sip out from the production wells. When water is injected to the reservoir through the injection well, more pressure is formed inside the reservoir that pushes oil out through the production wells. During
the process, the injected water also travels inside the reservoir towards the four production wells and after some time, it finally reaches the wells and is pumped out from the production wells as well. This moment when water is also pumped out from the production wells is what is called water breakthrough.

The time of water breakthrough might differ for different wells. This is due to the characteristics of the reservoir itself which affects the way fluid travels in it. For example, it is more difficult for the fluid to travel through impermeable rock rather than permeable rock. Hence, the time of water breakthrough of production well at the impermeable part of the reservoir is longer than well at the permeable part of the reservoir.

One may be curious about the reason behind the choice of the time of water breakthrough as a criterion in this method. (Some) people are interested in the information about the time of the water breakthrough for economic reasons. After water starts to sip out from the production wells, the production of oil becomes gradually less and less; thus it becomes less profitable while the cost of operation (injecting water at the injection well and pumping out oil (and water) from the production wells) stays pretty much the same. It means that after the time of water breakthrough, the activity of oil recovery from the production well becomes economically less valuable. It may be better to stop the activity in the production well (or close the production well) before the water breakthrough in order to avoid unnecessary costs of production.

Therefore, the ability to predict the time of water breakthrough is a precious feature for a method to possess. If a method is able to predict the time
of water breakthrough better, it translates to better cost-minimizing procedures. Thus, it also makes the method much more attractive to use. For this reason, water breakthrough is chosen.

6.2.2 Results

In the model, the time of water breakthrough can be observed by measuring the water flow rate that comes out from each of the production wells. Water flow rate is part of the total flow rate (that is used as the measurement taken at the four production wells) which is the summation of water and oil flow rate.

The model that is used in the experiment is a two-phase model where only two fluids are available, oil and water. Water enters the reservoir through injection at the injection well. Therefore, water spurs out some time after the injection starts. Before that time, the production wells only produce oil, which implies that the water flow rate is zero. Hence, the time of water breakthrough is the time the absolute value of the water flow rate in a production well increases to a positive number.

To measure the performance, the first step is to use the true permeability field so one can look at the behavior of the truth. To analyze the performance of both NPBN Method and the EnKF, the two estimated permeability fields at the last time step (day 480) are inputted to the model to generate the water flow rates plots. The differences between the water flow rates given by the two methods and the truth can be now compared.
Figure 6.1 presents the water flow rates of each of the four production wells. Note that because the flows in all production wells are upward, the values of the water flow rate are actually negative. Figure 6.1 shows the absolute value of these water flow rates.

**Figure 6.1:** Prediction of the time of water breakthrough. The curves correspond to water flow rate in each of the production wells

First of all, Figure 6.1 indeed shows that water is faster to reach the production wells at location 2 and location 3. This makes perfect sense because these two wells are located in parts of the reservoir that are relatively more permeable. Also, the paths from the injection well located in the middle of the field to these two locations are relatively more permeable than paths to the
other two wells. Therefore, it is plausible for the water breakthrough to occur earlier at location 2 and location 3 than at location 1 and location 4.

Figure 6.1 shows that the NPBN method performs really well at location 1 and location 2 where it predicts the time of water breakthrough accurately. EnKF is late by approximately 60 - 70 days at location 1 and early by approximately 15 - 25 days at location 2. However, both methods do not predict the time of water breakthrough as well at location 3 and 4. At these two locations, both methods’ predictions are too early than the actual water breakthrough with NPBN performs worse than EnKF. Respectively, ENKF is early by around 15 - 25 days and 100 - 120 days at location 3 and 4; while NPBN is early by around 100 - 120 days and 230 - 250 days at location 3 and 4.

This observation shows that indeed the NPBN method performs comparatively well. It is true that the prediction by NPBN at location 3 and 4 are further off than EnKF; but it is worth stating that it performs very well at location 1 and 2. This observation of water breakthrough supports the conclusion obtained by visual comparison of the permeability field where the NPBN method recovers the permeability field well. Therefore, there is more reason not to (blindly) trust the RMSE analysis as presented in the previous chapters. Indeed, in some ways, RMSE seems to be misleading.

As mentioned in the previous chapters, the NPBN method seems to be influenced by bad measurements much stronger than EnKF does. This may explain the inaccurate behavior at location 3 and 4 of both methods (the NPBN
method specifically) at Figure 6.1. At the end of Chapter V, the Gaussian noise used in [8] was brought back to the table to "avoid" the presence of bad measurements in the experiment. The same reasoning is used here. The prediction of water breakthrough is going to be performed one more time using the results obtained in that experiment with Gaussian noise (for details in Appendix C).

**Figure 6.2:** Prediction of the time of water breakthrough with Gaussian noise

Following the same steps as before, the plots of the absolute value of the water flow rate of each of the four production wells with the same measurements as used in [8] are presented in Figure 6.2.
Figure 6.2 shows that with this older set of measurements, both methods perform better. None of the errors (in any location) for any method is more than 100 days. In this case, however, the NPBN performs better than the EnKF in overall. At location 1, 3, and 4, the prediction of water breakthrough time given by NPBN is closer to the truth than EnKF. At location 3, EnKF performs better than NPBN with NPBN being late in predicting the water breakthrough. However, the error does not seem to be that big anyway.

As mentioned at the end of Chapter V, the main factor that makes the methods perform better here is not the distribution (of the measurement noise), but the variance. The Gaussian noise used in [8] only has 5% variance and so the probability of bad measurements to occur is much smaller than when using non-Gaussian noises whose variances are around 20% for four of the five measured variables.

Nevertheless, it seems that this measure of performance is better than the RMSE. The conclusion, somewhat, agrees with visual observation that the NPBN method performs quite well in recovering the truth.
CHAPTER VII
CONCLUSIONS AND RECOMMENDATIONS

7.1 Summary and Conclusions

7.1.1 Summary of Project

This section presents the summary of the thesis project that has been presented in Chapter I to Chapter VI of this thesis report. More detailed conclusions are presented in the next section.

The NPBN based approach in the history matching problem in reservoir simulation is still a very young approach. The project presented in this thesis report is meant to research the performance of this approach deeper and to compare further its performance with the well-known EnKF. Previous studies have indicated that the NPBN approach is a promising method and its performance is comparable with the performance of EnKF. Even though, with 900 available samples the NPBN seems to be able to handle, in an accurate manner, only around 100 variables, estimating only part of the reservoir does not affect the accuracy of the estimation. To estimate the entire field, interpolation between smaller grids approximations performed across the
entire field seems to be one possible way.

There are three key points of interest aimed with this thesis project. The first one is the distributions of the measurements noise, the second one is the inclusion of saturation to the NPBN approach, and the third one is a possible interpolation method to combine the approximations of different parts of the field on which the NPBN is performed upon. Along the way, a fourth objective is surfaced; and that is to perform an alternative way of measure of performance.

The first step towards the research is to build another set of measurements with different noise. To do that, the first thing on the list is to define the distribution of the noise. The chosen distributions for the measurement noise are all mixture of Gaussian and they are presented in Table 4.1. The process and reasoning of using this set of distributions are presented in Appendix B of this thesis report.

The variances of the "new" noise turn out to be different from the previous set of measurements' as used in [8] where the variances are all 5%. For bottom hole pressure, the variance is around 2%, while for the four total flow rate measurements, the variances are all above 20%. This (generally) "higher" variance means that bad measurements are more likely to occur. As pointed in Section 4.1, this is indeed the case. In the new set of measurements, bad measurements appear at time step 2 and time step 8. Deeper into the research, it is also revealed that some other measurements at time step 5 dan time step 7 are also quite troublesome. The emergence of bad measurements is pretty much unavoidable as they keep always appearing in all other attempts.
to build a set of measurements.

To start the research, the first step is to investigate the performance of the NPBN method (along with comparison with ENKF) at several grid blocks in the reservoir with this new set of measurements. Three grid blocks are chosen in the first part of Chapter IV. The first grid block is the same 7×7 grid block as used in [8], the second grid block is a 5×5 grid block located near the location of bottom hole pressure measurement, and the third grid block is a 5×5 grid block with several overlapping cells with the previous two grid blocks.

Moving forward, the next step is to consider the possibility to include saturation to the model. In [8], saturation is excluded from the model because of its constant-like behavior. A constant variable is independent of anything else and thus, it does not affect the other variables. However, saturation is not exactly constant and exclusion of it may result in loss of valuable information.

Before proceeding to the investigation, a necessary preparatory step has to be performed first; that is updating the Bayesian networks with the addition of extra nodes which represent saturation. The arcs directionality to/from the saturation nodes is based on the physical representation of the corresponding equations of saturation and the other variables. The updated NPBN is presented in Figure 4.14.

Once the Bayesian networks is updated, the investigation can be performed. The two 5×5 grid blocks as used in the previous experiment are going to be reused. This way, improvement resulted by addition of saturation can be observed immediately. The obtained results are not surprising.
Inclusion of saturation does not really add that much valuable information to the system, even though it indeed improves the accuracy a little bit. The estimated permeability fields look very similar to the estimated fields without saturation and RMSE analysis tells the same story with the RMSE of the two of them (with and without saturation) are very close to each other for both grid blocks.

Even though the improvement is not superb, for further research conducted in the thesis project, saturation is still going to be included. As a consequence, these further experiments can only use even smaller-sized grid blocks due to the extra dimensions introduced by addition of saturation to the model. For this reason, only grid blocks of size $6 \times 6$ are used in the next step of the research: the interpolation.

Before actually performing interpolation, it is key to first choose which interpolation method that is going to be used to combine the grid blocks across the reservoir. To initiate the research regarding possible methods of interpolation, the simplest possible ways of interpolation based on locations (and not taking into the account the 'randomness' of the field) are going to be considered. In the end, cubic interpolation is chosen.

After the method is chosen, interpolation can be performed. First of all, the grid blocks that are going to be interpolated need to be determined. Nine grid blocks with size $6 \times 6$ are going to be used. The locations of these nine grid blocks are depicted in Figure 5.6. Then, the NPBN method is performed at these nine separate grid block simultaneously. The resulted estimated fields
from these grid blocks are then interpolated together with cubic interpolation resulting in the estimated permeability field of the overall reservoir with the NPBN method. The results for every time step are presented in Figure 5.7.

Visually, the obtained permeability field by the NPBN method looks good and it seems to outperform the smooth field generated by the EnKF. However, the RMSE analysis states otherwise as the RMSE of the NPBN method increases over time and is (much) worse than the RMSE of EnKF. The presence of bad measurements is suspected to play a role on this RMSE "fiasco". Therefore, the previous measurement noise as used in [8] is brought back to the table to provide comparison. The same procedure is repeated with this older set of measurements (details in Appendix C). The obtained result in term of RMSE seems to be better and more stable than before; indicating that bad measurements indeed play some parts in causing the weird behavior of the RMSE. However, visual observation and RMSE analysis, even with this older set of measurements, still seem to be contradictory. Therefore, the concern about using the RMSE as a measure of performance is increasing since it can be misleading.

For this reason, a fourth objective is surfaced, which is to use an alternative measure of performance. In Chapter VI, a measure of performance based on the quality of the approximated permeability field is introduced, which is the forecast or prediction. The chosen criterion to measure the quality of the estimated permeability field is how well the methods (NPBN or EnKF) are able to predict the time of water breakthrough at the four production wells.
The NPBN method, indeed, performs well and outperforms EnKF at two of the four locations of the production wells. At the other two locations, both methods estimate the time of water breakthrough too early. Overall, with this measure of performance, the NPBN method performs well in comparison with the EnKF.

When this measure of performance is used upon the case with the older Gaussian noise, the NPBN method outperforms EnKF at three of the four production wells.

Stepping aside, highlighting the chosen interpolation method, a little further analysis reveals that cubic interpolation already works reasonably well for this particular case study. Therefore, for the time being, there may be no urgent need to research a more advanced interpolation method; even though, of course, a more advanced method is always welcomed in the development and improvement of the NPBN Method in the future.

7.1.2 Conclusions

A summary of the main conclusions of the research that has been performed in this thesis project is as follows:

1. With the set of measurements generated by the chosen noise distributions for each measured variables as presented in Table 4.1, it is very likely for bad measurements to occur. Indeed, bad measurements seem to occur at time step 2 (day 120) and time step 8 (day 480) at the generated set of measurements used in the thesis project.
2. Overall, the EnKF seems to perform in a steady behavior. Especially with the non-Gaussian noise, this is expected because the relatively higher variances of measurements make the Kalman Gain factor tell EnKF to trust the measurements less. It is good that EnKF damps the effect of bad measurements. NPBN method, on the other hand, is more sensitive to the measurement noise. Hence, the NPBN approach benefits much more than EnKF when good measurements are present, but its performance is affected negatively when bad measurements strike.

3. Moreover, the EnKF also does not seem to appreciate the proximity of the grid block to a good (or bad) measurement, unlike the NPBN method.

4. The NPBN approach has its merit in recovering impermeable cells and it estimates the very impermeable field more accurately. In fact, NPBN often overestimates the impermeability of the field. EnKF, however, produces a more smooth approximation of the field. Hence, it generally does not appear to perform very well when the truth is not smooth.

5. Comparison upon the overlapping cells between different grid blocks indicates the consistency of the NPBN approach. The behavior of one cell that is updated with two different grid blocks seems to still be similar from the first experiment to the second.

6. As expected, the inclusion of saturation does not improve the accuracy of the NPBN method significantly. The estimated permeability fields with and without saturation look very similar to each other, and this is
supported by the RMSE analysis as well. However, improvement is indeed still present, albeit small. Therefore, for further experiments in the project, saturation is included in the model; even though as a consequence, only smaller-sized grid blocks can be used in the experiments due to the extra dimensions added to the system.

7. From the two location-based interpolation methods that are researched, cubic interpolation seems to perform better. Therefore, for the interpolation step in this thesis project, cubic interpolation is used. Post-interpolating, a further analysis upon the obtained results reveals that cubic interpolation works reasonably well for this particular case study.

8. Performing interpolation to nine separated grid blocks across the reservoir that are updated with the NPBN method yields the full estimation of the permeability field by the NPBN. The obtained permeability field looks good as it seems to capture the geological behavior of the field better (the not smooth field). Further, based on visual observation, the NPBN approximation also seems to outperform the smooth EnKF approximation. However, RMSE analysis states otherwise. Not only the NPBN method performs worse than EnKF (in term of RMSE), the RMSE value of NPBN also increases over time while the RMSE of EnKF stays stable.

9. One suspected cause of this RMSE "fiasco" is the presence of bad
measurements. A deeper investigation shows that bad measurements, apparently, occur not only at time step 2 and time step 8, but some measurements at time step 5 and time step 7 are also considered not-good measurements. Because the NPBN method is much more sensitive to the presence of bad measurements (point 2 and 3 above), it seems reasonable to suspect these bad measurements as a cause of the increasing RMSE behavior.

10. Plotting the absolute difference between the truth and the estimated field by both EnKF and NPBN, as presented in Figure 5.11, it turns out that the cells with larger error are spread throughout the entire field. However, it is interesting to see that the NPBN method behaves quite similarly to the EnKF method in regard to the location of "relatively big" errors. Both methods have bigger errors (mostly) at the same locations in the field. It is just that in the location where large error occurs, NPBN overestimates (or underestimates) the truth much stronger than EnKF does.

11. Because the presence of bad measurements is suspected to cause the volatile performance of the NPBN method in term of the RMSE analysis, the old set of measurements as used in [8] is brought back to the table. The motivation behind this decision is that the noise’s nature used in this set of measurements sets bad measurement less likely to occur. The performances of the NPBN method and the EnKF with this noise are going to be used as comparison.
12. Indeed, it seems that bad measurements play a role in the RMSE fiasco of the NPBN method. With the set of measurements from [8], the RMSE of the NPBN behaves more steadily. However, the same phenomenon about the contradictory conclusions between visual observation and the RMSE analysis still emerges.

13. An alternative way to measure the performance of the methods, the forecast or prediction, seems to give more trustworthy information than the RMSE. The criterion used to measure the performance is the time of water breakthrough at the four production wells. The obtained result seems to agree with the visual observation, where it has concluded that the NPBN method performs well. In fact, the NPBN method outperforms EnKF at two of the four locations; even though at two other locations, EnKF performs better than NPBN (although the performance of EnKF itself at these two locations is not that superb either).

When applied to the obtained permeability estimations with Gaussian measurement noise (as used in [8]), the result gets even better as both methods predict the truth better. Further, more importantly, the NPBN method outperforms the EnKF at three out of the four production wells.

14. For this particular case study as presented in this thesis, it is not the distribution (of the measurement noise) that makes the difference between the performance of both methods when comparing the first results with non-Gaussian measurement noise and the second results
with the older Gaussian measurement noise as used in [8]. It is the
variances of the noises.

As indicated by all the results and analysis thus far, the NPBN method
seems to be an even more prospective alternative approach to solve the history
matching problem. However, the method is still in its infancy and therefore,
further research still needs to be done to investigate the method even further;
or even to improve its performance wherever possible.

7.2 Recommendations

Based on the research that has been performed in this thesis project, the
author proposes some recommendations for further research about this new
NPBN approach in history matching that are of interest. They are as the
following:

1. As already mentioned or implied several time, the presence of bad
measurements disturbs the performance of the NPBN method. Therefore,
it may be of future interest to somehow generate a set of measurements
with no presence of bad measurements in it. The analysis presented in
Appendix B, mainly the concerns in Appendix B.4, may be used as a
stepping stone to reach this objective.

2. In the Case Study in Chapter IV, it is said that the NPBN method is
consistent. This conclusion is drawn based on visual observation of the
overlapping grid cells only. Therefore, a more rigorous research to
evaluate the consistency of the method may be of future interest.

3. The addition of saturation does not improve the performance of the NPBN method that much. However, improvement is still present, albeit small, and this is why saturation is included in the model in this thesis project. However, inclusion of saturation means more dimension is involved in the system, while the NPBN method only seems to perform well with around 100 variables or dimensions at a time. Therefore, the size of usable grid-block has to be sacrificed a little bit. For a $6 \times 6$ grid block, 113 variables are already involved in the system with saturation included in it.

However, it all means that in the future, one can actually "trade" saturation for something else that may contribute to the model better or more significantly than saturation. This way, one needs not to reduce the size of the usable grid block even further.

Another possibility is to take saturation at only some cells of the grid block instead of at every cells of the grid block. This will not add that much dimensionality problem to the system. Another advantage of this alternative is that one may introduce another node/factor to the system without having to completely "trade off" saturation nor shrinking the size of the grid block(s) massively. However, which cells to pick (to represent saturation) and how to pick those cells definitely need further rigorous research.

4. The interpolation method that is used in this thesis project is a very
simple location-based method that does not take into the account the ‘randomness’ of the field. Analysis of the obtained results reveals that the cubic interpolation works reasonably well for this particular case study. However, for the interest of further development, a more advanced interpolation method is still not ruled out and might, in fact, also be of interest.

5. Taking a little step back, the interpolation method that is used in this thesis project may need to be tested and evaluated further. Sensitivity of the location and size of the grid blocks to be interpolated could be investigated. Will the results of difference choices be consistent?

6. It seems that one should not rely too much on RMSE to measure the performance of the method(s). RMSE is a measure of the average of the absolute difference (between the approximation and the truth) in the entire field and hence, it can be misleading. This seems to be the case here. Visual observation and forecast seem to be better measures of performance and they agree with each other. However, there are still a lot of other possible ways to measure the performance of the methods. Another alternative may be of future interest.


A.1 Several Distributions

In this section, several distributions that are used or referred to in this report, mainly Appendix B, are described briefly.

A.1.1 The Normal Distribution

The Normal (or Gaussian) distribution is a continuous probability function that has a bell-shaped probability density function given by the following form:

\[ f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \] (A.1)

where \( \mu \) denotes its mean and \( \sigma^2 \) denotes its variance. A normally distributed random variable \( X \) with mean \( \mu \) and variance \( \sigma^2 \) is denoted \( X \sim \mathcal{N}(\mu, \sigma^2) \). The normal distribution with \( \mu = 0 \) and \( \sigma^2 = 1 \) is called standard normal.

A.1.2 The Log-normal Distribution

A log-normal distribution is a continuous probability function which is the exponential of a normal distribution. In other words, if \( X \) is a log-normally
Figure A.1: Probability density function of standard normal distribution (left) and lognormal distribution with $\mu = 0$ and $\sigma^2 = 1$ (right)

distributed random variable, $Y = \exp(X)$ has a log-normal distribution. Hence, the parameters are the same as in the normal distribution, $\mu$, and $\sigma^2$. The mean and variance of this distribution are, respectively: $\exp(\mu + \sigma^2/2)$ and $(\exp(\sigma^2) - 1)\exp(2\mu + \sigma^2)$. By definition, it is obvious that a random variable log-normally distributed $Y$ is non-negative, i.e, it has support $y \in (0, +\infty)$.

One useful property of a log-normal distribution is that if $Y$ is a log-normally distributed random variable with parameters $\mu$ and $\sigma^2$, then $Y + c$, with $c \in \mathbb{R}$, is said to have a *shifted log-normal distribution* with support $y \in (c, +\infty)$. Its expectation is $\mathbb{E}(Y) + c$ and its variance is $\text{Var}(Y)$.

A.1.3 Logistic Distribution

A logistic distribution is a continuous probability function which shape resembles the normal distribution but with heavier tails. The probability
function is defined as:

\[ f(x; \mu, s) = \frac{e^{-(x-\mu)/s}}{s(1 + e^{-(x-\mu)/s})^2}. \]

The mean and variance are \( \mu \) and \( \pi^2 s^2/3 \) respectively.

### A.1.4 Mixture Distribution

A random variable \( X \) is called to arise from a (finite) mixture distribution if the probability density function \( f(x) \) can be written as a mixture density for all \( x \in \mathcal{X} \), where \( \mathcal{X} \) is the sample space, as:

\[ f(x) = p_1 f_1(x) + p_2 f_2(x) + \ldots + p_n f_n(x) = \sum_{i=1}^{n} p_i f_i(x). \tag{A.2} \]

In (A.2) above, \( f_i(x) \) denotes the probability density function for all \( i = 1, \ldots, n \) and \( p_1, \ldots, p_n \) denote the weights of each density with \( \sum_{i=1}^{n} p_n = 1 \). Of course (A.2) can be rewritten in the form of cumulative distribution function by taking the integral from \( -\infty \) to \( x \) on each side to yield the following form:

\[ F(x) = p_1 F_1(x) + p_2 F_2(x) + \ldots + p_n F_n(x) = \sum_{i=1}^{n} p_i F_i(x) \tag{A.3} \]

where \( F_i(x) \) denotes the cumulative distribution function for all \( i = 1, \ldots, n \). A single density \( f_i(x) \) is referred to as the component density. In most applications, it is assumed that all component densities are from the same parametric distribution family [7].

In case of a mixture of univariate normal distributions each with weight \( p_i \), mean \( \mu_i \), and variance \( \sigma_i^2 \); the mean and variance of the mixture are defined
as [7]:

\[
\mu = \sum_{i=1}^{n} p_i \mu_i \quad \text{(A.4)}
\]

\[
\sigma^2 = \sum_{i=1}^{n} \left( \mu_i^2 + \sigma_i^2 \right) p_i - \mu^2. \quad \text{(A.5)}
\]

**Figure A.2:** The density of a mixture distribution between two univariate normal distributions with \( \mu_1 = -1 \) and \( \mu_2 = 1 \), \( \sigma_1^2 = 0.5 \) and \( \sigma_2^2 = 1 \), and weight 0.4 and 0.6 respectively. The red dashed curves represent the weighted density of each component.

There is already a built-in class in the software MATLAB that deals with this mixture of (finite) univariate normal distributions. The class is called `gmdistribution`.

### A.2 Relative Information

If \( (p_1, \ldots, p_n) \) and \( (q_1, \ldots, q_n) \) are probability vectors with \( q_i > 0, i = 1, \ldots, n \), the relative information of \( p \) with respect to \( q \) is defined as:

\[
I(p, q) = \sum_{i=1}^{n} p_i \log \left( \frac{p_i}{q_i} \right). \quad \text{(A.6)}
\]
From (A.6), $I(p,q) \geq 0$ and $I(p,q) = 0$ if and only if $p = q$. $I(p,q)$ can be interpreted as measuring the degree of ‘uniformness’ of $p$ with respect to $q$. [4, 13]

Let $s = (s_1, \ldots, s_n)$ denote a sample distribution generated by $N$ independent samples from the distribution $p$. Let $\chi^2_d$ denote the cumulative distribution function of a chi-square variable with $d$ degree of freedom. Then:

$$\mathbb{P}(2NI(s,p) \leq x) \rightarrow \chi^2_{n-1}(x), \quad \text{as } N \rightarrow \infty. \quad (A.7)$$

Equation (A.7) above says that the statistic $2NI(s,p)$ is asymptotically $\chi^2$-distributed as the number $N$ of independent samples goes to infinity [4].

Relative Information will come in handy when a Goodness of Fit test between a dataset and a theorized distribution function of that dataset needs to be performed. This will be presented in Appendix B.
APPENDIX B
MEASUREMENTS GENERATION

The measurements synthesized in [8] are generated by adding Gaussian noise. However, a visual inspection of the histograms of the observable variables indicates that the measurements are not normally distributed. This phenomenon is indeed one of the main subjects of this thesis project. Another set of measurements using non-Gaussian noise will be generated. This new set of measurements bases the case study presented in Chapter IV of this thesis.

The process of the new measurements generation will be discussed and presented in detail in this appendix.

B.1 Data Spread of Observable Variables

As mentioned earlier, in this section, the goal is to find the distributions that match the observable variables better than normal distribution with variance 5% as done in [8]. Hence, the first objective is to find the data sets of observable variables which distribution fitting needs to be performed upon.

These data sets can be obtained by running the simulator simsim for 60 days. As pointed out in Chapter III of this thesis report, in this simulation,
measurements are performed once every 60 days for the period of 480 days. It
means that eight sets of measurements are generated during the entire 480 days
period. Further in this report, one ‘time step’ is meant to be a period of 60 days
(starting from day zero). Hence, time step 1 corresponds to day 60, time step 2
corresponds to day 120, and so on.

Running simsim for the first time step yields a synthetic data set
consisting of the values of all entries of the state vector in time step 1 each with
900 members (because 900 samples or ensemble members are used). Five of
these entries are none other than the five measurable variables and the
synthetic data will be used to describe the noise.

At this point, two key questions may arise:

1. For an observed variable, should the distribution of noise for each time
   step be different?

2. The general objective of this distribution fitting is to find distributions
   that match the noise of the observed variables. The data sets obtained by
   running simsim, however, consist of the value of the observed variables.
   How should the translation from the value of an observed variable to the
   noise of it be?

To answer the first question, it is assumed that there is no significant
apparent change in the shape of the histograms for all the variables over all
time steps. Hence, using this assumption, there is no need to use different
distribution for each time step for any of the observed variables.
For the second question, the translation from the *value* of a variable to its *noise* is done by the assumption that the "best estimate" of each variable should be well-approximated by the sample mean of that particular variable as generated by `simsim`. The mean of the data is easily approximated by simply taking its average. Hence, subtracting the sample mean from the value of each sample yields a data set consisting of the "noise" or "error". Performing this procedure on the five observable variables produces five empirical distributions for the noise, each corresponding to one observable variable. These data sets of noise are the ones distribution-fittings should be performed upon. The goal is to fit a parametric distribution to the empirical ones presented in Figure B.1.

**Figure B.1:** Histograms of the noise of the five measurements from data.
Therefore, running the simulator `simsim` for one time step and then performing the aforementioned procedure to obtain the "noise" generates the data sets for the noise of the observable variables.

**B.2 Distribution Fitting**

Now, using the obtained synthetic data sets for noise as presented in Figure B.1 above, the next step is to find distributions that match each of the five variables. This step shall be performed with the help of software MATLAB. In particular, two built-in commands in MATLAB shall be used: `fitdist` and the `gmdistribution` class.

Out of 20 parametric distributions that are supported in the class `fitdist` in MATLAB version R2011b, aside from the normal distribution, it seems that the log-normal and the logistic distributions are two distributions that may better represent bottom hole pressure (bhp) and the total flow rate, respectively, shape-wise. The command `fitdist` returns the best parameter(s) of a distribution to match a given data set fed as input. Hence, the (best) parameters for log-normal and logistic distributions to match the bhp data set and four total flow rate data sets, respectively, can be obtained by running this command.

One may notice that some of the data from the bottom hole pressure data set as in Figure B.1 have negative value. However, log-normal distribution, whose support is $(0, +\infty)$ or positive, is going to be fitted to this data set. This shall cause some problems. To deal with this situation, the shifting property of
a log-normal distribution are used. Instead of being-fitted to the noise of the bottom hole pressure, the log-normal distribution is fitted to the value of the bottom hole pressure instead. By definition, bottom hole pressure is non-negative and hence log-normal can be fitted. Then, the resulting fitted distribution will be shifted back to the noise using the same procedure of translation from value to noise as discussed in the previous section.

Aside from the log-normal and the logistic distributions, a mixture of normal distributions is also of interest. Gaussian mixture distribution fitting can be done by working with the gmdistribution class in MATLAB, in particular the gmdistribution.fit command. Three important outputs of this command are the mean, variance, and (normalized) weight of each of the individual Gaussian distribution. Because each of the data sets only contains 900 samples, a number that is not that large, a mixture of two normal distributions is of interest\(^6\).

The following figures illustrate the results of the distribution-fitting process. There are four plots in each figure. The first one is the data set, which is just the same as the corresponding data set as presented in Figure B.1. The second one is the fitted mixture of normal distributions with parameters generated by the command gmdistribution.fit. The third one is the fitted of log-normal (or logistic for the case of total flow rate) distribution with parameters generated by the command fitdist. Finally, the last one is a normal distribution with mean zero and variance 5%.

\(^6\)fitting more normal distributions to the data set triggers an error message in MATLAB
Figure B.2: Fitting bottom hole pressure (bhp).

Figure B.3: Fitting total flow rate at location 1 (Q1).
Figure B.4: Fitting total flow rate at location 2 (Q2).

Figure B.5: Fitting total flow rate at location 3 (Q3).
In all five cases, the mixture of (two) Gaussian distributions seems to be the best fit for the data. Shape-wise and based on observation upon the spread of the samples, the mixture of Gaussian distributions matches the data sets better than the log-normal or the logistic distributions. Moreover, in all cases, the fit is much better than the normal distribution with variance 5%. Hence, in this thesis project, a mixture of Gaussian distributions will be used in the noise generation process for the measurements. The parameters of each of the mixture of Gaussian distributions that are used in the five cases above are already presented in Table 4.1 in Chapter IV.

B.3 Goodness of Fit

In the previous section, mixtures of Gaussian distributions are picked to generate the noise of the measurements. The basis of these choices are
similarities between the plots of the data sets and the plots of the distributions as presented in Figure B.2 - Figure B.6. However, it seems that this reasoning is not a rigorous one in a statistical or mathematical sense, even though intuitively it seems right. To properly check the suitability of the Gaussian mixtures, one needs to perform a so-called goodness of fit test. This goodness of fit is what is going to be discussed in this section.

There are many available goodness of fit tests. In this report, two approaches will be used, a graphical based approach and hypothesis-testing based approach. The former will be performed by means of Q-Q plot and the latter by means of Relative Information based hypothesis testing.

B.3.1 Q-Q Plot Based Approach

A Q-Q plot is a graphical method that is used to compare two probability distributions by plotting their quantiles against each other. If the two distributions match each other (being similar), these points (of paired quantiles) in the plot shall lie around the 45° line (the \( y = x \) line).

To build the Q-Q plot for this case, sets of samples, each consisting of 900 samples, are generated from five mixture of Gaussian distributions using the parameters in Table 4.1. Then, each of them is compared with the corresponding data set in the form of Q-Q plot. To build the Q-Q plot, 150 quantiles are calculated from each of the data sets and sets of samples.
Figure B.7: Q-Q plot for bottom hole pressure (bhp).

Figure B.8: Q-Q plot for total flow rate at location 1 (Q1).
Figure B.9: Q-Q plot for total flow rate at location 2 (Q2).

Figure B.10: Q-Q plot for total flow rate at location 3 (Q3).
In all figures above, the red lines represent the $45^\circ$ line (or the $y = x$ line) and the blue lines represent the connected pair of quantiles. There are three plots in each figure. The first one is the Q-Q plot between the data set and a mixture of Gaussian. The second one is between the data set and a lognormal (for bhp) or logistic (for total flow rate). The last one is between the data set and a Gaussian distribution with mean 0 and variance 5%.

Observation on all five figures above reveals that indeed a Gaussian distribution with mean 0 and variance 5% is not suitable to represent any of the five data sets. The quantiles lines in all five figures deviate so far away from the $45^\circ$ line. This occurrence supports the evidence found in the previous section where this kind of noise overestimates the bottom hole pressure noise and underestimates the total flow rate noise. Further, from those five Q-Q plots, the log-normal distribution and the logistic distribution seem to perform
reasonably well. The corresponding quantiles lines lie relatively close to the 45° line. However, they still do not perform as good as the mixture of Gaussians.

Therefore, the Q-Q plot test concludes that the mixture of Gaussian distributions are the most suitable to represent the five data sets of measurements.

B.3.2 Relative Information Based Approach

One might still not be satisfied enough with a graphical approach, like Q-Q plots, without being able to quantify the goodness of fit in term of a number. One way to quantify the goodness of fit is by means of statistical test or hypothesis testing. There are many ways to perform hypothesis testing to a case; this case is no exception. For this project, hypothesis testing based on Relative Information will be used.

In short, Relative Information hypothesis testing tests the difference between two distributions with the null hypothesis being the difference is close enough to zero (with a predetermined confidence level). This null hypothesis is equivalent with just saying that the two distributions are very similar to each other. The characteristic of this test is that the difference between the two distributions is measured in term of the relative information as presented in equation (A.6). Then, using (A.7), the conclusion whether the null hypothesis should be rejected or not can be derived.

To be able to perform the calculation, the two probability vectors \( p = (p_1, \ldots, p_n) \) and \( q = (q_1, \ldots, q_n) \) must be defined first. To do that, each of the
data sets is divided into 45 parts (intervals) based on its 45 quantiles. Hence, each interval has 20 samples in it which means that the empirical probability of one random sample is located in any specified interval is $20/900$ or $1/45$. The first probability vector, $p$, can be built now. It consists of none other than the empirical probability of a random sample located in an interval. The first element if $p$ corresponds to the first interval, the second element to the second interval, and so on. There are 45 intervals; therefore the length of this vector is also 45, hence $n = 45$. Because each interval has the same probability, it means that $(p_1, \ldots, p_{45}) = (1/45, \ldots, 1/45)$.

The second probability vector, $q$, is built by combining the information obtained while building $p$ and the information regarding the parameter(s) of the tested distribution. For the mixture of Gaussian distributions that are used in this project, the parameters are presented in Table ???. This second probability vector shall consist of also 45 elements. It will represent the probability of a sample falls in each of the 45 interval based on the analytical (tested) distribution with a set of corresponding parameter(s). Hence, the same interval division as in $p$ must be used here. Of course the first element of $q$ also corresponds to the first interval, the second element to the second interval, and so on; just like in $p$. If the distribution matches the data set quite well, each element of $(q_1, \ldots, q_{45})$ should be close enough to $1/45$.

Now that both $(p_1, \ldots, p_{45})$ and $(q_1, \ldots, q_{45})$ are ready, the relative information of $p$ with respect to $q$, $I(p,q)$, can be calculated using equation (A.6). Multiplying this number with $2 \times 900$, with 900 being the number of
samples, based on (A.7), a number which is $\chi^2$-distributed with 44 degree of freedom is obtained. With this number, the $p$–value can be calculated. This $p$–value represents the probability to obtain a test statistics that is at least as extreme as $1800l(p,q)$. Hence, the null hypothesis is rejected if the $p$–value is less than a significant level $\alpha$. For this project, $\alpha$ is taken to be 5%.

The results of performing this Relative Information hypothesis testing to each of the five measurement variables are presented in Table B.1 below.

Table B.1: $p$–values of testing each of the variables with three possible distributions by means of Relative Information hypothesis testing

<table>
<thead>
<tr>
<th>Variables</th>
<th>Mixture of Gaussian</th>
<th>Lognormal/Logistic</th>
<th>Gaussian $\mathcal{N}(0,5%)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>bhp</td>
<td>0.0348</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Q1</td>
<td>0.7608</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Q2</td>
<td>0.9201</td>
<td>0.0088</td>
<td>0</td>
</tr>
<tr>
<td>Q3</td>
<td>0.0600</td>
<td>0.0001</td>
<td>0</td>
</tr>
<tr>
<td>Q4</td>
<td>0.0025</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table B.1 above states that a Gaussian noise with mean 0 and variance 5% is strongly rejected to fit any of the data sets. Again, this conclusion goes along well with the same conclusion yielded from Q-Q plot based goodness of fit and by manual histograms comparison. Log-normal or logistic distributions are also rejected in all five of the cases, though not as strongly as Gaussian especially in total flow rate at location 2 and 3. Mixture of Gaussian

\footnote{there is no connection whatsoever between the name $p$–value and the first probability vector which coincidentally is also presented as $p$.}
B.4 : CONCERNS AND POSSIBLE FURTHER QUESTIONS

distributions perform marginally better than the other distributions. The $p$-values are much higher in all cases; even though for bottom hole pressure and total flow rate 4, the $p$-values are technically below the significant level of 5%. For bottom hole pressure, though 3.48% is technically below 5%, it is not that much lower. Hence to some degree, it is still acceptable to not reject the null hypothesis. For total flow rate 4, the $p$-value is 0.25%. However, even with this $p$-value, mixture of Gaussian for this case still outperforms logistic distribution and Gaussian distribution; hence mixture of Gaussian should still be used. This might be caused by the lack of samples used during the distribution fitting process.

Two approaches of goodness of fit tests have both concluded that mixtures of Gaussian distributions are suitable enough to represent the five data sets of noise. Also, both confirm that the mixture of Gaussian distributions fit the data sets much better than log-normal/logistic or the Gaussian distributions. Therefore, the measurement values for this project will be generated using noise distributed according to mixtures of Gaussian distributions with parameters as presented in Table 4.1.

B.4 Concerns and Possible Further Questions

There are two assumptions that are used in the measurements generation process while determining the distributions of the measurement noises as described in Section B.1 above:
1. The distribution of noise is the same over all eight time steps.

2. The true value of each observable variables is well approximated by the mean (average) of the 900 data of that particular variable.

The first assumption leads into the action of taking the distributions of the five measurable variables at the first time step to generate the distributions of the noises (using the second assumption) of all time steps. This action heavily favors the first time step. Next, if in reality the first assumption is not quite correct, this may lead to error in noises generation of the next time steps because not-so-correct distributions are then used.

Table B.2: $p$–values of the statistical tests testing the first assumption's validity on every time step.

<table>
<thead>
<tr>
<th>Var</th>
<th>Time 1</th>
<th>Time 2</th>
<th>Time 3</th>
<th>Time 4</th>
<th>Time 5</th>
<th>Time 6</th>
<th>Time 7</th>
<th>Time 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>bhp</td>
<td>0.0348</td>
<td>0.0011</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>Q1</td>
<td>0.7608</td>
<td>0.0541</td>
<td>0.0793</td>
<td>0.0025</td>
<td>0.1982</td>
<td>0.0410</td>
<td>0.1714</td>
<td>0.0010</td>
</tr>
<tr>
<td>Q2</td>
<td>0.9201</td>
<td>0.7327</td>
<td>0.6590</td>
<td>0.4110</td>
<td>0.2480</td>
<td>0.3319</td>
<td>0.3474</td>
<td>0.0003</td>
</tr>
<tr>
<td>Q3</td>
<td>0.0600</td>
<td>0.1845</td>
<td>0.2273</td>
<td>0.3546</td>
<td>0.1149</td>
<td>0.1152</td>
<td>0.4214</td>
<td>0.0341</td>
</tr>
<tr>
<td>Q4</td>
<td>0.0025</td>
<td>0.0167</td>
<td>0.0038</td>
<td>0.2232</td>
<td>0.0710</td>
<td>0.0409</td>
<td>0.0129</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

Unfortunately, this seems to be the case. Running the program simsim further for all eight time steps without any measurement updates, the distribution of noise of the five measured variables for every time step can be obtained (by using the second assumption). Then, a statistical test identical to the relative information based approach goodness of fit as presented in the
previous section is performed to test whether the distributions in Table 4.1 are suitable to represent the measured variables for each of the time step or not. The results in term of their \( p \)-values are presented in Table B.2.

Table B.2 indicates that the noise distribution of the first time step gradually becomes unsuitable to represent the distribution of noise of the further time steps. The further a time step is from the first time step, its noise becomes more unsuitable to be approximated by the noise distribution of the first time step. Running the program \texttt{simsim} further in time and extending Table B.2 above to time step 12 (day 720) confirms this conclusion even further. The \( p \)-values for these further time steps are available in Table B.3.

\textbf{Table B.3:} \( p \)-values of the statistical tests testing the first assumption’s validity on further time steps.

<table>
<thead>
<tr>
<th>Var</th>
<th>Time 1</th>
<th>...</th>
<th>Time 7</th>
<th>Time 8</th>
<th>Time 9</th>
<th>Time 10</th>
<th>Time 11</th>
<th>Time 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>bhp</td>
<td>0.0348</td>
<td>...</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>Q1</td>
<td>0.7608</td>
<td>...</td>
<td>0.1714</td>
<td>0.0010</td>
<td>0.0037</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>Q2</td>
<td>0.9201</td>
<td>...</td>
<td>0.3474</td>
<td>0.0003</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>Q3</td>
<td>0.0600</td>
<td>...</td>
<td>0.4214</td>
<td>0.0241</td>
<td>0.0206</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>Q4</td>
<td>0.0025</td>
<td>...</td>
<td>0.0129</td>
<td>0.0001</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

We can see that the further a time step is from the first one, indeed the noise distribution at time step 1 is getting more unsuitable to be used to model the noise distribution of that particular time step. One may be curious in seeing the actual difference in the distribution between the first time step and a further time step. Figure B.12 below provides the five noise distributions on time step
1 (day 60), time step 8 (day 480), and time step 12 (day 720).

**Figure B.12**: Comparison of the noise distribution between different time steps.

(a) Time step 1 (day 60)  
(b) Time step 8 (day 480)  
(c) Time step 12 (day 720)

The changes in distribution are, indeed, observable. The bottom hole pressure appears to become more and more uncertain as time goes on, as indicated by the "expansion", shape-wise, of the noise distribution of this variable. Hence, taking the noise distribution of bhp to be the same as time
step 1 at the further time steps is underestimating the uncertainty. Quite the contrary, the four total flow rates appear to become less and less uncertain as time goes on. The noise distributions seem to have tendency to become "thinner". Therefore, taking the noise distribution of the four total flow rates to be the same as time step 1 at the further time steps is overestimating the uncertainty. This simple visual observation supports the conclusion from the statistical test above, as presented in Table B.2 and Table B.3.

Even so, as already pointed out in [8], the (absolute) correlations between variables tend to become smaller and smaller as time goes on, after the first time step. Lower (absolute) correlation means weaker effect (or less sensitivity) one variable has on another. Therefore, even though the first assumption is heavily criticized here, taking different noise distributions in each time step might not improve the performance of both (EnKF and NPBN) methods significantly either. Hence, it may not be urgent for one to be too bothered with this and go on more trouble by generating unique noise distributions for every time step.

Stepping to the second assumption, the validity of the second assumption also needs to be investigated. It is assumed that the true value is well approximated by the mean of the 900 data. By taking the second assumption, the noise is constructed as the difference between each of the data with this mean. However, if the true value is actually nowhere near the mean, the generated noises might not illustrate the actual ones very well. This may further lead into error in the measurements generation process.
Figure B.13: Histograms of the five measurable variables with the red bin containing the mean value and the yellow bin containing the truth

Figure B.13 above illustrates the locations of the mean of the synthetic data (red bin) and the true value (yellow bin) in the empirical distribution of each of the five noises at time step 1. It looks a bit concerning with regard to the validity of the second assumption. The true values seem to be located quite far away from the mean of the 900 samples, especially in the case of the four total flow rates.

Let $X_{bhp}, X_{Q_1}, \ldots, X_{Q_4}$ denote random variables each corresponding to each of the five observed variables and $t_{bhp}, t_{Q_1}, \ldots, t_{Q_4}$ denote each of their true
values. To look deeper into the situation, the empirical probability of a corresponding random variable to be less than its true value can be calculated, based on the 900 samples. This probability also represents to which percentile each of the true value belongs to in the corresponding distribution. The values of this probability are the following:

\[ P(X_{bhp} < t_{bhp}) = 0.7844 \quad P(X_{Q3} < t_{Q3}) = 0.05 \]
\[ P(X_{Q1} < t_{Q1}) = 0.9344 \quad P(X_{Q4} < t_{Q4}) = 0.9256 \]
\[ P(X_{Q2} < t_{Q2}) = 0.0422 \]

Basically, these numbers are just another way of representing the same thing as in Figure B.13 above. These numbers, just like the figure, look quite concerning.

Even though Figure B.13 and the probability values do not look very promising, there is still not enough evidence to conclude anything. Further investigation regarding the validity of this second assumption still needs to be performed.
As mentioned in Chapter V, the set of measurements with Gaussian noise as used in [8] is reused to produce a comparison for the obtained result when the non-Gaussian noises, as described in the beginning of Chapter IV and Appendix B, are used. The result in term of the RMSE has been provided in Chapter V through Figure 5.12. However, details of the experiment with Gaussian noise have been left out from discussion in that chapter.

This Appendix consists of the more detail process of performing the experiment using this set of measurements with Gaussian noise as well as a little bit of the analysis of the outcomes.

C.1 The Measurements

In [8], all five measured variables are disturbed noises from the same distribution: a Gaussian noise with mean zero and variance 5%. In other words, the noise is $V \sim \mathcal{N}(0, 5\%)$. There is no reason to use different set of measurements from the one used in [8]. Therefore, the same set of measurements shall be used in this Appendix.
The generated measurements are presented in the form of Figure C.1 below. The black lines and circles correspond to the truth while the blue stars correspond to the measurements.

**Figure C.1:** The generated measurements with Gaussian noise as used in [8]

---

**C.2 Applying the NPBN Method**

With this set of measurements, the same procedure of the NPBN Method as described in Section 5.2.1 is performed. To provide fair comparison, the same nine $6 \times 6$ grid blocks, as presented in Figure 5.6, are used. Running exactly the same procedure (but with different measurements) yields the following result as presented in Figure C.2.

From Figure C.2, it turns out that the NPBN Method produces a rather smooth approximation of the permeability field when Gaussian noise is used in the measurements generation.
**Figure C.2**: The estimated permeability field of the entire reservoir by NPBN at all time steps with Gaussian Measurement Noise

**C.3  Analysis and Comparison with EnKF**

Comparison between the obtained permeability field as presented in Figure C.2 above and EnKF when Gaussian noise is used for time step 4, 7, and 8 (day 240, 420, and 480 respectively) is presented in Figure C.3 below. The RMSE of the two approaches, which is already provided in Figure 5.12, is represented in Figure C.4.

Analysis upon Figure C.3 reveals that both methods indeed produce rather smooth approximations of the permeability field. This is not surprising coming from EnKF. Visual observation yields a promising result for the NPBN Method as it seems to recover the truth better than EnKF. Both methods seem to underestimate the permeability at the permeable part of the reservoir, but
the NPBN Method overestimates the impermeable part more than EnKF does. The NPBN Method also seems to distinguish permeable area and impermeable area in the field stronger than in EnKF. This can be seen from the relatively low amount of "medium-permeability" cells (indicated with the color yellow and light green) in the approximated field of NPBN. Most cells in the NPBN approximations are either red (permeable) or blue (impermeable). In some ways, this behavior is actually quite similar to the behavior of the truth.
**Figure C.4:** RMSE of the estimated permeability field of the entire reservoir EnKF and saturated NPBN with Gaussian measurement noise

However, the RMSE, once again, states the otherwise. Based on the RMSE, the EnKF method performs better than the NPBN method. However, the difference between the two is not that much (only to the magnitude of 0.1 or less). Both RMSEs also seem to behave quite constantly through time.

An analysis based on the absolute error of both methods (the absolute difference between the truth and the approximation) of each cell in the reservoir field can also be performed. Figure C.5 below presents the absolute difference of each cell with both methods.

Figure C.5 indicates that cells with larger error (yellow cells) are quite spread throughout the entire reservoir for both methods. As has been observed in the previous experiment with non-Gaussian measurement noise (Figure 5.11 in Chapter V), both methods behave similarly in regard to the locations of cells
Figure C.5: The absolute difference between the truth and the approximation of each cell in the reservoir with both methods (NPBN and EnKF) with Gaussian noise with relatively large error. It is just that in the location where large error occurs, NPBN overestimates (or underestimates) the truth much stronger than EnKF. This can be seen by the stronger yellow in the NPBN figures. As a result, this brings the RMSE, which is the average of the absolute error of all cells in the field, of the NPBN Method a little bit higher than EnKF.

Another measure of performance (forecast or prediction) to test the results of the NPBN method and the EnKF method with this set of measurements has also been performed. The analysis and comparison between the two results are already presented in Chapter VI.
C.4 Evaluation of the Cubic Interpolation Method

This section is meant as a supplement to the analysis that is performed in Section 5.4 in Chapter V regarding the evaluation of the chosen interpolation method in this project, which is a simple cubic interpolation. As presented in that section, the evaluation is done by observing the consistency between the permeability of the interpolation cells and the blocks cells. The consistency itself is examined by looking at the deviation between the approximated permeability and the corresponding truth of one cell to another.

The left column of Figure C.5 provides visual observation for this purpose. Figure C.6 below provides the same plots as the left column of Figure C.5 but with black lines to help distinguishing the blocks and the interpolation cells.

**Figure C.6:** The absolute difference between the truth and the approximation of each cell in the reservoir with the NPBN method with Gaussian noise

Based on Figure C.6, visually, there appears to be no problem with the
cubic interpolation method. The absolute error of the interpolation cells and the blocks cells appears to behave quite similarly with no distinguishable pattern between the two nor observed "strange" behavior.

To quantitatively evaluate this consistency, just as performed in Section 5.4, RMSE of the blocks cells, the interpolation cells, and the "relevant" blocks cells are calculated separately. The result is presented in Figure C.7 below.

**Figure C.7**: The RMSE of the interpolation cells, blocks cells, and "relevant" block cells

![Graph showing RMSE comparison](image)

As encountered in Section 5.4, directionality-wise, all RMSEs agree with each other. However, there is noticeable gap between the RMSE of the entire blocks cells (blue line) and the RMSE of the interpolation cells (red line) where the RMSE of the entire block cells is always bigger than the RMSE of the interpolation cells with magnitude of around 0.15 – 0.2. Again, the cause of this phenomenon is some "non-contributing" cells whose approximations are way off from their truths compared to the other; thus, bringing the RMSE of the
blocks cells up. These non-contributing cells with higher absolute error are visible in Figure C.6, for example in the middle west and middle south part of the field. For this reason, the analysis of the consistency uses the third RMSE line (the black line) instead, which is the RMSE of the "relevant" cells in the grid blocks, i.e., cells that are involved in the cubic interpolation performed in the experiment.

Directionality-wise and value-wise, the relevant blocks cells’ RMSE appears to behave much more similarly to the interpolation cells’ RMSE. Even though for this Gaussian case the RMSE of the relevant blocks cells is always higher than the RMSE of the interpolation cells, the difference is not that much ("only" with magnitude of less than 0.5). This is a good news because it means that the interpolation cells behave hand in hand with the relevant blocks cells. Therefore, for this Gaussian case, it appears that a simple cubic interpolation also performs reasonably well.