Evaluation of a Machine Learning methodology for spatio-temporal induced seismicity forecasts within the Groningen field

Machine Learning

IBM and Shell Research

F. Lanz, K. Bisdom, E. Barbaro, J. Limbeck, T. Park, C. Harris and K. Nevenzeel

Date January 2019

Editors Jan van Elk & Dirk Doornhof

General Introduction

The prime objective of the Hazard and Risk Model is to forecast seismic hazard and risk for the local population living on and around the Groningen gas field and in particular the risk metrics of the Meijdam safety norm. This model has been extensively documented.

As an alternative approach for forecasting production induced seismicity, a methodology based on Machine Learning (ML, a branch of artificial intelligence in the field of computer science) to forecast seismicity for the Groningen field was explored, in the context of the measurement and control protocol.

This report documents the Machine Learning study into developing a methodology for spatiotemporal induced seismicity forecasts within the Groningen field. Machine learning has been used also in other studies of the earthquake in the Groningen field (Ref. 1 to 3).

References

1. J. Limbeck, F. Lanz, E. Barbaro, C. Harris, K. Bisdom, T. Park, W. Oosterbosch, H. Jamali-Rad and K. Nevenzeel, Evaluation of a Machine Learning methodology to forecast induced seismicity event rates within the Groningen Field.
2. Seasonality analysis for induced seismicity event rate time series within the Groningen Field Machine Learning, IBM and Shell Research, Park T., H. Jamali-Rad, W. Oosterbosch, J. Limbeck, F. Lanz, C. Harris, E. Barbaro, K. Bisdom & K. Nevenzeel, October 2018.
3. A Simulation Study into the Detectability Threshold for Seasonal Variations in Earthquake Occurrence Rates within the Groningen Field - Machine Learning, T. Park and K. Nevenzeel, IBM/Shell, 2019.

NAM

Title	**Evaluation of a Machine Learning methodology for spatiotemporal induced seismicity forecasts within the Groningen field** **Machine Learning**	**Date**	January 2019
		Initiator	NAM
Autor(s)	**F. Lanz, K. Bisdom, E. Barbaro, J. Limbeck, T. Park, C. Harris and K. Nevenzeel**	**Editors**	Jan van Elk and Dirk Doornhof
Organisation	IBM and Shell Research	**Organisation**	NAM
Place in the Study and Data Acquisition Plan	Study Theme: Seismicity Forecasting Comment: The prime objective of the Hazard and Risk Model is to forecast seismic hazard and risk for the local population living on and around the Groningen gas field and in particular the risk metrics of the Meijdam safety norm. This model has been extensively documented. As an alternative approach for forecasting production induced seismicity, a methodology based on Machine Learning (ML, a branch of artificial intelligence in the field of computer science) to forecast seismicity for the Groningen field was explored, in the context of the measurement and control protocol. This report documents the Machine Learning study into developing a methodology for spatiotemporal induced seismicity forecasts within the Groningen field. Machine learning has been used also in other studies of the earthquake in the Groningen field (Ref. 1 to 3).		
Directliy linked research	(1) Gas Production (2) Machine Learning (3) Measurement and Control Protocol		
Used data	KNMI Earthquake catalogue Groningen gas production data		
Associated organisation	NAM		
Assurance			

Evaluation of a Machine Learning methodology for spatiotemporal induced seismicity forecasts within the Groningen field

by

F. Lanz (IBM Services)

K. Bisdom (GSNL-PTX/S/RM)

E. Barbaro (IBM Services)

J. Limbeck (GSNL-PTX/D/S)

T. Park (GSNL-PTX/D/S)

C. Harris (SUKEP-UPO/W/T)

K. Nevenzeel (IBM Services)

Executive Summary

Business purpose:

Decades of gas production caused induced seismicity in the Groningen gas field, located in the Northern part of the Netherlands. The capability to forecast induced seismicity depending on production strategy is an essential element of the Probabilistic Seismic Hazard and Risk Assessment (PSHRA) for the exposed population. As part of the Study and Data Acquisition Plan in the context of the Measure and Control Protocol, this study evaluates a methodology based on Machine Learning (ML, a branch of artificial intelligence in the field of computer science) to forecast production induced seismicity for the Groningen field. This study is a direct extension of the methodology developed in Limbeck et al. (2018), in two ways:

- *The range of validity is extended* by (i) validating forecast performance on a hold-out set, (ii) qualitatively validating the similarity between short-term (1-3 months) and long-term (1-5 years) forecasts and (iii) based on synthetic "ultimate states" (assumptions about reservoir conditions long after shut-in), ensuring that our methodology forecasts an approximately zero seismicity rate long after shut-in.
- *The range of applicability is extended* from temporal (event rate) forecasts to spatiotemporal forecasts.

Methodology Evaluation:

The range of validity of the event rate model of Limbeck et al. (2018) is extended by the following:

- ML models can capture the negative trend in seismicity observed during the 2013-2016 hold-out period, like the depletion thickness and vertical strain thickness Moving Average baselines.
- Model selection based on short-term (1-3 months) or long-term (1-5 years) performances in terms of mean bias error remains unchanged, despite the decrease in performance observed for longer-term forecasts.
- The model performance has been evaluated using a Poisson likelihood. The Poisson likelihood of the mean forecast model is analytically the same as the likelihood of a bin-wise constant Poisson point process model. Hence, we have constructed, examined, and evaluated a wide range of probabilistic models.
- The pipeline can generate a probability density function if used with the Poisson link. The distribution, per field, is Poisson with the parameter that is predicted.

The range of applicability is extended to spatiotemporal:

- ML models can now forecast event rates per cell over a freely defined grid of adjustable resolution. Hence it is now possible to investigate the regional response of seismicity rates to certain production scenarios. The aggregated performance of these forecasts is comparable in the temporal domain to that of the event rate report (Limbeck et al., 2018).
- The spatiotemporal performance of the ML models, in particular, that of random forests and support vector machines, is significantly better than that of the Depletion Thickness and vertical Strain Thickness Moving Average baselines if evaluated using the likelihood ratio method for combined spatiotemporal performance evaluation.
- ML models can now make use of spatially varying dynamic and spatial features per grid cell rather than using averages over the entire reservoir, or predefined regions to then make forecasts at the individual grid cell level.

Limitations:

The methodology developed in this study has the following main limitations:

- The methodology generates spatiotemporal induced seismicity forecasts. An important additional seismicity dimension, event magnitude, is not considered beyond a minimum magnitude. As such, a separate magnitude model would be required to complement the current spatiotemporal model.
- Part of the spatiotemporal input data and the "ultimate states" are based on reservoir flow model history matches and forecasts. Consequently, the ML results are intrinsically linked to the accuracy of the reservoir model. Furthermore, the current ML pipeline is only guaranteed to forecast negligible seismicity long after the field shut-in within a specific range of ultimate states (and given no-depletion and steady-state situations).

Table of Contents

Table of Figures

Table of Tables

1. Introduction: Overview, Earlier Work & Study Goals

1.1. Earlier Induced Seismicity Forecasts for Groningen

Discovered in 1959, with an initial recoverable reserve estimate of 2900 billion m^3 gas, the Groningen gas field is amongst the largest gas fields in the world (TNO, 2017). Production commenced by NAM in 1963. By 2015, around 2000 billion m^3 have been produced. The reservoir of the Groningen field is the Upper Rotliegend Group of Early Permian age, consisting of porous sandstone and located at a depth between 2600 m and 3200 m, with the water zone around 3000 m deep. The gas in the reservoir is sealed by a thick impermeable salt and anhydrite layer of the overlying Zechstein Group, as depicted in Figure 1. The Groningen field has several fault systems with around 1500 known faults, whose existence does not impact permeability in a significant way.

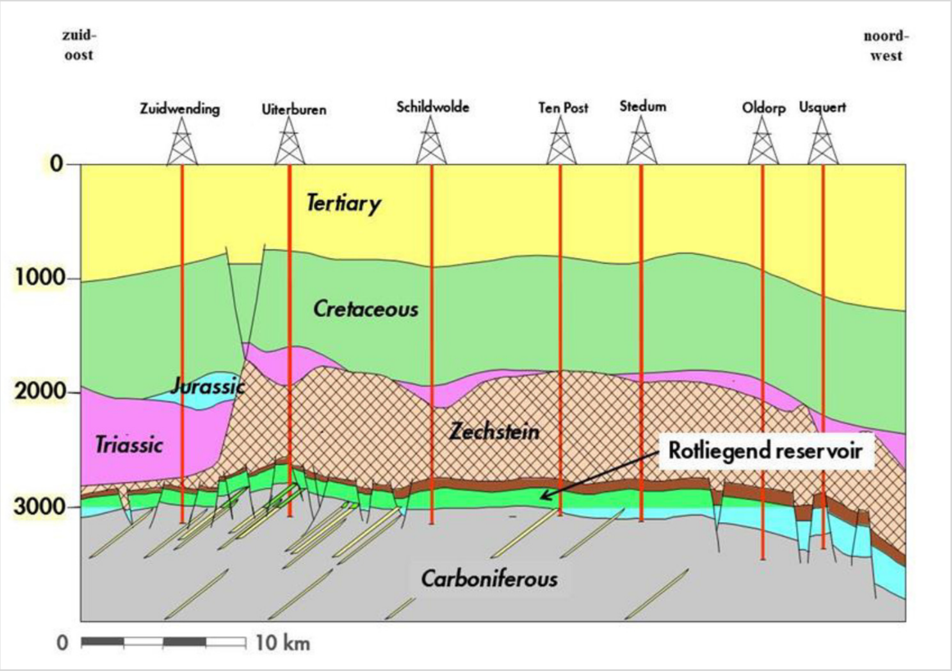


Figure 1: Geological cross-section of the Groningen field (NAM, 2016)

Following decades of gas production, the historically aseismic region experienced induced earthquakes for the first time in 1991. The frequency and intensity of earthquakes increased steadily to around ten or more earthquakes per year with a magnitude equal to or larger than 1.5 as of 2003. Following an earthquake of magnitude 3.6 on the Richter scale with an epicenter in the village of Huizinge in 2012, a Study and Data Acquisition Plan (Nederlandse Aardolie Maatschappij, 2016a) was put in place to better understand how gas production at reservoir depth affects safety at the surface, and to test the effectiveness of mitigation measures. This led to an integrated Probabilistic Seismic Hazard and Risk Assessment (PSHRA) starting from gas production, sequentially followed by compaction, seismicity, ground motion, exposure, building strengthening and finally risk and safety of inhabitants, see Figure 2.

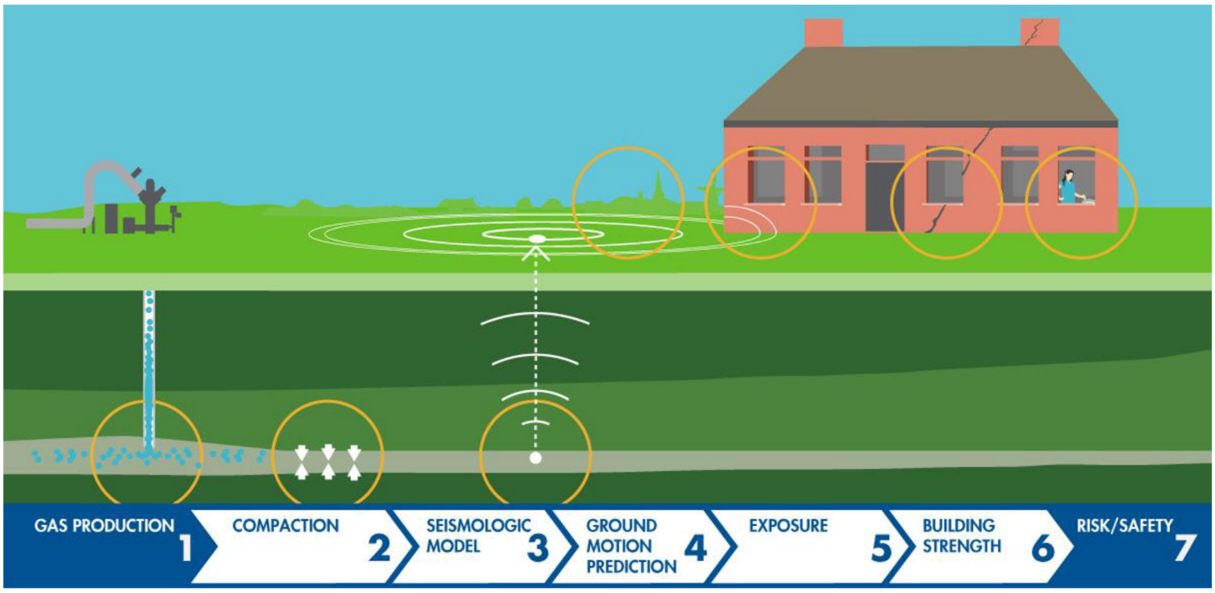


Figure 2: Causal chain from gas production to the safety of people in or near a building (NAM, 2016)

The Study and Data Acquisition Plan encompasses this study in the context of the Measure and Control Protocol (Nederlandse Aardolie Maatschappij, 2016b). The focus of this study is seismologic modelling (element 3 of PSHRA) and forecasting.

Here, we explicitly use the word "forecast" instead of "prediction" in the context of seismicity as within seismology both terms refer to different approaches to gain more quantitative insights in future seismicity (Marzocchi and Zechar, 2011). Predictions refer to high confidence statements about the location, timing, and magnitude of a future seismic event, whereas forecasts are used to describe quantitative statements about future event statistics. This study and all other studies to date which have been able to provide reliable statements (DeVries et al., 2018) about future seismicity in the Groningen field are forecasts – due to limitations in both available data and human understanding of geophysical mechanics triggering earthquakes. To avoid any confusion between audiences versed in different scientific fields, we observe that within the field of machine learning the term "prediction" is applied in a broader sense than in seismology, as in the context of statistics, 'prediction' refers to the process of determining the magnitude of statistical variates at some future point of time (Marriott et al., 1990). In a few sections on machine learning (e.g., chapter 6), some instances of the word 'prediction' are used for machine-learning oriented expositions in line with common practice in that community.

1.2. Machine Learning Seismicity Forecasts Elsewhere

Machine Learning (ML) is a branch of statistical computer science which over the last decade has been applied successfully in a wide variety of domains (Jordan and Mitchell, 2015). In the context of physics, ML allows for experimental control over a vast number of factors (Langley, 1988) making it suitable for physical modelling (Liu, 2018). Due to their nature, ML models can perform well in situations where underlying processes are not fully understood (Melnikov et al., 2018) or are complex (Carrasquilla and Melko, 2017). The use of ML concomitantly with information about physical phenomena is called theory-guided machine learning (TGML; Karpatne et al., 2017). TGML allows for the introduction of scientific consistency, knowledge, and protocols in ML models. Also, these models often are more (physically) interpretable than a standard ML model. Despite being a relatively new area of research, many traditional scientific fields are taking advantage of this paradigm. ML and physical knowledge have proven to accurately forecast the behaviour of a large spatiotemporal chaotic physical system where the mechanical description of the dynamics is limited (Pathak et al., 2018) – accurate forecasts up to eight times the regular forecast horizon could be achieved. Deep learning applied to forecast the locations of aftershocks yielded better results compared to a physics-based model and provided insights as to the relative importance of different physical parameters in aftershock localization (DeVries et al., 2018). Several studies have shown that pure ML models without the guidance of domain knowledge may lead to incorrect models (e.g., Wagner and Rondinelli, 2016).

Given that, ML seems a viable tool to complement physical and statistical seismicity modelling efforts and has become increasingly popular for seismic analysis. Three main ways in which ML has been applied within seismicity studies are (i) earthquake identification (Perol et al., 2017; Rouet-Leduc et al., 2017), (ii) catalogue-based seismicity forecasting (Panakkat and Adeli, 2009) and (iii) model parameter inference (e.g. the Gutenberg-Richter b-value, Asencio-Cortés et al., 2016). On top, we note that in the context of PSHRA machine learning is used already for optimisation of the production distribution over the Groningen field to reduce seismicity, see Nederlandse Aardolie Maatschappij (2017). A non-exhaustive review aimed to shed light on the role of ML in seismicity analyses is given below.

Earthquake identification is often made by acoustic or ground vibration wavelet analysis of seismic detection sensors. A recent study of Rouet-Leduc et al. (2017) estimated time to fault failure based on a local moving time window signal emitted by laboratory faults. In their study, a vast number of potential features was computed for every single time window (e.g., 0th/1st/2nd order statistics), and the most useful features are used in a Random Forest (RF) model achieving a high determination coefficient $(R^2 = 0.89)$. Interestingly, the RF model accurately forecasted failure not only when failure is imminent but throughout the failure cycle. Features which quantify signal amplitude distribution (e.g., variance and higher-order moments) are highly effective for forecasting, despite their high variability. The authors acknowledge that this effort remains academic, however. We note that if a connection between seismic wavelets and fault properties could be identified, it would help the development of deterministic geomechanical models. Perol et al. (2017) employed a scalable Neural Network to consistently detect and localize earthquakes based on a single waveform. They claim to detect 20 times more earthquakes than previous methods, which is important to make seismic catalogues more complete, in turn improving Hazard and Risk Assessments for induced seismicity in Oklahoma. A possible caveat of this study is the fact that it requires a pre-existing history of catalogued seismicity and is therefore less suitable for areas of lower activity or more recent instrumentation. Ramirez and Meyer (2011) used a kernel ridge-regression algorithm to study seismic phases from seismic recordings. Their method consists of a multi-scale potential predictor extraction on low-dimensional manifolds. In addition, they merged their regression scores across the potential predictor manifolds. The authors concluded that their algorithm could correctly forecast around 75% of the classification rates for seismic data collected in the US during 2005 and 2006.

Seismicity forecasting via earthquake catalogues uses dates, locations, and magnitudes of earthquakes to forecast future earthquakes. Panakkat and Adeli (2009) forecast earthquake times and locations for earthquakes for magnitude $M \geq 4.5$ using a wide variety of Neural Networks. These networks were offered multiple seismicity indicators derived from an earthquake catalogue (e.g., Gutenberg-Richter's b-value, the average magnitude of the last n events, the mean-square deviation about the regression line based on Gutenberg-Richter's inverse power law curve for n events, etc.) as parameters. The magnitude of their error in forecasting the epicentral location of high magnitude events was always within 20-40 miles, which the authors claim to be useful for emergency management and planning. Rouet-Leduc et al. (2017) utilized an RF algorithm on labinduced earthquakes to investigate hidden signals preceding the events. They suggest that previous literature only based on earthquake catalogues may be incomplete.

Asencio-Cortés et al. (2016) proposed a meta-analysis setup to find out the best set of parameters and concluded that it is possible to use ML techniques to calculate the b-value. Last et al. (2016) focused on understanding whether future maximum earthquake magnitude exceed the median of maximum yearly magnitudes (for the same region). Several ML algorithms used here are also utilized in their study (Decision Trees, K Nearest Neighbours, Support Vector Machines and Neural Networks). Their results point out to a variant of a decision tree as the most accurate ML model. Their features are based on observed earthquake catalogues and derived relations, e.g., the Gutenberg-Richter law.

1.3. Open questions from Event Rate report

The goal of the event rate report Limbeck et al. (2018) was to develop an ML-based methodology to forecast production induced seismicity event rates for the Groningen field. That methodology allowed probing of a wide variety of possible linear and non-linear combinations and interaction terms between physical variables without assuming a priori knowledge on the nature of the relationships between these variables. A two-step approach was employed: a factorial experimental setup followed by meta-analysis (analysis of the effectiveness of the experimental setup) was used to select robust and relatively well-performing models and meta parameters. The selected models and meta parameters were used for seismicity event rate forecasts.

The event rate forecasts were evaluated in three ways: (i) quantitatively; (ii) qualitatively and (iii) the range of validity. Quantitatively, we noted that with the data used in that setup, in general, the ML models were not statistically significantly better than baseline models. Qualitatively, we observed that for the Winningsplan 2016 (Nederlandse Aardolie Maatschappij, 2016c), the selected models and meta parameters forecast a relatively stable seismicity event rate for the coming five years, in line with the default PSHRA forecasts. For the average production scenario announced by the Ministry of Economic Affairs and Climate in March 2018 (Ministry of Economic Affairs and Climate Policy, 2018) (hereafter the post-March 2018 average production scenario) model behaviour diverged qualitatively. Both in the Event Rate report (Limbeck et al., 2018) and the current report, ML models are divided into two classes, namely (i) extrapolating and (ii) non-extrapolating. According to our definition (see Glossary) non-extrapolating models are unable to forecast outside their target range of calibration – see also Kneale and Brown (2018), while theoretically extrapolating models can – see Appendix A for details.

The range of validity of the methodology described in Limbeck et al. (2018) was influenced by three key aspects, which all might have played a role in the unphysical forecasts of, in particular, the non-extrapolating models for the post-March 2018 production scenario.

- First, the model evaluation strategy was geared towards maximum statistical power, thereby decreasing uncertainties and hence improving the ability to statistically distinguish between the forecasts of various models. Models were evaluated and retrained after each step forward when uncertainties were still relatively small. An implied (although not technically fundamental) assumption is that one step forward (1-3 months) forecast performance is indicative for many steps forward (1-5 years) forecast performance. This assumption was not always satisfied in Limbeck et al. (2018) and could lead to selecting models which performed well on the short term (1-3 months) but not on the long term (1-5 years).
- Second, physical variables like P and HCT are monotonically evolving features and thus are guaranteed to go out-of-bounds of the convex hull of the past values of the feature set. Given the fact that some of these features were expected (and for RFs were shown) to play a major role, this posed a challenge for non-extrapolating models which might result in

unphysical behaviour, for example by forecasting a constant increase in seismicity after production rates decreased to zero.

- Third, the model and evaluation selection criteria measured concordance with physical constraints only indirectly – through concordance with past (physically-sound) observations.

Three concrete steps which could mitigate the limitations on the range of validity are:

- Investigate usage of long-term (1-5 years) forecasts to validate model performance, instead of the short-term (1-3 months) forecast performance evaluations used in this study.
- Analyse the degree to which the event rate forecast feature values exceed the convex hull of historical feature values and analyse the impact thereof.
- Extend the model evaluation and selection criteria with rules encoding ultimate states about reservoir conditions long after shut-in from multiple reservoir-model forecasts running up to the year 2100.

Regarding the definiteness of the conclusions reported in Limbeck et al. (2018), we note that in light of the (from an ML perspective) relatively limited number of events, all data available at the start of the study was used for model meta-analysis (and thus model selection). Although various safeguards were put in place, forecast performance estimates might be on the optimistic side, and a hold-out set would be required to validate these estimates. Ideally, the training and testing period of the hold-out set ends around a moment that the production strategy changes. Models which forecast a change in seismicity following a change in production strategy probably capture underlying mechanisms driving seismicity better. Two approaches were identified. First, a validation set will be obtained naturally over time. An appropriate cut-off moment between training/testing and validation might be before the post-March 2018 production scenario is enacted. The disadvantage is that it will take quite some years to obtain a large enough validation set. For this reason, a second approach, training/testing the model up to the production shut-ins following the Huizinge earthquake in 2012 and using the remaining years for validation was proposed. A disadvantage of this second approach is that it roughly halves the number of events, which might impact our ability to discriminate between models. Despite that, the second approach was chosen because it could be directly implemented.

1.4. This Study: Machine Learning-based Spatiotemporal Seismicity Forecasts for Groningen

This study is a direct continuation of the Event Rate work initiated in Limbeck et al. (2018). The primary goals of this continuation study are to:

i) Address the outstanding limitations of the temporal analysis listed in the previous section, most notably addressing the unphysical model behaviour;

ii) Extend the methodology for ML-based induced seismicity event rate forecasts for the Groningen field to also include a spatial forecasting component to achieve spatiotemporal forecasts over the field.

The study uses and extends the ML forecasting pipeline as developed in Limbeck et al. (2018) as shown high-level in Figure 3, where the elements stand for the following:

- **Data sources** are selected, including earthquake event catalogues, static, dynamic and geomechanical reservoir model data and seismic data, and potential predictors (features) are generated from them.

- **Meta-parameters** define the experimental setup within which models are trained and do forecasts. Our meta-parameters can be divided into two sets: (i) those related to our forecasting target like minimum magnitude (ii) those describing our experimental setup, grid size and smoothing bandwidth.
- **The model evaluation strategy** is based on a walk-forward evaluation strategy with spatial cross-validation with two standard evaluation metrics and an additional likelihood method[1], including the associated standard error estimates for the standard metrics.
- **Machine learning models** are generated for each experiment that is carried out - loosely based on empirical performance studies at least the following algorithms are tested: RFs, KSVMs, KNNs, and GLM variants.
- **Meta-Analysis** is employed on top of factorial runs of experiments to analyse the impact of model and meta-parameter choices on forecast performance. Based on the meta-analysis robust models with meta-parameter sets are selected for each target. These models are subsequently trained and used for seismicity forecasts.
- **Automated orchestration framework** enables, as the title says, automated runs of a factorial setup of experiments.

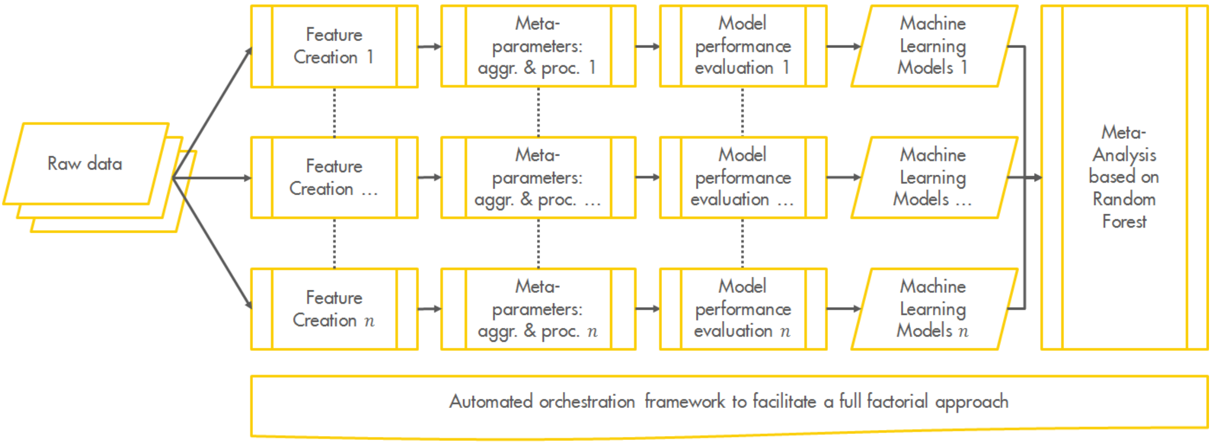

Raw data
Feature Creation 1
Feature Creation …
Feature Creation n
Meta-parameters: aggr. & proc. 1
Meta-parameters: aggr. & proc. …
Meta-parameters: aggr. & proc. n
Model performance evaluation 1
Model performance evaluation …
Model performance evaluation n
Machine Learning Models 1
Machine Learning Models …
Machine Learning Models n
Meta-Analysis based on Random Forest
Automated orchestration framework to facilitate a full factorial approach

Figure 3: High-level overview of the forecast methodology of this study.

Report Structure

The report is structured to first address the open questions of the event report as these also impact the spatiotemporal models presented in this study, to then describe the spatiotemporal model extension. Specifically, this study is structured as follows:

- The event rate model limitations are addressed in chapter 2, which:
  - Validates the ML models using as hold-out set the period between 2013-2016;
  - Demonstrates that the model performance for both short-term and long-term forecasts remains alike;
  - Indicates that the nature of 5-year hold-out forecasts is similar to the forecasts performed during the training period;

[1] The evaluation metrics guiding us throughout this study are the Mean Absolute Error (MAE), a standard choice in machine learning, and the Root Mean Square Logarithmic Error (RMSLE), particularly useful for count data with a large low-end tail as is the case here. There is a third error metric that we use in the form of the likelihood of a model given certain input data which is then used to compare two models using the likelihood ratio.

    - Confirms that the ML models forecast negligible seismicity given no gas production and assuming an ultimate state (long after shut-in) reservoir condition.
- Data sources for input feature definition for spatial forecasting are discussed in chapter 3 (potential predictors) and chapter 4 (the target: earthquakes).
- The spatiotemporal experimental setup is described in chapter 5. The aggregation, gridding, and smoothing of spatial predictors and forecasting targets are discussed, in addition to the feature and experiment down-selection approach that is applied to identify and exclude experiments from the factorial setup that are deemed as ineffective.
- The extensions required to enable spatiotemporal model validation are detailed in chapter 6.
- Ultimate-state constraints as introduced in chapter 2 to overcome event rate model limitations are enabled for a spacetime setting in chapter 7, followed by evaluation of the spatiotemporal seismicity forecasts in three ways:
    - A quantitative evaluation based on forecast performance;
    - A qualitative analysis of spatially aggregated event rates, in comparison with the pure event rate forecasts, developed earlier;
    - Full spatiotemporal forecasts.
- Conclusions and a discussion are presented in chapter 8. We note that while both the range of validity and the range of applicability have been substantially progressed, pending confirmation of both and more definite conclusions on forecast performance, the models reported on here should not be used for business decisions.

2. Extensions to the Event Rate model proposed by Limbeck et al. (2018)

In this chapter, we discuss additions to the experimental set-up presented in Limbeck et al. (2018) to evaluate in detail the range of validity of seismicity event rate forecasts using the machine learning methodology proposed by Limbeck et al. (2018). Our prime objective is to show that the ML methodology can forecast accurate seismicity event rates in the case of strong and sudden changes in production scenarios. This topic becomes especially relevant given the post-March 2018 production scenario announced by the Ministry of Economic Affairs and Climate (Ministry of Economic Affairs and Climate Policy, 2018) – see chapter 10.2 in Limbeck et al. (2018). We start by highlighting that the accuracy and the (range of) validity of the event rate forecasts are mainly determined by four aspects, namely:

i) The quality of the forecast during the hold-out validation period;

ii) The short-term forecast performance being indicative of long-term performance;

iii) The degree by which monotonically evolving features will be outside of the observed model training range, and;

iv) ML algorithms being, in principle, data-driven and not explicitly aware of physical mechanisms.

We advise reading Limbeck et al. (2018), in order to understand the extensions proposed to their methodology and how the above aspects impact the forecast of event rates. Regarding (i), we perform a validation forecast using hold-out data between 2013-2016 on a model trained on data between 1995-2012. We pay particular attention to the fact that the ML models capture the observed decay in earthquake rate (see Glossary) observed from 2013 onward. In (ii), we investigate the evaluation strategy of assessing models using their short-term performance – when uncertainties are smallest – henceforth ensuring maximum statistical power and an altogether better capability to distinguish among the various forecasts. We qualitatively show that the performance of all ML models remains alike for both short-term (1 to 3 months) and long-term (1 to 5 years) forecasts. Concerning (iii), features such as Pressure (*P*) and Hydrocarbon Column Thickness (*HCT*) are monotonically decreasing and increasing, respectively. That means, these – and likely other features as well – may eventually be out of the variable range used to train our models. To address that, we compute the so-called convex hull (Balas, 2010) for the training data and calculate the minimum distance between every unique point in the forecast set and the training convex hull. With that in hand, we show that our features during the forecast period lie outside the convex hull which encompasses the training data. We show that this out-of-the-convex-hull behavior is also observed within the forecasts performed within the training and testing period. Therefore, we argue that the overall quality of our forecasts should remain alike to the hold-out set forecasts performed during the training and testing period. Lastly, we further scrutinize the model general validity by including a physics-based analysis besides the mathematical ones discussed in (i-ii-iii). We highlight that our models – to some extent – already naturally encode underlying physical laws since our input features come either from observations or physically-sound numerical models. That alone, however, does not prevent the ML pipeline to relate these features in an unphysical fashion – since ML models are not naturally aware of physical models. Hence, we address (iv) by guaranteeing non-negative seismicity forecasts and by constraining[2] the models to an ultimate post-shut-in state where seismicity rate is set to zero, production is zero, pressure changes and other derivatives are also zero and reservoir pressure is within a possible range of depleted states under different production scenarios. We show that with these additions, the ML models can consistently forecast negligible seismicity when gas depletion ceases (i.e., all first and second order derivates of time-

[2] See Glossary

dependent features are zero) and pressure/HCT stabilize within the range obtained from the different production scenarios. In contrast, a continued-depletion situation would still result in non-zero derivatives, and hence seismicity rates would remain non-zero.

2.1. Validating forecast quality using a hold-out set of observations between 2013 and 2016

Limbeck et al. (2018) used all available data between 1995 and 2016 to perform an ML model meta-analysis. There, it was asserted that this is needed to ensure an optimal choice in the meta-parameters to model seismicity. Since the objective here is not to tune or choose a new set of meta-parameters, part of the data is used as a hold-out set, hence creating a validation set. The period between 2013 and 2016 shows a pronounced decrease in earthquake rate, making this period suitable to evaluate the quality of the (*hold-out*) forecasts. To do so, an experiment for $M_{min} = 1.2$ is designed, identical to the one presented in Limbeck et al. (2018) – their Table 17 - but spanning from 1995 up and including 2012 (ID: **FC95-12-1.2**). That leaves the leftover period containing observations (2013-2016) as validation of our methodology. We are interested to evaluate to what degree ML models can capture the steep decrease in earthquake rate observed in that period.

Echoing the results extensively discussed in Limbeck et al. (2018) – see their Appendix 8 – we observe in Figure 4 that weighted mean pressure (weighted mean *P)* and weighted mean hydrocarbon column thickness (weighted mean *HCT)* are the main seismicity rate drivers for the period 1995-2012. These are followed – to a much lesser extent – by temporal derivatives of gas produced (*Q* in m^3), *HCT* (m), and *P* (bar). As discussed in Limbeck et al. (2018), these features are representatives of a highly correlated group of features, which would, in turn, yield equally performing models.

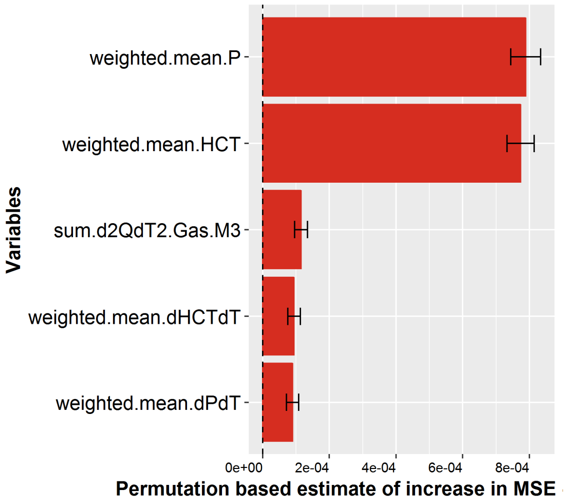


Figure 4: Variable importance plot (with uncertainties) for experiment FC95-12-1.2.

Aiming to understand the model responses to changes in important individual variables, we show in Figure 5 the ICE plots for weighted mean *P* and weighted mean *HCT*. Seismicity rates increase as weighted mean *P* decreases. The max-min ranges for both variables as well as the respective general trends are similar to the ones discussed in Limbeck et al. (2018) for their similar experiment RF-FC35-1.2 (1995-2016). Do notice that by not including the period 2013-2016 as training data, we are roughly halving the number of events used to train our models. However, regardless of small differences, these results suggest that the main features explaining seismicity between 1995 and 2012 remain consistent throughout at least 2016.

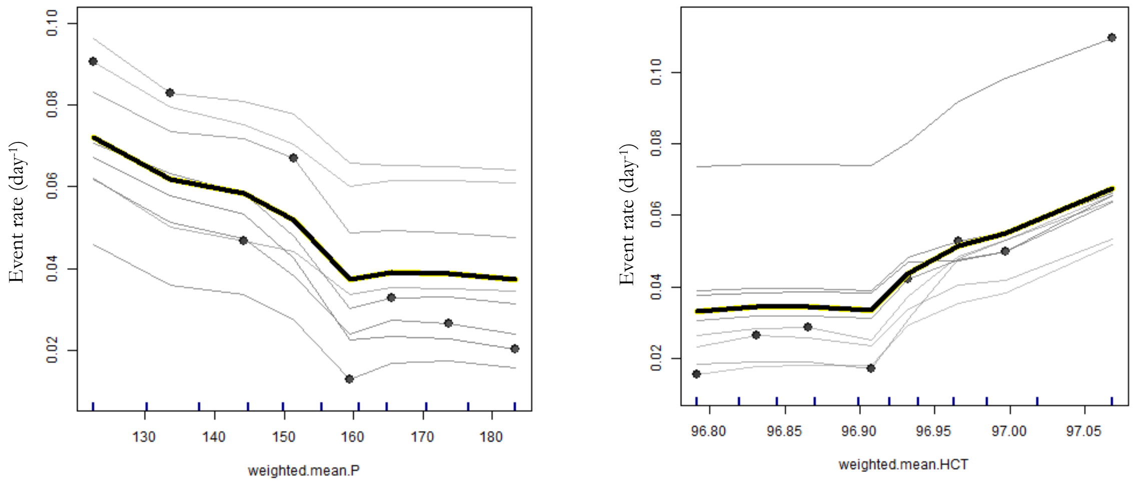


Figure 5: ICE plots for the main seismicity drivers of FC95-12-1.2 for (a) weighted mean P and (b) weighted mean HCT.

Next, seismicity is forecasted for the period between 2013 and 2016 using training data from the period 1995-2012. To illustrate the results, we show in Figure 6 the seismicity evolution calculated by the KSVM model.

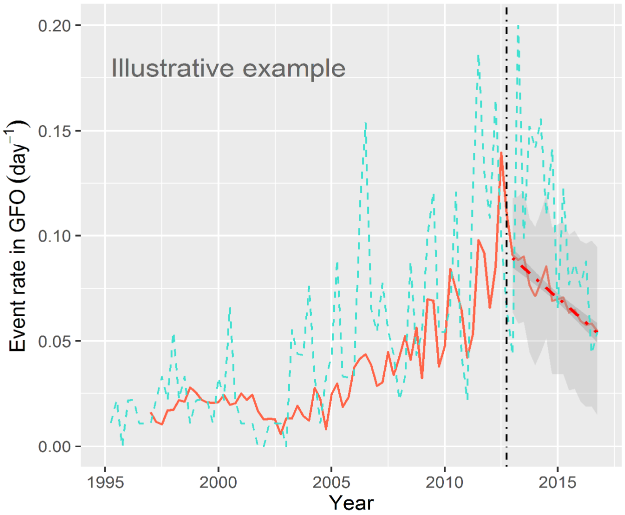


Figure 6: 3-month average seismicity forecasts (red line) compared to observations (blue dotted line) for the KSVM model within GFO. Results obtained for the production plan 2016 scenario. The vertical black line indicates the end of 2012, i.e., the end of the training period.

Note that the KSVM can capture the decay in seismicity observed from 2013 onwards. Besides the KSVM, all non- and extrapolating ML models are able – to different extents – to forecast the decay in seismicity observed from 2013 onward. In Table 1, we show the MAEs and their respective standard errors for the selected non- and extrapolating models for the period 2013-2016 compared against the observations. The standard errors are calculated based on the Jackknife resampling method, as described in Limbeck et al. (2018) and section 6.

Models which are heavily dependent on the pressure variation with time, such as the Depletion MA, perform well – since it is known that activity rate scales with reservoir pressure depletion.

According to the Wilcoxon signed-rank test, the MAE results obtained with the Depletion MA model are not statistically significantly better than any of the other individual ML models ($p > 0.05$ for all individual models) but are better than the moving average baseline ($p = 0.01248$, test statistic $V = 25$). The same is observed if we individually compare the ML models to the moving average model ($p \ll 0.05$ for all the individual comparisons).

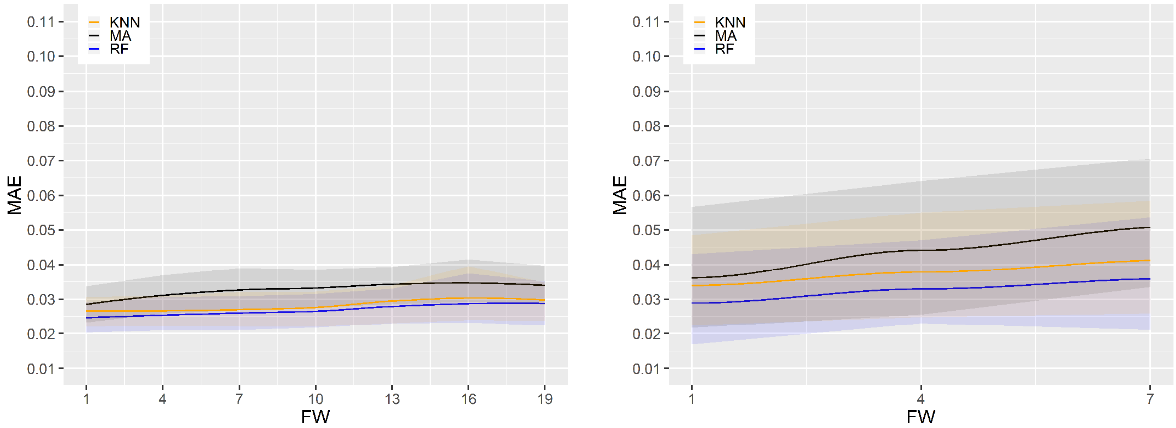
Table 1: MAE with standard errors (day^{-1}) between the hold-out set forecasts and observations for the period 2013 and 2016 for non-extrapolating and extrapolating models

Non-extrapolating models		Extrapolating models	
RF	0.0353 (±0.0061)	**SVM**	0.0401 (±0.0082)
KNN	0.0375 (±0.0073)	**GLMtop**	0.0377 (±0.0085)
MA (baseline)	0.057 (±0.011)	**Dep MA (baseline)**	0.0319 (±0.0055)

The results for the validation set (2013-2016) indicate that our current ML pipeline can capture the decay in seismicity observed from 2013 onward statistically indistinguishable from the Depletion MA baseline and significantly better than the moving average baseline. Consistently, these results echo the ones already discussed in Limbeck et al. (2018). From an MAE perspective, these values are considerably higher than the ones for the period 1995-2012 (see Table 2) because of the significant variability in seismicity between 2013 and 2016 and the high frequency of events on average during that period compared to 1995-2012.

2.2. Evaluating Concordance Between Short and Long-term Forecast Performances

The strategy proposed by Limbeck et al. (2018) of assessing the ML models using their short-term performance is further evaluated by qualitatively investigating if the long-term and short-term forecasts behave similarly. Also, the impact of altering initial conditions on the MAE is quantified by reproducing the same base experiment discussed in the previous sub-section but shifting the starting date in monthly steps by one year in total, i.e., 12 different experiments per individual forecast window (FW). In Figure 7, we present the min/max and the average MAE for the 12 different starting dates per forecast window (FW) for both short-term (small FWs) and long-term (large FWs) forecasts. The results are shown for both the periods 1995-2012 and 2013-2016.



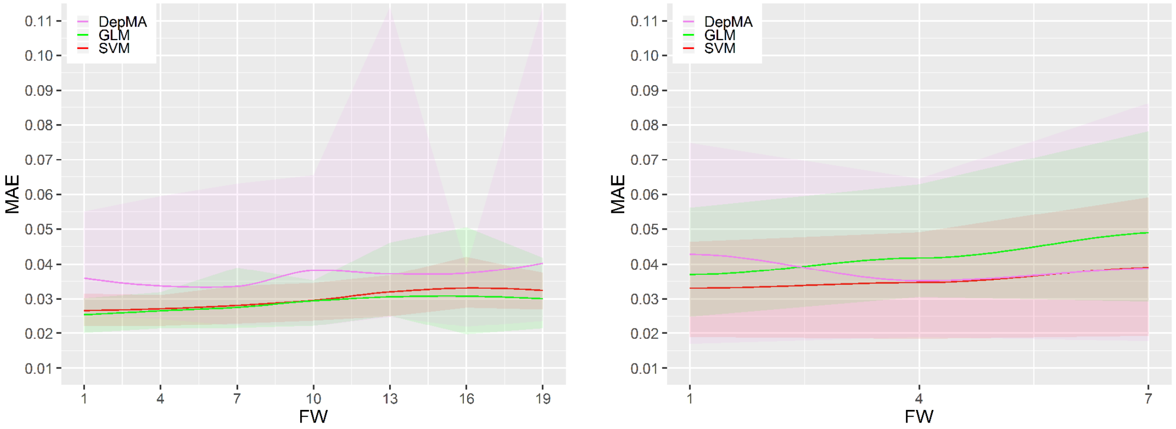


Figure 7: MAE (in day-1, vertical axes) per different forecast windows (horizontal axes). The solid lines indicate the average MAE for different starting dates for every FW. The shaded areas describe the total MAE range for the different experimental initial conditions per FW. Small FWs represent short-term forecasts while large FWs represent long-term forecasts. Left plots are for 1995-2012 and right plots for 2013-2016.

As expected, for all ML models shown in Figure 7 (RF, KNN, KSVM, GLM Top), there is a loss in performance (i.e., higher MAE) for long-term (large FWs) forecasts compared to short-term (small FWs) for both 1995-2012 (left plots) and 2013-2016 (right plots) periods. For the period 1995-2012 (left plots), all ML models are within an average MAE (thick lines) roughly between 0.025-0.035. For the period 2013-2016 (right plots), the average MAE is larger (around 0.030-0.050) – indicating a loss in performance compared to the 1995-2012 period. The consistent increase in MAE variation for all FWs (shades) during 2013-2016 (right plots) compared to 1995-2012 (left plots) can be partially explained by the more significant variability in seismicity after 2012 and by the frequency of events being on average higher during 2013-2016 if compared to 1995-2012. By visually inspecting the non-extrapolating model's evolution (upper plots), the model choice remains unchanged for all different FWs (i.e., short- and long-term forecasts) for both periods 1995-2012 (left plots) and 2013-2016 (right plots). For the extrapolating models (lower plots), we reach similar conclusions except for the Depletion MA baseline results. That said, Limbeck et al. (2018) showed that the MAE results of their ML short-term forecasts are statistically indistinguishable. Here, the average MAEs and the ranges observed in Figure 7 considerably overlap, qualitatively indicating a similar outcome. That is valid for both short (small FWs) and long-term (large FWs) forecasts. In conclusion, these results qualitatively indicate that the performance of all ML models remains alike for both short-term (1 to 3 months) and long-term (1 to 5 years) forecasts for all ML models.

2.3. Convex-Hull analysis to evaluate model validity for the forecast period

We use convex hulls to evaluate if/when earthquake rate forecasts may become unreliable due to out-of-bound features. A convex hull is commonly defined as the smallest convex set that contains a given set of points (Efron, 1965; Balas, 2010). Here, convex hulls are used to evaluate if the features used by the ML models lie within or outside a convex hull calculated for the same features during training, which might be important for the validity of the so-called non-extrapolating ML models used in this study.

The out-of-bounds issue motivates the use of the convex hull to ensure the forecast range does not require excessive extrapolation. We acknowledge that there are other methodologies than the convex hull available in the literature to perform this type of "change-point detection" (Liu et al., 2013) or "novelty detection" (Markou and Singh, 2003). Notably, we highlight the works from

Chandola et al. (2009) and Pimentel et al. (2014) for providing extensive reviews on other viable approaches (e.g., Gaussian models, likelihood range, clustering, nearest neighbours, among others). Here, we select the convex hull technique as our tool of choice because it is efficient, easily interpretable across disciplines and provides insights on temporal dependencies by calculating distances between newer features and the ones used to calculate the training convex hull. According to (Majumdar et al., 2010), convex-hull applications are widely found in many different instances, such as medicine, biology, computer sciences, mathematics, physics, astrophysics, geography, and many others – see Getz and Wilmers (2004; Stout et al. (2008); Van De Weygaert et al. (2011), Chupeau et al. (2015) and Wolsey and Yaman (2018) for recent applications.

There are many convex-hull algorithms available in literature – see Avis et al. (1997) for a review on different methodologies. We take advantage of the Quickhull algorithm Barber et al. (1996) to calculate the convex hulls for both the training and forecast sets. A d-dimensional convex hull can be represented by its vertices and facets. Each facet comprises vertices, neighbouring facets, and a hyperplane equation. According to Barber et al. (1996), Quickhull uses two different geometric operations: oriented hyperplane through d-points and signed distance to the hyperplane. A hyperplane defines points with negative distances to it. Analogously, if the distance is positive, the point is classified as above the hyperplane. The algorithm assigns every new point to an outside set and Quickhull locates a visible new facet for each point. In case a point is above these facets, one of the new facets is selected. If the point is below, it is then assigned as already inside the convex hull and therefore discarded.

A caveat of this method is that it is not able to identify specific areas inside the polytope without feature distribution support. That means that a particular combination of features may be inside the outer hyperplanes of the convex hull but in an area that is outside of the feature training range. Hence, being outside of the convex hull does imply being outside the feature training range, but that does not guarantee that a point inside the convex hull lies inside the feature training range.

Besides using Quickhull to obtain the convex hull for the training set and to verify if the points in the forecast period fall inside/outside the polytope, we also calculate the minimum distance between every forecast point and the training-convex-hull. Distance calculations provide a better understanding as to how far – in the feature domain – training and forecast sets are and if there is any temporal dependence in the forecast set. To calculate distances, we use an algorithm from the CGAL library (CGAL Project, 2018) that calculates the (squared) distances between two convex polytopes. Briefly, the polytope algorithm allows for efficient calculations of the smallest distance between two convex hulls. This problem is formulated as an optimization problem with linear constraints and a convex quadratic objective function. The solution is obtained using an efficient solver for quadratic programs (Gärtner and Schönherr, 2000). We highlight that the execution time increases approximately linearly with the number of points for any fixed dimension, which allows for the inclusion of all the significant features for the entire training and forecast periods in the distance calculation while keeping the execution time within a few seconds.

In CGAL and Quickhull, all features (for training and forecast periods) are centred and scaled based on their respective means and standard deviations for the training period, followed by calculating whether the features - used to make the forecasts - lie inside the convex hull made with the (centred-scaled) training data. This approach is in alignment with how our data are used within the model training procedure first described in Limbeck et al. (2018). If forecast features lie outside, we calculate their minimum (squared) distance to the training convex hull. To understand which of the features explain these distances, we also calculate for every individual feature the number of times it lies outside the min-max training range during the forecast period. Note that individual forecast features lying inside their min-max ranges are not guaranteed to be inside the convex hull. Therefore, the range is used solely to highlight features *outside* of the min-max range since that implies to be outside of the convex hull as well. The results for the baseline experiment (FC95-12-1.2) are displayed in Figure 8.

The results in Figure 8a show that all points in the forecast period (2013-2017) display a (squared) distance larger than zero. Here, every point represents the squared distance between the vector containing the forecast features for that specific date and the convex hull encompassing all the training features. That means all forecasts have features which lie outside the training convex hull. That is not surprising, however, since Figure 8(b) shows that weighted mean *P* and weighted mean *HCT* are always outside the range observed in training. As these variables are monotonically evolving, they get further away from the training range, which also explains the positive trend for the squared distances. The dotted-line (median) and shade (IQR) in Figure 8(a) represent the variation in distance found in every forecast window within a 5-year period from 1995 until 2012: These statistics are calculated starting with a training set containing the minimum number of points, according to section 5.5 in Limbeck et al. (2018) – for all forecast windows up to 5 years ahead. Then, one point is added to the training set, and the process is repeated until the training set encompasses all the points within 1995-2012. These statistics show that the distances observed for the forecast period starting in 2013 are within the range of distances previously observed during the period 1995-2012. This *qualitatively* indicates that the nature of the forecasts should not obviously differ from our hold-out set forecasts during the training period. In addition to that, the ection in which the convex hull is extended seems to remain mainly dependent on *P* and *HCT*.

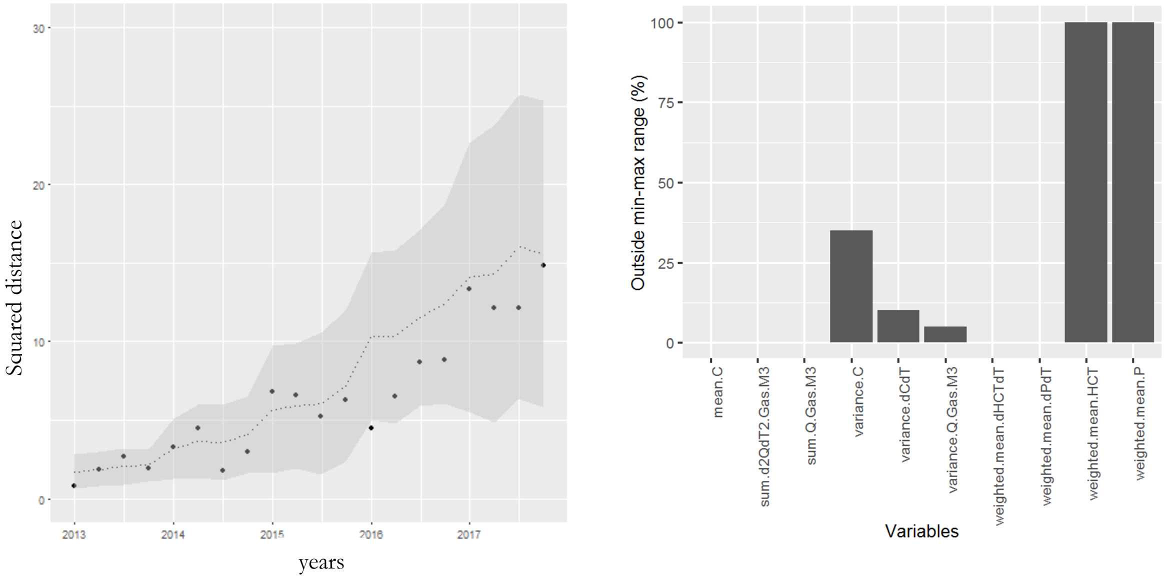


Figure 8: (a) Squared distance between every vector containing the forecast features in the forecast period (2013-2016) to the convex hull encompassing all training features (1995-2012). The dotted-line represents the median distance between every forecast window within a 5-year period from 1995 until 2012. The shades indicate the respective IQR – see text – and (b) fraction of features in the forecast period outside their respective min-max training range. Results obtained for the production plan 2016 scenario for a 5-year forecast window.

Extending on these results, a natural next step is to repeat this simulation excluding weighted mean *P* and weighted mean *HCT* from the set of available features – since those are monotonically declining variables and will consistently increase the distance between forecast and training periods. The idea is to identify a set of features which would remain inside the training convex hull during forecast. We show these results in Figure 9.

Here, 80% of the forecast features lie outside the training convex hull, however roughly within the observed variation in distance (IQR) found during training. The outlier observed at the beginning of 2017 – with a squared distance roughly equal to 5 – is explained by an 11-fold-larger-than-average variance of the temporal derivative of compaction. We highlight the much smaller magnitude of the squared distances if compared to the experiments including weighted mean *P* and weighted mean *HCT.* Also note that the prominent upward trend observed in Figure 8 is not present here, confirming that weighted mean *P* and weighted mean *HCT* are the primarily responsible mechanisms for increasingly distancing the forecast feature values from the training feature values. The same is also found in Figure 9b, where only some variances are observed, for a much shorter period, outside the min/max training range.

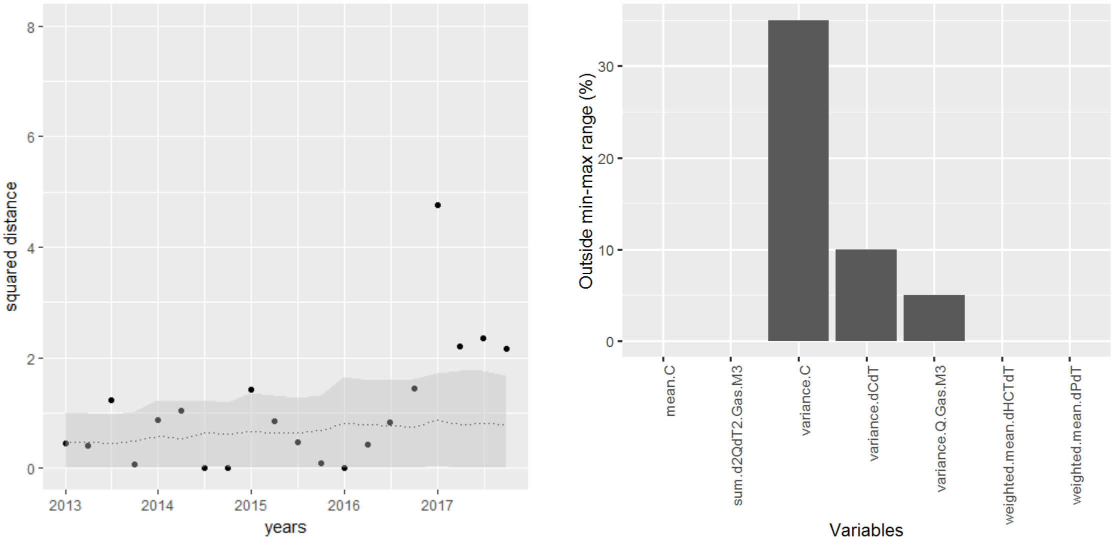


Figure 9: Squared distance between every vector containing the forecast features in the forecast period (2013-2016) to the convex hull encompassing all training features (1995-2012). Results obtained for the experiment excluding the weighted means of P and HCT and the production plan 2016 scenario, for a 5-year forecast window. The dotted-line represents the median distance between every forecast window within a 5-year period from 1995 until 2012. The shades indicate the respective IQR – see text – and (b) fraction of features in the forecast period outside their respective min-max training range.

Another valid approach aimed at having the forecast features inside the training convex hull is only to use features we can guarantee will remain inside the training range. Here, we show results for an experiment using solely the weighted mean of the temporal variation of P and the scaling relation between seismicity rate and depletion. These are essentially the features used by the Depletion MA model (see Limbeck et al. (2018) for the formulation details). The forecast results for this experimental setup, as expected, always remain inside the training feature convex hull – i.e., squared distances equal to zero throughout the entire forecast period. That approach, however, comes associated with a strong decay in forecasting performance – as discussed next.

In Table 2 to Table 7, we summarize and compare the performance of numerical experiments (*ALL*) containing all available features, (*noPnoHCT*) all but weighted means *P* and *HCT*, and (*onlyDeltaPc*) only scaling and weighted mean of the temporal variation of *P*, for both periods 1995-2012 (Table 2-Table 4) and 2013-2016 (Table 5-Table 7).

Table 2: MAE with standard error (day^{-1}) for experiment *ALL* - (1995-2012)

Non-extrapolating models		Extrapolating models	
RF	0.0231 (±0.0028)	**KSVM**	0.0253 (±0.0029)
KNN	0.0249 (±0.0028)	**GLM Top**	0.0257 (±0.0027)
MA (baseline)	0.0266 (±0.0033)	**Dep MA (baseline)**	0.0291 (±0.0037)

Table 3: Same as Table 2, but for experiment *noPnoHCT* - (1995-2012)

Non-extrapolating models		Extrapolating models	
RF	0.0289 (±0.0037)	**KSVM**	0.0318 (±0.0040)
KNN	0.0298 (±0.0039)	**GLM Top**	0.0309 (±0.0036)
MA (baseline)	0.0266 (±0.0033)	**Dep MA (baseline)**	-

Table 4: Same as Table 2, but for experiment *onlyDeltaPc* - (1995-2012)

Non-extrapolating models		Extrapolating models	
RF	0.0284 (±0.0039)	**KSVM**	0.0301 (±0.0041)
KNN	0.0288 (±0.0039)	**GLM Top**	0.0265 (±0.0037)
MA (baseline)	0.0266 (±0.0033)	**Dep MA (baseline)**	-

Table 5: Same as Table 2, but for experiment *All* - (2013-2016)

Non-extrapolating models		Extrapolating models	
RF	0.0353 (±0.0061)	**KSVM**	0.0401 (±0.0082)
KNN	0.0375 (±0.0073)	**GLM Top**	0.0377 (±0.0085)
MA (baseline)	0.057 (±0.011)	**Dep MA (baseline)**	0.0319 (±0.0055)

Table 6: Same as Table 2 but for experiment *noPnoHCT* - (2013-2016)

Non-extrapolating models		Extrapolating models	
RF	0.0484 (±0.0086)	**KSVM**	0.0497 (±0.0087)
KNN	0.0511 (±0.0094)	**GLM Top**	0.0519 (±0.0073)
MA (baseline)	0.057 (±0.011)	**Dep MA (baseline)**	-

Table 7: Same as Table 2, but for experiment *onlyDeltaPc* - (2013-2016)

Non-extrapolating models		Extrapolating models	
RF	0.078 (±0.010)	**KSVM**	0.081 (±0.010)
KNN	0.082 (±0.010)	**GLM Top**	0.075 (±0.010)
MA (baseline)	0.057 (±0.011)	**Dep MA (baseline)**	-

Non-extrapolating and extrapolating ML models perform similarly within each experiment. This conclusion is valid for both training and forecast periods. Specifically, for the 1995-2012 training period, we use a paired Wilcoxon test to show that the performance of the best ML model (RF for *ALL* and *noPnoHCT*, and GLM Top for *onlyDeltaPc*) is not significantly different from the other ML models nor baselines. For the period 2013-2016, the paired Wilcoxon tests show that for experiment *ALL*, all individual ML models are statistically indistinguishable from the Depletion MA ($p > 0.05$) and statistically better than the moving average. Since weighted mean P is not a feature in experiments *noPnoHCT* and *onlyDeltaPc*, there are no results for Depletion MA in these cases. For the experiment *noPnoHCT,* the MAE results indicate that only the RF model performs significantly better than the moving average ($p = 0.009125, V = 23$). For all other ML models, $p > 0.05$ suggests that these models perform similarly to the moving average. Hence, we conclude that removing weighted mean *P* and weighted mean *HCT* fairly worsens our modelling performance, albeit decreasing the distance between the feature spaces of the training and forecast periods. While that is an important result, we show in Figure 8 that the out-of-range magnitudes for training and forecast periods remain similar throughout the entire forecast period. For experiment *onlyDeltaPc*, MAEs are much higher than for previous experiments, and the Wilcoxon tests confirm that all individual ML models perform significantly worse than the moving average baseline ($p \ll 0.05$ for all individual models).

In conclusion, we have qualitatively shown that the nature of the forecasts discussed here should remain similar to the hold-out set forecasts performance during the training period. Removing weighted mean *P* and weighted mean *HCT* from the list of features reduces the magnitude of the squared distances between forecast and training points considerably – as well as removes the upward trend in squared distances observed throughout the entire forecast period. That, however, worsens the overall modelling performance. In the next section, we discuss how we address negative forecasts and non-zero seismicity rates even after the field has been long shut-in.

2.4. Evaluating seismicity forecasts given an ultimate post-shut-in field state

This section introduces extensions to our ML models that ensure seismicity forecasts are not producing negative rates and that seismicity rates decrease towards zero in the time after the field has been shut in. We trivially guarantee seismicity forecasts greater than or equal to zero throughout the entire forecast period, by forcing seismicity to zero in case of any spuriously generated negative seismicity forecast, to ensure the applicability and validity of the long-term (1 to 5 years) forecast results. Note that this procedure was already in place and is discussed in Limbeck et al. (2018). Second, positive and/or increasing seismicity rates long after the field shut-in date or complete depletion of the field is addressed by providing input training features to the ML models for when the system has theoretically reached its "ultimate state". We define "ultimate state" as the stable state the system reaches long after gas production has ceased (leaving aside processes that operate on geological time scales) – see Glossary. This "ultimate state" of the field is calculated using scenarios from the reservoir flow model (Van Oeveren et al., 2017).

We aim to extend the training range and therefore try to limit the ML model capacity to relate variables in a way that is known to be physically impossible. In doing so, we train the ML models with first and second derivatives of time-dependent features set to zero and with features from reservoir flow model forecasts, such that the ML model seismicity forecasts decrease to approximately zero after the field is shut in and the pressure in the field approaches a new equilibrium state. From an ML perspective, we turn an extrapolation problem (no ultimate state provided) into an interpolation problem (ultimate state provided). Concretely, the MoReS asset model for Groningen (Van Oeveren et al., 2017) and subsequent compaction and subsidence models (Bierman, S., Kraaijeveld, F., Bourne, 2015) are used to obtain forecasts for the post-March 2018 average production scenario up to 2100, when it is expected that the reservoir and the overburden have reached an "ultimate state" situation and induced seismicity is significantly decreased or ceased in its entirety.

According to the MoReS simulation for the post-March 2018 average production scenario, weighted mean P and weighted mean HCT are in 2100 equal to 86 bar and 97.3 m, respectively. To minimize the risk of overfitting or teaching the model a training feature field to be again observed in the forecast period, we do not provide these exact values as training data to our ML pipeline. Instead, we use data obtained from two extreme production scenarios: *shut-in 2019* and *maximum depletion*. The former accounts for a field shut in happening in 2019 (maximizing e.g., weighted mean P), while the latter describes a situation where the maximum amount of gas is depleted (minimizing e.g., weighted mean P). In addition to P and HCT, gas production rate and all temporal derivatives are also provided as training features and assumed to be zero (i.e., a steady-state situation). Under these steady-state conditions, with no production, *seismicity rate is also set to zero*. In Figure 10, we sketch our setup using P as an example:

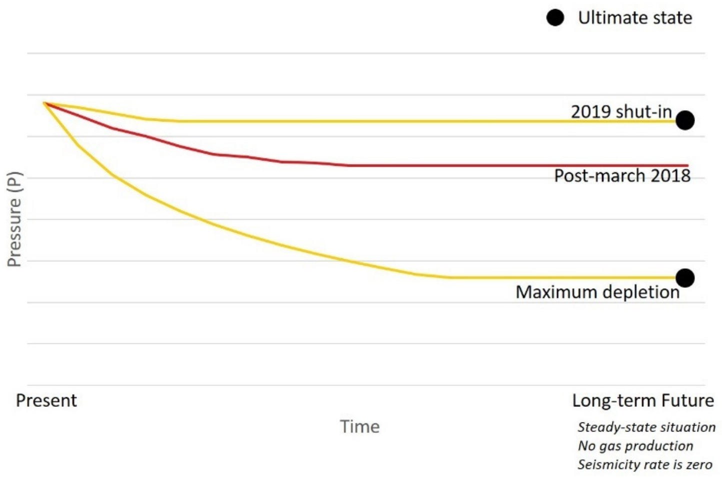


Figure 10: Sketch of the ultimate states for weighted mean pressure provided as training data to the ML pipeline.

By using these as training data, we are essentially ensuring the ML models to forecast no seismicity under the conditions mentioned above. Specifically, the feature values for weighted mean P and weighted mean HCT representing these two-possible extreme definite states are shown in in Table 8**Error! Reference source not found.**:

Table 8 Values of the ultimate-state features for the production scenarios shut-in 2019 and maximum depletion.

Production scenarios	P (bar)	HCT (m)
Shut-in 2019	94	97.3
Maximum depletion	37	97.8

By using both weighted mean P and weighted mean HCT obtained from the MoReS extreme scenarios as ultimate feature states, it is guaranteed that the post-March 2018 average production scenario falls within *this range*. We highlight that under a different production scenario, where for example ultimate pressure does not fall within the range given in Table 8, the pipeline needs to be readjusted. Hence, our ML models are guaranteed to be able to forecast negligible seismicity within the ranges provided in Table 8 (and given no-depletion and steady-state situations), but in case a new production scenario is used, the pipeline needs to be re-validated. The features simulated for both production scenarios are added 10 times each to our ML pipeline. We arrive at ten by means of a sensitivity analysis where we empirically tried different repetitions (2, 4, 6, 8, and finally 10) and concluded that ten repetitions per production scenario is the minimum number of repetitions the ML pipeline needs to learn that the features within the range displayed in Table 8 (and no depletion and steady-state) should lead to negligible seismicity. Hence, the reason we add the features exactly 10 times each is solely to provide the ML models with sufficient amount of synthetic training data and ensure the forecast of negligible seismicity rates. In addition, by adding ten points, we are not significantly impacting model performance – as shown in Table 9. In Figure 11, we show the temporal evolution of seismicity from 1995 until 2100 for the cases with and without ultimate states.

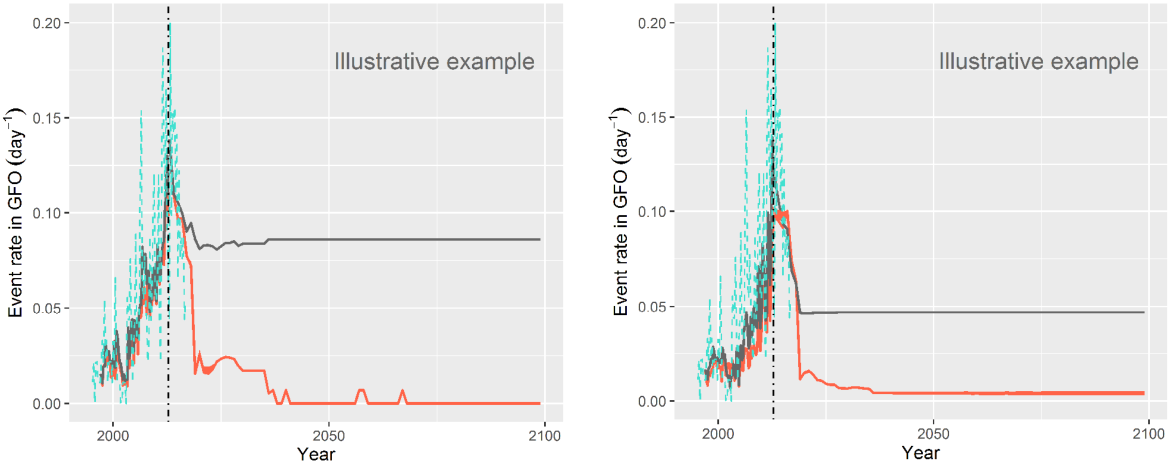


Figure 11: Hold-out set seismicity forecast (2013-2100) for (left) RF, and (right) KSVM. Results obtained for the post-March 2018 production scenario. Turquoise dashes indicate the observations (1995-2012). The black line represents the forecast without ultimate states. The red band represent the maximum variation between the experiment described in Table 8 and the sensitivity analysis for the different forecasts with ultimate states (Table 10) – see text.

The seismicity forecasts for both RF and KSVM models converge to zero around 2037. The inclusion of ultimate states does not significantly impact model performance during both 1995-2012 and 2013-2016 periods if compared to models without the ultimate states, as shown in Table 9. A paired Wilcoxon test confirms that the MAE values for the experiments are not significantly different.

Table 9 Comparison between MAE with standard error (in day^{-1}) for experiments with and without ultimate state constrains for the periods 1995-2012 and 2013-2016.

Without ultimate state		With ultimate state	
RF (1995-2012)	0.0231 (±0.0028)	**RF (1995-2012)**	0.0224 (±0.0030)
KSVM (1995-2012)	0.0253 (±0.0029)	**KSVM (1995-2012)**	0.0282 (±0.0035)
RF (2013-2016)	0.0353 (±0.0061)	**RF (2013-2016)**	0.0328 (±0.0065)
KSVM (2013-2016)	0.0401 (±0.0082)	**KSVM (2013-2016)**	0.0365 (±0.0076)

Finally, we further extend the robustness of these conclusions by performing a sensitivity analysis on the values discussed in Table 8. A series of individual experiments (S1-S5) is designed where we independently add different weighted mean P and weighted mean HCT pairs as ultimate state points to the training set, as shown in Table 10. Consistent with the experiment described in Table 8, all other input features (including gas production), and seismicity rate itself are set to zero. The aim of these five independent runs is twofold: (i) to show that results remain similar for any production scenario within the ranges shown in Table 10, and (ii)) to evaluate if deviations around the MoReS forecast values affect our results. We set the weighted means P and HCT values for every experiment to gradually depart from the *Shut-in 2019* values. The difference between the sensitivity analysis experiments and the experiment shown in Table 8 is the fact that sensitivity experiments (S1-S5) are independent of each other, meaning that they provide the ultimate state points to the training set one experiment at the time, instead of as a range as in Table 8. Hence, the points in experiment S1 are used to run solely one experiment, and points in experiment S2 are used to run another independent experiment, and so on.

Table 10 Sensitivity analysis on the synthetic points for pressure and HCT provided to the training set for five different independent experiments.

Experiments	P (bar)	HCT (m)
S1 (Shut-in 2019)	94	97.30
S2	90	97.35
S3	86	97.38
S4	82	97.40
S5	78	97.42

The results for all the five independent experiments (S1-S5) are virtually indistinguishable if compared to the experiment described in Table 8. Note in Figure 11 that the red band indicating the variation in seismicity rate for these experiments is very narrow, showing the almost negligible impact of this sensitivity analysis for both training (1995-2012) and forecast (2013-2100) periods. Despite results being discussed until 2100, we highlight that no quantitative claims are made about model forecast performance for such long-term forecasts. Consistently, we limit our long-term analysis up to 5 years. The interest with the ultimate state analysis is to show that in the very long term, models forecast no seismicity given no gas production and an ultimate state field situation. That, as expected, agrees with current physical expectations. Besides allowing the ML models to learn new relationships among the variables, the ultimate state constraining is especially important for the non-extrapolating models. That is because the inclusion of these points widens the feature training range and by extension the convex hull for the training period. That means P and HCT throughout the entire forecast period now remain within the training convex hull. That claim, however, is only valid for these monotonic variables and is not applicable for variances or derivatives since these can very well deviate from previously observed values.

3. Spatiotemporal features related to induced seismicity

The extended event rate model discussed in the previous section forms the basis for further extensions towards a spatiotemporal model, where seismicity rates are forecasted in both time and space throughout the Groningen field. The spatiotemporal models require the use of the spatial component of the existing temporal features (e.g., reservoir pressure) as well as new spatial (time-invariant) features. The selection of spatiotemporal input features is guided by the physical understanding of the seismicity in the Groningen field, using the Coulomb stress model (3.1). Using this model, we distinguish between spatially varying time-invariant features (3.2) and spatially and temporally varying features (3.3). Finally, a summary of all spatiotemporal model features is presented in 3.4, in addition to a summary of how these features were generated and from which data or models they were obtained. Additional details regarding data sources are provided in Appendix B.

3.1. Overview of seismicity mechanisms in the Groningen field

Induced seismicity in the Groningen field is compaction-driven, resulting from production-driven pore pressure decrease in the reservoir. The pressure decrease initially is strongest near the gas production wells. The volume changes associated with gas production in the direct vicinity of the producer wells generate a pressure decrease near these wells, which propagates throughout the reservoir via a diffuse process controlled by the rock matrix porosity and permeability. Provided that the rock mechanical and flow properties have been accurately measured and modelled, the pore pressure diffusion throughout the reservoir can be captured by a reservoir flow model, which is calibrated to pressure, rate and temperature measurements at the producer wells.

The average reservoir pore pressure in the Groningen field has dropped from a pre-production pressure of 180 bars to below 100 bars in the present day. The decrease in pore pressure increases vertical effective stress, as the compressible reservoir rock compacts with depletion whereas the vertical overburden volume, and stress, acting on the reservoir remains constant. The increase in effective stress in the reservoir also increases the normal and shear stresses acting on faults within the reservoir. The Coulomb stress definition can be used to describe the stress state of a fault, and its proximity to failure. The Coulomb stress is the amount of stress (in Pascal) that is acting on a fault surface, as a function of the principle stresses in the earth subsurface and the geometry and orientation of the fault. The Coulomb stress can be defined as a constant value on the entire fault surface, assuming a planar fault, or at each individual point on a fault surface, if a more detailed fault geometry and rock property description is available. We use the following definition for Coulomb stress on faults in a homogeneous elastic thin-sheet model (Bourne and Oates, 2017):

$$C = C_i + \Delta C,$$

where C is the maximum Coulomb stress (in Pa) on the fault plane at failure conditions, C_i is the initial Coulomb stress (i.e. the combined effect the tectonic stress regime and the fault orientation) and ΔC is the change in Coulomb stress induced by a change in pore pressure ΔP:

$$\Delta C = -Y\Delta P,$$

where Y is a material constant capturing the poroelastic reservoir rock behaviour and the fault frictional properties:

$$Y = (1+\mu)\gamma\alpha - \mu, \text{ with}$$

$$\gamma = \frac{1-2\nu}{2(1-\nu)},$$

where ν is the Poisson's ratio (dimensionless ratio of radial strain to axial strain measured in e.g., a core plug), α is the Biot coefficient (dimensionless poroelastic rock property constant describing grain vs. bulk volume compressibility) and μ is the dimensionless fault friction coefficient, which defines the shear stress required to induce slip on a fault. Following this model, the Coulomb stress at any point on a fault is defined by a combination of matrix rock properties (ν, α), fault properties (geometry, friction angle μ), pore pressure and the magnitude and orientation of the three principle stress components in the subsurface. All these properties are expected to be a function of space, but constant in time, at least at the production time scale (as opposed to geological time scales), with the exception of pore pressure, which changes as a function of production.

In the case of Groningen, it has been demonstrated that this model needs to be extended from homogeneous elastic to a model that incorporates pre-existing fault geometries using geometric heterogeneities and elastic heterogeneities to account for observed lateral variations in reservoir compressibility (Bourne and Oates, 2017):

$$C = C_i + \mu \Delta P - \left(\mu + \sqrt{1 + \Gamma^2}\right) \gamma H \epsilon_{zz},$$

where H represents the reservoir bulk modulus (in Pa), ϵ_{zz} is the purely vertical strain (dimensionless) and Γ captures the other nonzero elements of the vertically averaged strain tensor, ϵ_{xz} and ϵ_{yz}:

$$\Gamma^2 = 4 \frac{\bar{\epsilon}_{xz}^2 + \bar{\epsilon}_{yz}^2}{\bar{\epsilon}_{zz}^2}.$$

The pre-existing fault geometries are incorporated by defining the fault friction range that leads to fault instability as a function of Γ:

$$\mu = \frac{\gamma \sqrt{1 + \Gamma^2}}{1 - \gamma + \frac{H_r}{H_s}},$$

with,

$$H_r = \Delta P / \epsilon_{zz}$$

$$H_s = 3K_s \left(\frac{1-\nu}{1+\nu}\right),$$

where K_s is the bulk compressibility (Pa^{-1}) in poroelasticity theory. Note that these equations assume that the fault orientation is ‘optimal’ from a failure perspective. As the overburden stress and rock and fault friction properties remain constant, fault failure is controlled by changes in the pore pressure in the reservoir, and the initial Coulomb stress of each fault:

$$C_i > -\Delta C.$$

It follows that even if the reservoir would be homogeneously depleting, some faults are more prone to failure than others as their initial Coulomb stress state may differ because of geometrical and elastic heterogeneities discussed above.

The above equations provide a physical model for induced seismicity, defined by Coulomb stress. However, the rock properties and stress parameters can at best only be measured at the locations of the wells, and fault friction properties cannot be measured directly in the reservoir. Hence, we identify in the following sections data-driven features that can serve as proxies for the two main mechanisms, i.e. C_i and ΔC in the above equations. The spatiotemporal features to be considered in the machine learning approach should capture spatial variations in the reservoir and overburden that may influence the initial Coulomb stress of the faults (spatially-varying time-invariant features) and the change in Coulomb stress induced by gas production, which can be varying in space and time. Examples of such features have already been investigated in other works, such as the Hazard and Risk Assessment model (Figure 12). This section builds upon those earlier findings in guiding the selection of features for the ML models. The following subsections discuss the relevant ML predictor features that have been identified based on their impact on either the initial Coulomb stress or change in Coulomb stress.

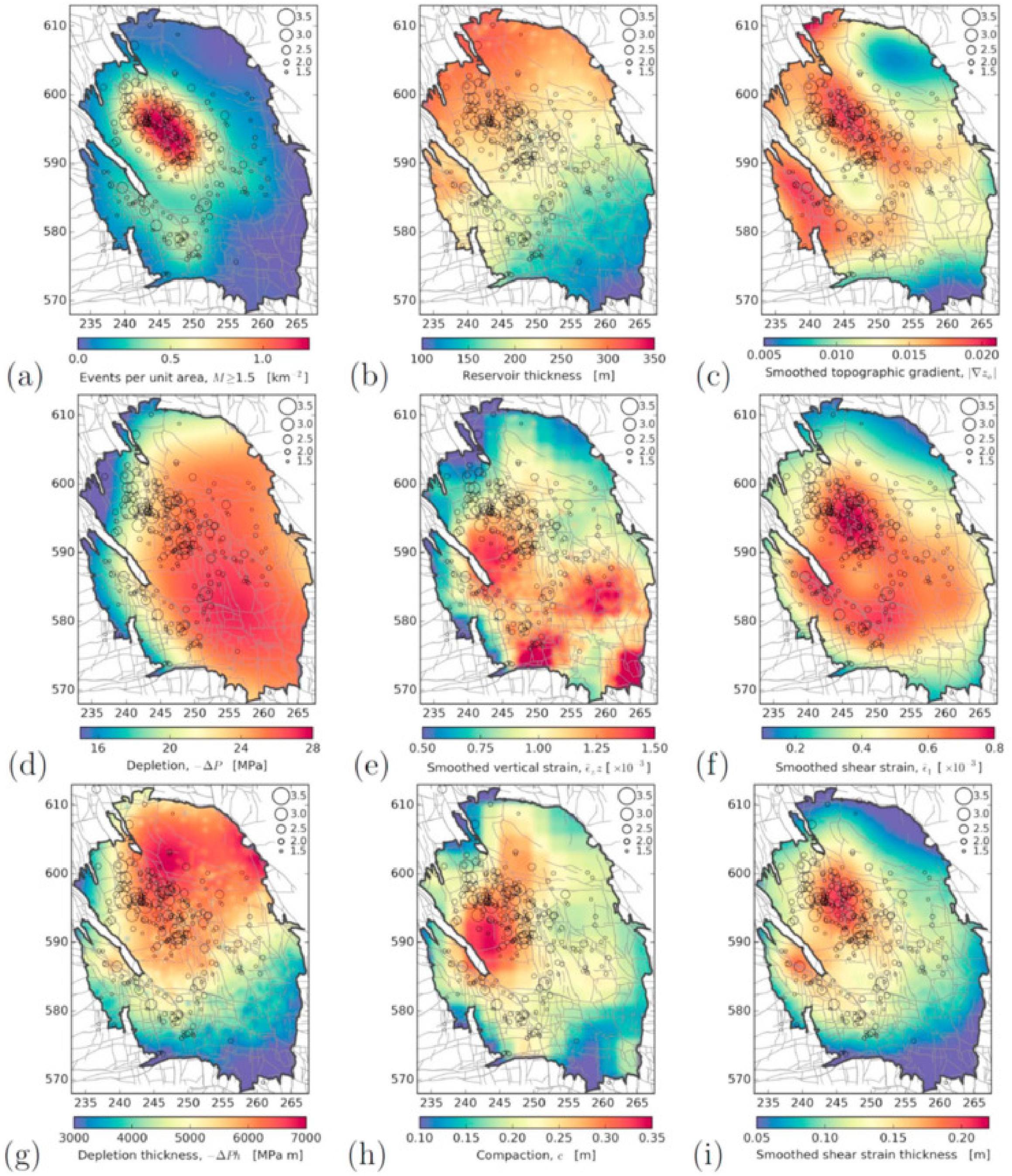

610
600
590
580
570
235 240 245 250 255 260 265
3.5
3.0
2.5
2.0
1.5
(a)
0.0 0.5 1.0
Events per unit area, $M \geq 1.5$ [km^{-2}]
(b)
100 150 200 250 300 350
Reservoir thickness [m]
(c)
0.005 0.010 0.015 0.020
Smoothed topographic gradient, $|\nabla z_s|$
(d)
16 20 24 28
Depletion, $-\Delta P$ [MPa]
(e)
0.50 0.75 1.00 1.25 1.50
Smoothed vertical strain, $\bar{\epsilon}_{zz}$ [×10^{-3}]
(f)
0.2 0.4 0.6 0.8
Smoothed shear strain, $\bar{\epsilon}_1$ [×10^{-3}]
(g)
3000 4000 5000 6000 7000
Depletion thickness, $-\Delta Ph$ [MPa m]
(h)
0.10 0.15 0.20 0.25 0.30 0.35
Compaction, c [m]
(i)
0.05 0.10 0.15 0.20
Smoothed shear strain thickness [m]

Figure 12 Spatial comparison of seismic event density (a) and other features using data from the period between April 1995 and January 2017 (Bourne and Oates, 2017).

3.2. Spatial features as proxies of initial Coulomb stress

We analyse three groups of features that may influence the initial local Coulomb stress state acting on faults:

1. The fault geometry, including offset and fault orientation attributes;
2. Poroelastic reservoir rock properties;
3. Pre-production geological variations in overburden stress.

Fault properties

The initial Coulomb stress is a function of the fault orientation relative to the orientation of the paleostress field principal stress components (Barton et al., 1995). The overburden stress in the Groningen field is the largest principal stress component (i.e., normal faulting regime), and the smallest horizontal principal stress component is oriented in an N070 direction, based on in-situ stress measurements and focal mechanisms of a subset of studied induced seismic events (van Gent et al., 2009). The vertical stress component acting on the fault plane is, therefore, a function of the overburden stress and the fault dip angle, and the horizontal stress component is a function of the angle between the fault strike and the regional minimum principal stress direction. Although there is a scatter in the orientation distribution of faults, most faults are observed to strike in NNW-SSE and ENE-WSW directions (Figure 13). Note that the ENE-WSW striking faults are sub-parallel to the regional minimum principal stress direction (N070) and may, therefore, be more prone to failure.

Second, the fault geometry influences the peak Coulomb stress as irregularities on a fault plane can form weak points that fail relatively easily, whereas a perfectly planar fault surface may sustain a larger Coulomb stress change before failing, but the subsequent slip patch of such a perfectly planar fault is likely larger than the slip patch of a fault with an irregular geometry.

In addition to surface size and orientation, fault offset can impact the initial Coulomb stress. Faults contained within the reservoir have sand-on-sand contacts. These contacts are unlikely to result in peak normal or shear stress concentrations along the fault, but faults that extend into the Zechstein formation overlying the reservoir can experience salt migrating into the fault zone, changing the fault frictional properties of those faults compared to the faults with sand-on-sand contacts. Fault frictional properties can also vary depending on the gas saturation of the fluid that permeates through the fault (Candela et al., 2018) and on the amount of shale in the fault zone, as shale can 'lubricate' faults resulting in aseismic rather than seismic slip (Buttinelli et al., 2016; Candela et al., 2018). As in-situ friction properties of faults are not directly measurable, we use the shale volume in the reservoir rock, the ratio between the gas column and water column heights and the gas saturation in the aquifer as possible proxies for fault friction variations throughout the field (Figure 14).

Last, fault density is a relevant parameter impacting spatial initial Coulomb stress variations in the reservoir: In the hypothetical situation where all faults have the same friction and orientation properties, the spatial distribution of induced seismicity is likely correlated to the fault density.

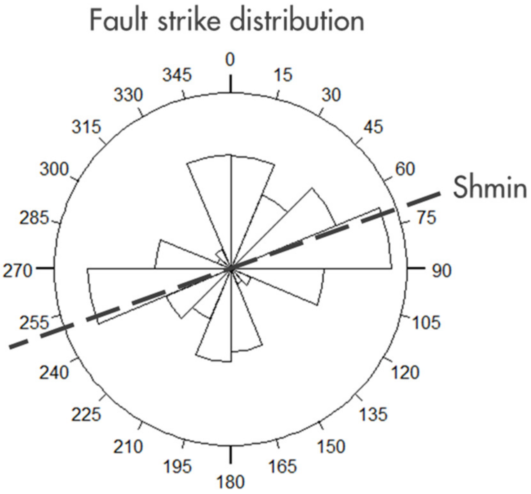


Figure 13: Rose diagram of the fault azimuth distribution sampled from the Petrel model (100 m spacing). As longer faults will have more sampling points, the rose diagram is length-weighted. The dashed line indicates the regional horizontal minimum principal stress direction.

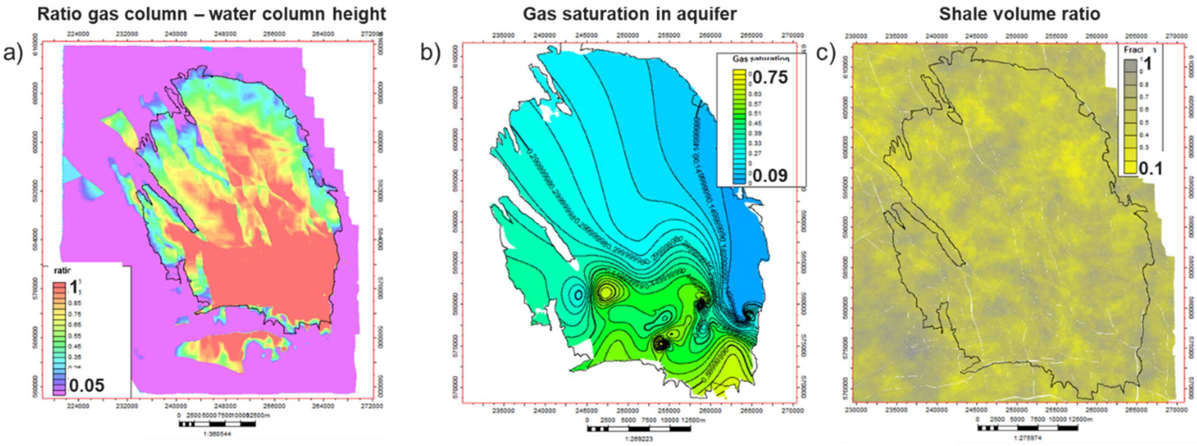


Figure 14 Potential reservoir proxies for fault friction variations: a) Ratio between the gas column and water column heights, normalized over reservoir thickness, as an indicator of how much of the reservoir and fault rock is exposed to water versus gas. Note that the regions outside the Groningen field outline (black outline) are not taken into account; b) Gas saturation in the carboniferous, underlying the reservoir, based on extrapolation of saturation measurements in several wells that penetrated the carboniferous. Data based on observations from five wells, mostly in the south of the field; c) Shale (vs sand) ratio. Grey indicates 100% shale.

The beforementioned fault properties are used as input features for use in the Machine Learning workflow and are obtained from two data sources:

1. The Petrel geological model of the Groningen field (Figure 15). In this model, faults have been manually interpreted from 3-D seismic reflection data. The interpreted faults have been converted into a gridded model with a spatial resolution of 100 x 100 meters, and for each cell, fault attributes are calculated. The fault intensity is calculated by taking the cumulative length of fault traces within an area (P_{21} following the intensity definitions by Dershowitz and Herda (1992)). The conversion of fault interpretations to a fault model has as benefit that fault attributes including orientation (strike and dip), offset and depth are available at a constant 100 m spacing, but the disadvantage of this approach is that in the

conversion from interpretation to model, manual adjustments to the fault network may have been made to avoid gridding issues. To ensure that the anisotropy in the fault network is maintained in the ML models, the faults and their attributes are divided into two orientation families, based on the paleostress orientation. One set contains all faults with a strike of N070±45° or N250±45° ("ESE-WNW"), and the other set contains faults striking N160±45° or N340±45° ("NNW-SSE").

2. Additional features are derived directly from the seismic reflection data: To minimize data sampling artefacts resulting from human bias when interpreting faults and modelling bias when converting the interpreted faults into a fault network model, seismic attributes are extracted from the 3-D seismic data that act as proxies for fault attributes (Figure 16). The selected attributes are the surface gradient, the mean amplitude in the reservoir interval and variance, which is calculated using an inline/crossline range of three, and mild vertical smoothing (15 samples). These attributes are extracted from the seismic amplitude data around the level of the top reservoir surface and represent proxies for seismic-scale structural deformation (i.e., faults). The disadvantage of these attributes is that they are dimensionless proxies for faults and do not provide absolute measures of fault orientation, density or offset.

Further details regarding the origin and uncertainty of these sources are given in section 3.4 and Appendix B.

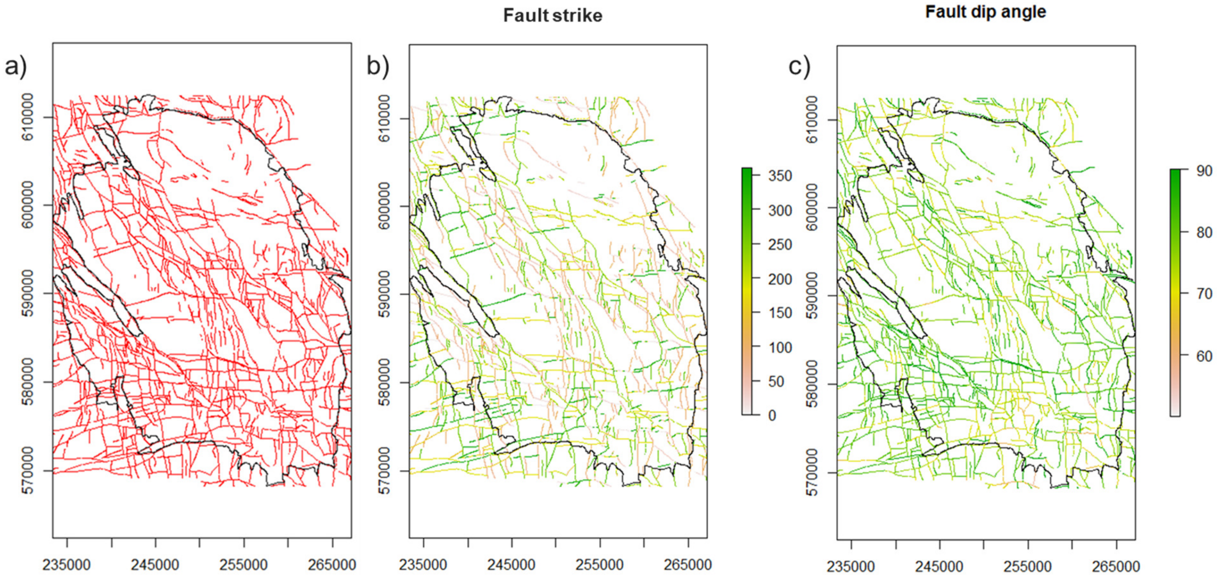


Figure 15 Fault network in the Petrel reservoir model: a) Top view of the fault network at the intersection with the top reservoir surface; b) Fault strike orientation between 0 and 360°; c) Fault dip angle orientation.

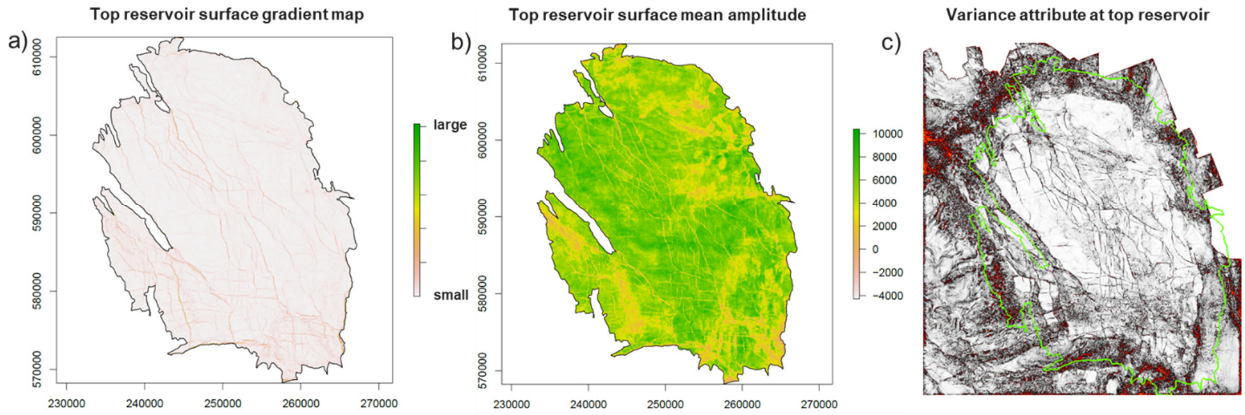


Figure 16 Seismic attributes as proxies for structural deformation in the reservoir: a) Gradient map of the top reservoir surface; b) Mean amplitude of the reservoir interval; c) Variance attribute calculated around the top reservoir surface. Hot colours indicate high variance.

Reservoir rock properties

The initial Coulomb stress may be influenced by rock property variations throughout the reservoir. There are several rock properties, measured in wells and interpolated between wells using correlations with acoustic impedance, that may be used as proxy features for these Coulomb stress variations. These features are available from the geological reservoir model, at a 100 m resolution:

- Shale volume (Figure 14c), as shale and sand have different Poisson's ratios;
- Porosity, which affects the diffusivity of pressure changes in the reservoir (Figure 17a);
- Uniaxial compressibility, which is partly constrained by inversion of geodetic data (i.e., not correlated to acoustic impedance) (Figure 17b-d).

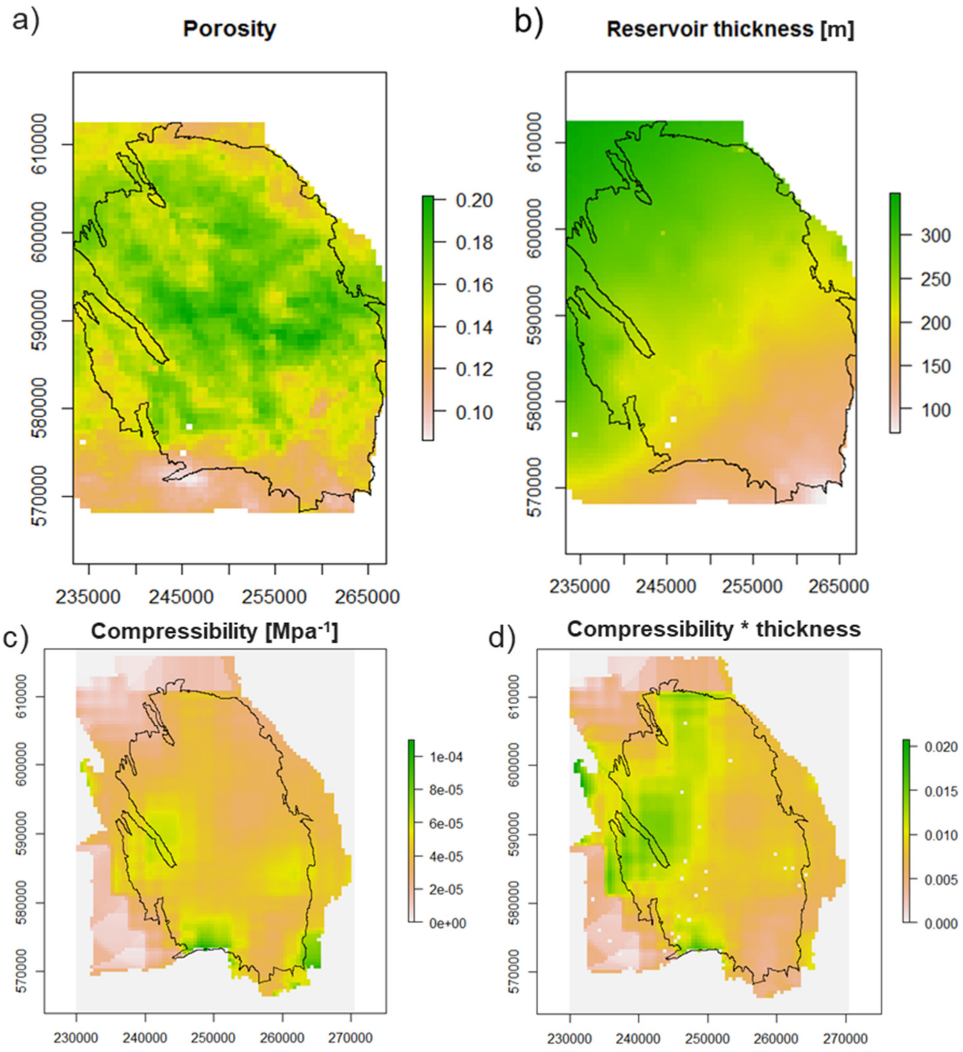


Figure 17 Reservoir properties: a) Porosity [-]; b) Thickness [m]; c) Compressibility [MPa^{-1}]; d) Compressibility multiplied by reservoir thickness [m/MPa].

Overburden properties

The vertical stress is a function of depth and overburden density ($\sigma_v = \rho g z$; Figure 18). The overburden density is the largest uncertainty for quantifying overburden stress, as the thickness of the overburden is relatively well-constrained by seismic reflection and well data, whereas density is only calculated from petrophysical logs in the overburden. The overlying Zechstein formation is the main source of overburden density uncertainty, as its density is lower than that of other overburden formations, and its thickness and density (using velocity as a proxy) are spatially varying as the Zechstein consists of a combination of halite and anhydrite (Figure 19; Table 11). The Zechstein thickness and velocity maps are used as proxies for overburden density variations, in addition to a geomechanical overburden stress model obtained from a geomechanical Finite-Element model of the Groningen reservoir (Appendix B; Figure 18b).

Table 11: Properties of the different salt types in the Zechstein Fm. (Romijn, 2017), compared to the density of other overburden formations.

Formation	Density [kg/m^3]	Velocity (V_p) [m/s]
Halite	2090	4400
Anhydrite (floaters and basal layer)	2810	5900

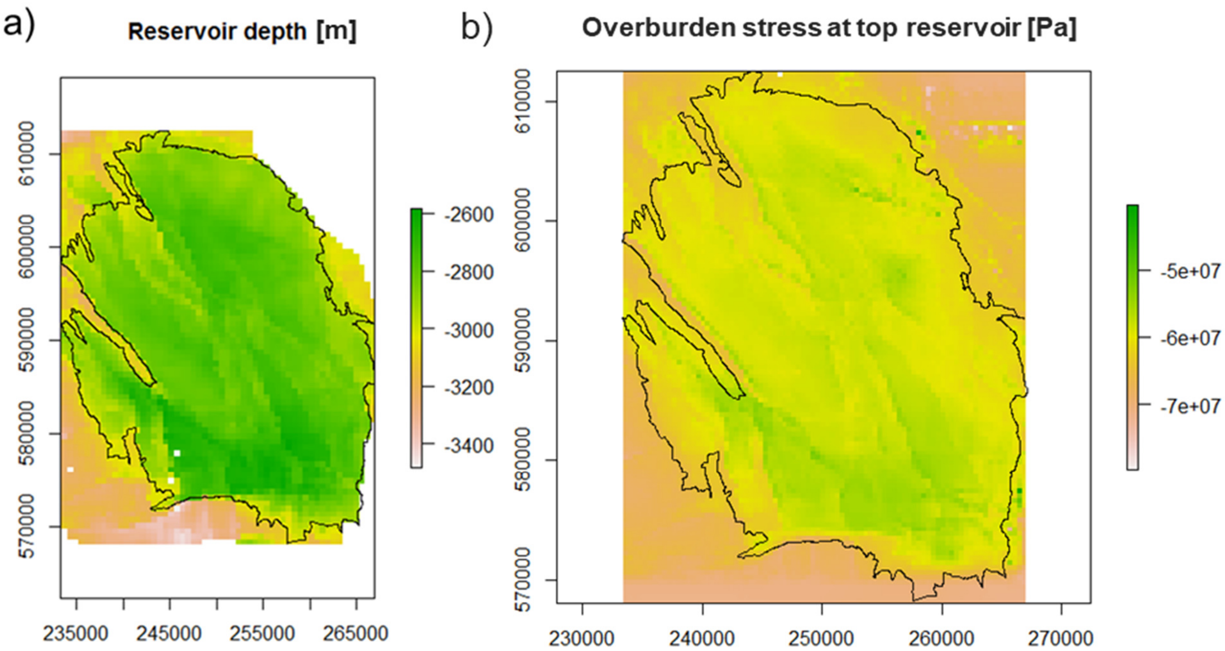


Figure 18: Overburden stress distribution in Pascal at the depth of the top reservoir horizon. Compressional stresses are represented as negative values.

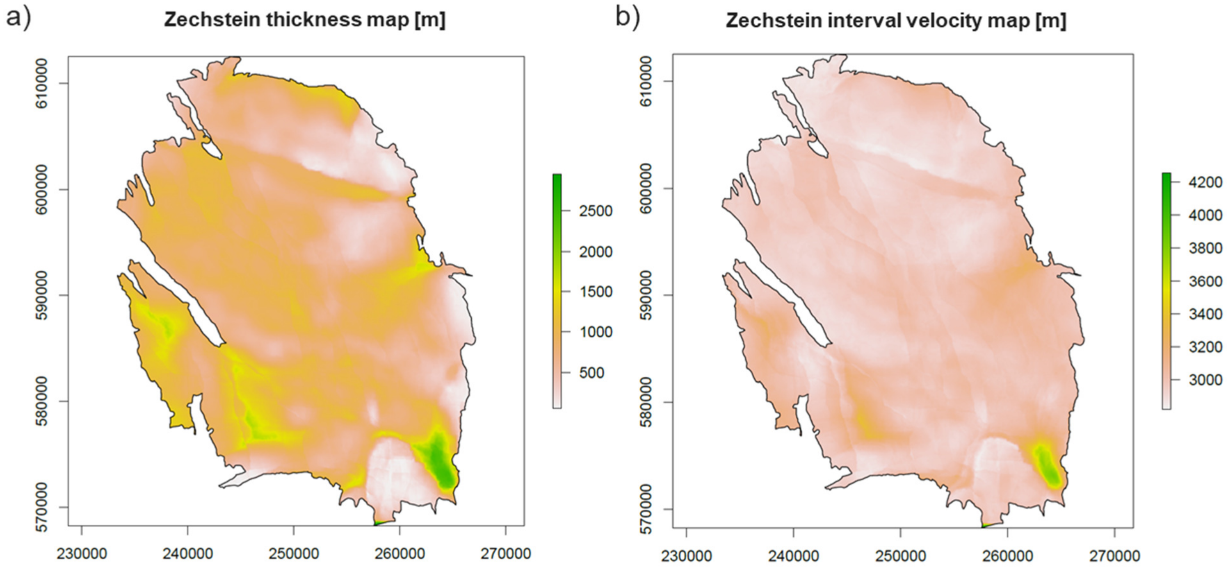


Figure 19 Zechstein properties obtained from seismic data (25 x 25 m resolution): a) Zechstein thickness mapped from seismic reflectors; b) Zechstein interval velocity obtained from seismic.

3.3. Spatiotemporal features as proxies for Coulomb stress changes

The change in Coulomb stress ΔC depends on changes in pore pressure ΔP, which is driven by the volume of gas production and the pressure diffusion through the reservoir, combined with the distribution of geometric and elastic heterogeneities throughout the reservoir. The main sources of these heterogeneities, the pre-existing fault network and reservoir compressibility, do not change significantly during gas production, but pore pressure does change during production. Gas production is measured at a high temporal resolution, but spatially sparse as data is only available for clusters of wells. Production-related features are therefore only included using a field-wide averages.

The spatial pressure changes resulting from gas production are available throughout the reservoir, as pressure change is modelled using a reservoir flow simulation code (MoReS). The static geological model forms the input data for the initial pre-production (static) reservoir state and describes the geometry of the reservoir and the time-invariant flow properties of the rock (i.e., porosity, permeability). The dynamic model is initialized using initial pressures measured from wells and using a fluid property model based on Pressure, Volume and Temperature (PVT) data of fluid samples. Based on these initial pressures and fluid and rock properties, a numerical finite-difference solver calculates the pressure gradient throughout the reservoir as a function of well production rates and volumes. The resulting model describes in 3-D for each grid cell the changes in saturation, fluid composition and pressure, calibrated to available historical well data.

The temporal analysis of ML for seismicity (Limbeck, et al., 2018) made extensive use of a range of dynamic features from the reservoir flow model, such as pressures, production rates, changes in hydrocarbon column thickness (*HCT*) and compaction. However, in the spatiotemporal ML models, the pressure is the only spatiotemporal feature used from the dynamic reservoir flow model, as other features are primarily derived from pressure. Production data is only used as a field-wide aggregated average.

3.4. Overview of spatiotemporal features for ML models

We summarize here all features that were presented in the sections above and that form the predictor inputs for the ML models, with a brief description summarizing how these features relate back to the Coulomb stress model, and from which data source they were obtained (Table 12). For each data or model source from which the features are derived, the model or data origin and the calculation methods for obtaining the different features are summarized (Table 13). See Appendix B for a more elaborate discussion of data and model origin, and for several alternatives regretted features and the motivation for discarding these features.

Table 12: Features used in the spatiotemporal ML approach

FEATURE NAME	DESCRIPTION	DATA SOURCE	MODEL LABEL
Fault density	P21 (cumulative fault length per grid cell area) fault intensity per reservoir grid cell	Fault model from the Petrel geological model	F_ALL.Density
Fault dip angle	Fault dip angle between 0 and 90°		F_ALL.Dip.mean
Fault strike angle	Fault strike angle between 0 and 360°		F_ALL.Dip.Azimuth.mean

Fault offset	Vertical reservoir offset along faults (in meters)		F_ALL.Reservoir.Offset.mean
Fault reservoir thickness	Average reservoir thickness at the location of the faults		F_ALL.Av.Reservoir.Thickness.Mean
NNW-SSE fault density	The density of the orientation subset of faults with a strike within the range N160±45° or N340±45°		F_NS.Density
NNW-SSE fault dip angle	Fault dip angle between 0 and 90° for the orientation subset of N160±45° (or N340±45°) striking faults		F_NS.Dip.mean
NNW-SSE fault strike angle	Fault strike angle between 0 and 360° for the N160±45°/ N340±45° subset of fault strikes.		F_NS.Dip.Azimuth.mean
NNW-SSE fault offset	Vertical reservoir offset along faults (in meters), for the N160±45°/N340±45° striking group of faults		F_NS.Reservoir.Offset.mean
Reservoir thickness at the location of NNW-SSE faults	Average reservoir thickness at the location of the N160±45°/N340±45° striking faults		F_NS.Av.Reservoir.Thickness.Mean
ENE-WSW fault density	The density of the orientation subset of N070±45°/N250±45° striking faults		F_EW.Density
ENE-WSW fault dip angle	Fault dip angle between 0 and 90° for the orientation subset of N070±45°/N250±45° striking faults		F_EW.Dip.mean
ENE-WSW fault strike angle	Fault strike angle between 0 and 360° (N070±45°/N250±45° striking faults)		F_EW.Dip.Azimuth.mean
ENE-WSW fault offset	Vertical reservoir offset along faults (in meters), for the N070±45°/N250±45° striking group of faults		F_EW.Reservoir.Offset.mean
Reservoir thickness at the location of ENE-WSW faults	Average reservoir thickness at the location of the N070±45°/N250±45° striking faults		F_EW.Av.Reservoir.Thickness.Mean
Surface gradient	Surface gradient (seismic dip map) of the top reservoir surface as a proxy for fault locations	Seismic attribute	SeisDip.val.mean
Mean amplitude	Mean seismic amplitude of the reservoir interval as a proxy for faults and other structural deformation features		SeisMeanAmp.val.mean
Variance volume attribute	Seismic volume attribute is capturing the variance around the depth of the top reservoir formation, as a proxy for structural deformation in the reservoir.		SeisVar.val.mean
Interval velocity Zechstein formation	Interval velocity (in m/s) for the seismic interval of the Zechstein formation, as a proxy		SeisVint.val.mean

	for lateral density variations in the Zechstein, caused by the anhydrite floaters.		
Zechstein formation thickness	Thickness (in meters) of the Zechstein formation, as a proxy for lateral overburden density variations resulting from the relatively low-density salt.		SeisZechThick.val.mean
Vshale	The amount of shale versus sandstone in the reservoir rock. The shale ratio may affect the friction behaviour of faults, as shale in the fault core increases the probability of aseismic versus seismic slip, and the ratio between shale and sand is a potential proxy for spatial variations in the elastic rock properties (Poisson's ratio).	Petrel geological model (well data interpolated using acoustic impedance)	avg_vsh.val.mean
Gas column height versus water column height	The ratio between the gas column height and water column height at each individual location in the reservoir (prior to production), as a potential proxy for lateral variations in the fault friction behaviour.	Petrel geological model (interpolated well data)	gc_vs_wc.val.mean
Gas saturation in the aquifer	Gas saturation in the Carboniferous (aquifer), as a potential proxy for lateral variations in the fault friction behaviour.	Petrel geological model (limited well data, interpolated using kriging without constraint to other features).	sg_carb.val.mean
Porosity	Mean reservoir porosity (weighted vertical average)	Petrel geological model (well data interpolated using acoustic impedance cross-correlation)	statRes.porosity2D.mean
Compressibility	Reservoir rock compressibility [MPa^{-1}] for calculating compaction and strain thickness from pressure changes.	Inversion from subsidence data.	statRes.cm
Reservoir thickness	Reservoir thickness in meters, used for calculating compaction and strain thickness.	Petrel geological model interpreted seismic	statRes.thickness2D.mean
Top reservoir depth	The depth of the top reservoir surface in meters.		statRes.Ztop.mean
Top reservoir surface gradient	The average gradient at each location calculated from the gradient in the adjacent cells. Only considers average absolute gradient without orientation.	Calculated from top reservoir surface data.	statRes.topoGrad.mean

Mean overburden stress	Overburden stress at the top reservoir level in Pascal (before production).	3-D Finite Element model (COMSOL)	Sv.Sv.mean
Absolute reservoir pore pressure	Average reservoir pressure, based on the vertical weighted average	MoReS reservoir flow model	weighted.mean.P
Change in pore pressure over time	The first temporal difference of pressure		weighted.mean.dPdT
Rate of pore pressure change over time	The second temporal difference in pressure		weighted.mean.d2PdT2
Field-averaged pressure	Field-wide average reservoir pressure		weighted.mean.P.agg
Field-average pressure change	Field-wide averaged change in pressure		weighted.mean.dP.aggdT
Field-average rate of pressure change	Field-averaged second derivative of pore pressure		weighted.mean.d2P.aggdT2
Produced gas volume	Field-wide total volume of produced gas in a period of time (in m^3)		sum.Q.Gas.M3
Average production rate	Field-wide average production rate over a time period (in m^3)		sum.dQdT.Gas.M3
Variance in production rate	Field-wide variance production rate (m^3)		variance.dQdT.Gas.M3
Change in production rate	The second derivative of field-averaged production		sum.d2QdT2.Gas.M3
Variance in production rate change	The variance of the second derivative of field-averaged production		variance.d2QdT2.Gas.M3
Compaction	Amount of compaction within a time step (i.e., incremental compaction) in meters, using MoReS pressure, reservoir thickness and compressibility.	Calculated from MoReS pore pressures using compressibility and reservoir thickness features.	mean.C
Change in compaction rate	The second temporal difference in compaction		mean.d2CdT2
Cumulative compaction	Cumulative compaction since the start of production		mean.cumC
X coordinate	Coordinate in meters, using the Rijksdriehoek coordinate system. Regularly spaced grid.	Calculated from resampled spatial input grids.	RD_X
Y coordinate	Coordinate in meters, using the Rijksdriehoek coordinate system. Regularly spaced grid.		RD_Y

Table 13 Summary for all the data/model sources from which features in the previous table are extracted, including the origin of the model or data, the format, and resolution in which the data is available, and how ML features are calculated from this data format.

DATA SOURCE	ORIGIN	FORMAT	RESOLU TION	FEATURE CALCULATION
Fault model from the Petrel geological model	Petrel geological grid (i.e., regular grid at 100x100 horizontal resolution and variable vertical resolution). The 3-D fault model is converted to 2-D by extracting fault line traces at the intersection between top reservoir and faults. The Petrel fault model is based on seismic fault interpretation with manual post-processing to avoid gridding issues.	Fault length traces discretized into unstructured point set at a 100 m resolution, with fault attributes for each point.	100x100 m	• Mapping of fault points per fault orientation set to a grid. • Averaging of fault attributes (e.g., orientation, thickness) for multiple points within a cell. • Fault density is calculated using the sum of points within a cell.
Seismic attribute	3-D seismic volume attributes calculated from seismic amplitudes around the top reservoir (surface attributes) or from the reservoir interval.	2-D point set with x, y data, and attribute values.	25x25m	Mapping of the points to the ML grid, using averaging for multiple points within a single cell.
Petrel geological model (well data interpolated using acoustic impedance)	Petrel geological 3-D model. Property values are calculated from well log data and interpolated in 3-D using a correlation with acoustic impedance. Converted to map features using (cell) volume weighted averaging in the vertical direction. The only exception is sg_carb.val.mean, which is interpolated using kriging instead of acoustic impedance.	Regular gridded 2-D point set with x, y center coordinates per cell and attribute values.	100x100m	• Mapping of the points to the ML grid, using averaging for multiple points within a single cell. • The topographic gradient in each cell is calculated from the average surface gradient of the 8 neighbouring cells.
Compressibility inverted from subsidence data.	Compressibility model inverted from geodetic (subsidence) data.	2-D regular point set with x, y and compressibility values.	500x500m resolution	Mapping of the points to the ML grid, using averaging for multiple points within a single cell.
3-D Finite Element model (COMSOL)	Model geometry is based on the horizon and fault surfaces interpreted from seismic. Mechanical rock properties are obtained from the static geological model.	2-D point set of the stress state at the top reservoir, variable resolution.	Variable, an average of 150 m, minimum of 25 m (seismic resolution)	Mapping of the points to the ML grid, using averaging for multiple points within a single cell.

MoReS reservoir flow model	Model geometry and static properties are upscaled from the geological model. The fluid model obtained from PVT data, and pressure, saturation, temperature and flow are modelled based on calibration to measurements at wells.	3-D point set with cell center points, cell volume, and dynamic attributes.	Variable, an average of 650 m, minimum of 80 m.	• Vertical aggregation from 3-D to 2-D using cell-volume weighting, using averaging (e.g., pressure) or summation (e.g., production) • 2-D upscaling or downscaling of reservoir simulation grid to ML grid using averaging or summation. • The depletion thickness and vertical strain thickness baselines are calculated using compressibility and thickness data at the ML grid scale.

4. Spatiotemporal Earthquake Data and Target Definition

The goal of this study is to forecast seismicity event rates in space and time: the number of earthquakes within a certain time interval, within a certain region, above a certain minimum magnitude. This section describes earthquake measurements and the choices made to generate the target from these measurements.

4.1. Earthquake measurements

The KNMI (the Royal Netherlands Meteorological Institute) has seismicity monitoring stations throughout the Netherlands and specifically in Groningen[3]. The network is described in more detail in, e.g., Dost et al. (2012) and Dost and Haak (2002). Measurements from this network are automatically processed by KNMI and earthquakes detected are formally published in a catalogue[4], which we use as a source for seismic events. The induced seismicity catalogue has a straightforward structure as shown in Table 14: The data is provided in a tabular form with each row representing an event, with event date and time, location, latitude, longitude, depth, magnitude and evaluation mode. Most of these fields are self-explanatory, possibly except for the location field[5] but that field is not used in our analysis.

Table 14 KNMI induced earthquake catalogue data structure

Date	Time	Location	Lat	Lon	Depth	Mag	Eval mode
1986-dec-26	07h47m51s	Assen	52.992	6.548	1	2.8	Manual
1987-dec-14	20h49m48s	Hooghalen	52.928	6.552	1.5	2.5	Manual
…	…	…	…	…	…	…	…

4.2. Uncertainties

The number of sensors in the seismic sensor network, their locations and the data processing procedures used influence detection sensitivities and location uncertainties. As the network has been extended over time, detection sensitivity and location uncertainties vary over time. Table 15 provides an overview of sensitivities as reported by the KNMI (see, e.g., Dost et al., 2012, 2017; Kraaijpoel et al., 2015; Spetzler and Dost, 2017). In general, the horizontal location uncertainty is around 1 km, and the vertical uncertainty is between 1-2 km. Given the sizeable vertical uncertainty, vertical locations are pre-set to 3 km for nearly all events.

[3] For an overview of these stations, see https://www.knmi.nl/nederland-nu/seismologie/stations.

[4] Catalogue available at https://www.knmi.nl/kennis-en-datacentrum/dataset/aardbevingscatalogus.

[5] Up to November 30, 2016 the location field described the city or village centre nearest to the event, whilst as of December 1, 2016 the municipality border within which the event took place is registered.

Table 15 KNMI Seismic Sensor Network developments over time

Time	Detection	Localisation	Comments
Since 1995	≥ 1.5	≥ 2.3-1.5	Network installed (8 borehole stations in Northern Netherlands)
±2010	Processing software upgrade, real-time continuous data transmission		
2009-2010	≥ 1.0	≥ 1.5	6 additional borehole stations in Northern Netherlands
2015-2017		≥ ~0.5	Major extension: 64 additional borehole stations in Northern Netherlands

4.3. Choice of minimum magnitude M_{min}, temporal interval and temporal aggregation period T_{agg}

The magnitude of completeness M_c of a sensor network is usually defined as the lowest value of the moment magnitude of an event for it to be detected with 100% reliability. Event counts with a moment magnitude below M_c are incomplete, which in principle does not pose a problem for machine learning algorithms as long as M_c is constant over time: Algorithms would simply forecast observed seismicity. However, with the detection sensitivity increasing over time, an increase in the detection of earthquakes is a combination between a change in seismicity and a change in detection sensitivity. As this effect is strongest for low magnitude seismicity a minimum magnitude cut-off M_{min} is chosen, only earthquakes with a magnitude equal to or higher than M_{min} are taken into account. A sensible choice for M_{min} is the magnitude of completeness M_c – this choice would ensure that all signal picked up comes from seismicity instead of sensor network sensitivity changes. Given the improvements in the sensor network over time, the choice of M_c and the start of the temporal interval T_{start} are coupled: a later T_{start} might allow for a lower M_c and vice versa. The choice for both parameters is, of course, driven by the desire to use as much of the data as possible, while avoiding the introduction of bias.

Following the extensive analysis of M_{min} in section 3.2 of Limbeck et al. (2018), we proceed with the following choices for M_{min}, T_{start} and T_{agg}:

- Following both KNMI reported M_c values and the PSHRA default $M_{min} = 1.5$ with $T_{start} = 1995$ are used with a $T_{agg} =$ 3 months.
- For consistency with earlier work, $T_{end} = 2016$. Event data from 2017 and 2018 is hence not used for model training, such that all model forecasts from January 1 2017 can be compared to observations to assess out-of-sample performance.

4.4. Choice of geospatial interval and forecasting target

Following Limbeck et al. (2018), the geospatial area of interest is delineated by the outline of the Groningen field Outline (GFO; Figure 20). Note that only reservoir-related induced events are considered when using this boundary and that events in the aquifer are excluded.

The forecasting target used in this study is the earthquake rate, i.e., the number of earthquakes in three months (equivalent with earthquake count for uniform temporal intervals).

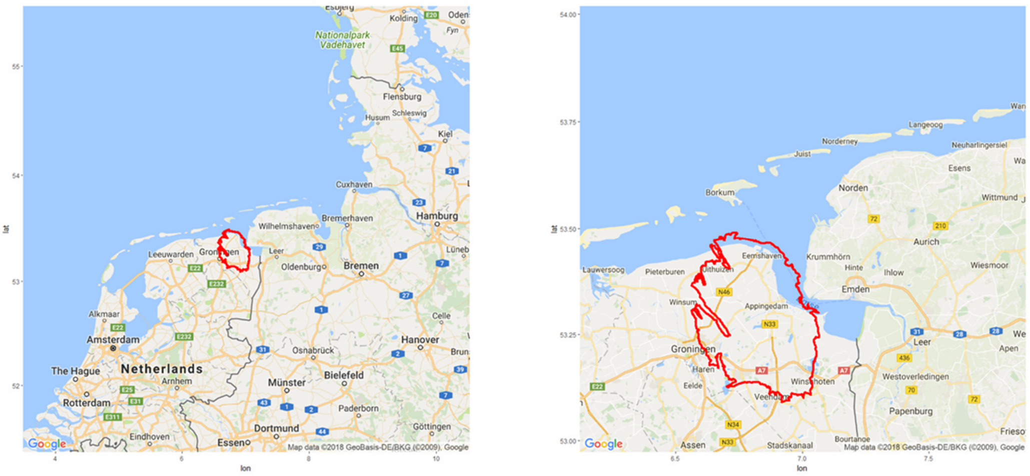


Figure 20: Groningen field Outline (GFO) geospatial view Google Maps (2018)[6]

[6] Map data © GeoBasis-DE/BKG (© 2009) Google. Google Maps image retrieved from: http://maps.googleapis.com/maps/api/staticmap?center=53.5,7&zoom=9&size=640x640&scale=2&maptype=roadmap&language=en-EN&sensor=false

5. Spatiotemporal Experimental Setup

The spatiotemporal models require that the spatial and temporal features defined in the previous sections are combined into a single feature data table. The processing of different spatial data at different resolutions into a data table suitable for ML modelling is handled through gridding (discussed previously in section 3.4) and smoothing of data (section 5.1), and conversion of the data into machine learning features (section 5.2).

To optimize simulation times, feature down-selection is applied based on the extent to which features are correlated to the forecasting target and to each other (section 5.3). In addition, conclusions from earlier work help to down-select the meta-parameters (5.4), resulting in a final subset of experiments summarized in section 5.5. Referring back to the high-level workflow (Figure 3), the experimental setup to go from raw data to feature creation and aggregation can be visualized in more detail (Figure 21). The steps from this figure are explained in more detail in the next subsections.

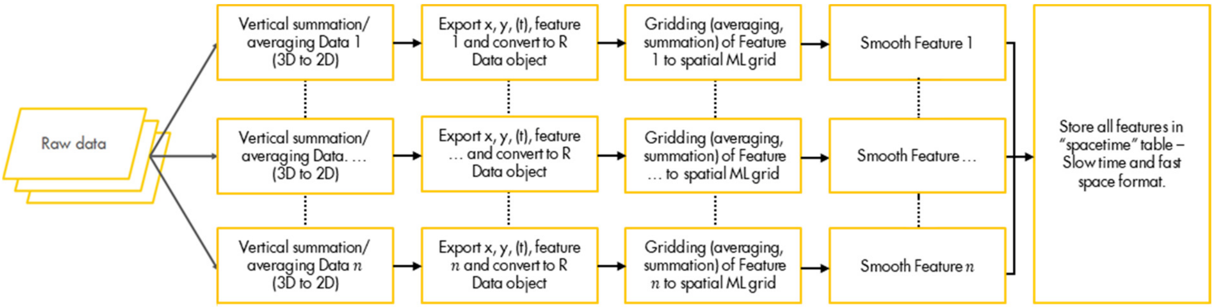

Raw data
Vertical summation/ averaging Data 1 (3D to 2D)
Export x, y, (t), feature 1 and convert to R Data object
Gridding (averaging, summation) of Feature 1 to spatial ML grid
Smooth Feature 1
Vertical summation/ averaging Data. ... (3D to 2D)
Export x, y, (t), feature ... and convert to R Data object
Gridding (averaging, summation) of Feature ... to spatial ML grid
Smooth Feature ...
Vertical summation/ averaging Data *n* (3D to 2D)
Export x, y, (t), feature *n* and convert to R Data object
Gridding (averaging, summation) of Feature *n* to spatial ML grid
Smooth Feature *n*
Store all features in "spacetime" table – Slow time and fast space format.

Figure 21 Flowchart of the spatiotemporal experiment setup

5.1. Spatial smoothing of the target and input data

Smoothing is applied to account for varying resolutions of the spatial features. Spatial smoothing requires several trade-offs, between ensuring that each cell contains data for all features, avoiding discretization errors where the results of the models are influenced by the grid resolution, keeping computation times within acceptable limits and dealing with spatially sparse seismic events. The spatiotemporal predictor features are obtained from different data sources with varying resolutions (Table 13):

- Seismic reflection data: 25 m;
- Geological reservoir model (based on well data interpolated using seismic): 100 m;
- Dynamic reservoir model: On average 400 m, with local refinements around the wells;
- Geomechanical model: variable resolution between seismic resolution and dynamic reservoir model resolution;
- Geodetic data has a non-uniform but relatively high spatial resolution at the surface, but the reservoir compaction inverted from this data has an approximate bandwidth of 3000 m, constrained by the depth of the reservoir.

All input features are projected onto a single regular-spaced grid with a 1500 m resolution through averaging or summation depending on the feature. At this resolution, the property distributions are sufficiently smooth without generation of discretization artefacts. Note that further increasing the number of cells comes at a significant computational cost. The value of each feature at the grid cell area and the grid cell centre locations form the actual features for the ML models. The gridding approach for each data source is described in Table 13 in section 3.4. Once all data is mapped to the specified grid resolution, spatial smoothing is applied to account for spatial uncertainties (Figure 22). A symmetric Gaussian kernel smoother is used for the predictor features. To account for anisotropy in the fault network, smoothing is applied separately to the two fault orientation groups of NNW-SSE and ENE-WSW striking faults. The smoothing bandwidth is guided by the lowest resolution data, i.e., the compressibility and compaction features, as reservoir compressibility is modelled through inversion of subsidence data. The distance between subsidence data and reservoir compressibility (i.e., reservoir depth) limits the maximum resolution in compressibility that can be resolved. We investigate a range of bandwidths between 1500 and 5000 meters. This range of bandwidths includes the value of 3500 m, which was found to be the optimal bandwidth by Bourne and Oates (2015).

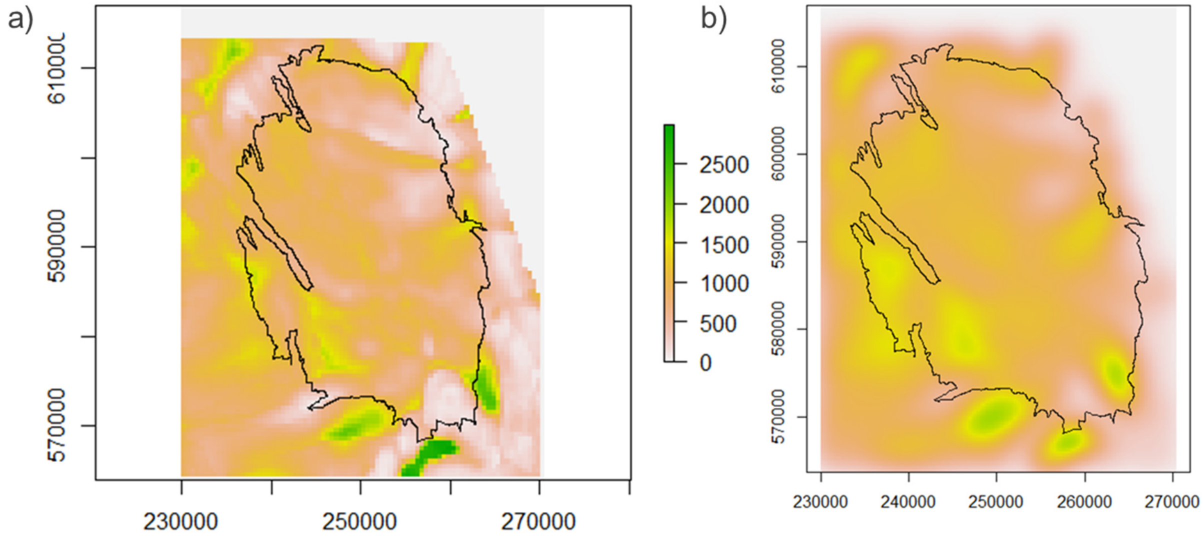


Figure 22: Example of smoothing of the Zechstein thickness map (thickness in meters), using a grid resolution of 500 m (a) and a kernel smoothing bandwidth of 1500 m.

In training the models, the target is also smoothed to a resolution identical to the resolution used for the predictor feature smoothing, but model validation is done using the actual event counts without smoothing. It is important to note that with smoothing there could be some loss of information within the field outline which might cause the sum of smoothed event counts within the field outline to not be the same as the recorded event count. Several checks have been implemented to ensure that the impact is as minimal as possible. Firstly, we have made sure that only events that occur within the limits of the Groningen field outline as seen in Figure 23 are taken into account. Second, we have added extra padding (i.e., additional) cells to the rectangular grid over which events are initially smoothed to make sure that events that occur close to the borders are not cut out, and finally, an additional buffer was added to the final cropping to make sure that no empty cells are taken into the final experiment.

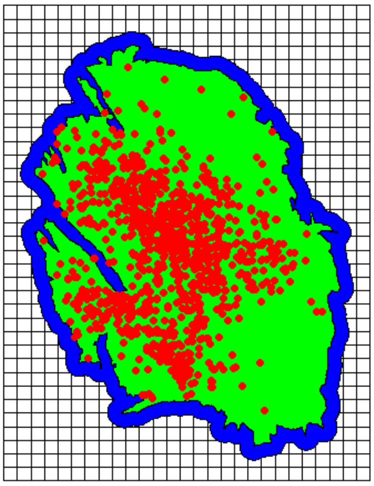


Figure 23: Events considered are in red, the Groningen outline in green with the buffer in blue and the rectangular grid over which events are smoothed to the 500 m spatial event resolution in the background.

After taking these preventive measures into account, we find that there is still some small but to the best of our knowledge immaterial difference between the sum of smoothed event counts and the observed event counts, with an error of 0.6% on average. This error is calculated as the difference between the field-wide average of a feature value compared to the post-smoothing field-wide average. Based on this error, the spatially aggregated temporal event count forecast is expected to be slightly lower than the temporal event count from Limbeck et al. (2018). However, we note that if the sum of smoothed event counts is rounded to the nearest integer then both counts, smoothed and observed match 100%. One aspect to bear in mind is that further increasing the buffer could most likely eliminate the error altogether, however here the caveat is that by selecting our observation cells based on such a more extensive area would not only inflate the number of cells hence increasing computational time significantly on an already time consuming process but also that there may be events associated with these cells that will not be taken into account as the models consider only events within the field outline.

5.2. Spatiotemporal Binning with Spatial Coordinates as Features

The spatiotemporal dataset is integrated into a full spatiotemporal matrix using the same data conversion conventions used for the event rate report (Limbeck et al., 2018). The data has been structured in long form, where the full spatiotemporal information is stored for each location for each timestamp, with other columns providing the specific spatiotemporal features (Table 16). More specifically the data is stored in slow time and fast space format to use spacetime terminology, meaning that for each date all locations are shown before moving on to the next date rather than all dates for one location and then moving on to the next location**Error! Reference source not found.**. This matrix takes the form of a full spacetime grid of observations for spatial features (points, lines, polygons, grid cells) s_i , $i = 1,...,b$ and observation b time t_j , $j = 1,...,d$ is obtained when the full set of $b \times d$ set of observations z_k is stored, with $k = 1,...,bd$.

In ML terms this results in a matrix of size $b \times d$ whe

re all the features have been aggregated as per their description in sections 3.4 and 5.1 (averaged, summed, smoothed) per grid cell and time. The allocation of feature values to a specific grid cell is done by using the coordinate information available. The methodology used is contained in the “sp” and “maptools” packages in R by which we assign the feature values to the cell that contains the coordinates linked to those values. Table 13 shows a diagram of how the table is structured, each row in the table corresponds to a cell in the grid at a specific time and all the features have been aggregated at that level. E.g., the mean pressure (weighted.mean.P) at a given row in the dataframe is the mean pressure value at the cell with centroid RD_X, RD_Y at a given Date.

Table 16 Diagram showing spacetime format used. Each row corresponds to a cell in the grid at a specific time, and all features are aggregated per cell per time.

Date	RD_X	RD_Y	Features
1995-01-01	20000	55000	…
1995-01-01	21000	56000	…
1995-01-01	22000	57000	…
1996-01-01	20000	55000	…
1996-01-01	21000	56000	…
1996-01-01	22000	57000	…

Furthermore, the centroid coordinates of each cell in the grid are passed to the model as features, to ensure that most of the models that had already been implemented such as KSVM, Random Forest and GLM can be reused while enforcing the explicit use of location for earthquake forecasting. The rationale behind using the coordinates as features is that if the location has a significant impact in the ability of the model to forecast seismicity, the models will be able to automatically derive spatial areas where seismicity is likely to occur, by also leveraging potential non-linear relations between location and the other spatial and dynamic features. However, the real impact of the coordinates must be considered since they are likely dependent on the mapped spatial properties like faults and could, in fact, obfuscate or act as a proxy for these properties. Further experiments without using the coordinates as features could provide a clearer picture of whether the coordinates are important in and by themselves, this idea is expanded upon in the recommendations section of this report.

Moreover, we would like to emphasize the fact that models are not being trained on a cell-by-cell basis. Each model used has access to all the cells in the grid that are available for training within the given resampling iteration then to make forecasts on a cell by cell basis. Notice how this approach is dramatically different from one in which models are trained with data within a specific region (with no access to information about other regions) then to make forecasts on that same region. What we are trying to do is feed the model with information about all parts of the reservoir, so that if spatial differences exist that can improve the forecasting performance, that they could be picked up by the models.

5.3. Feature (down)selection

The ML input features presented in section 3 were selected based on their relation with the Coulomb stress model for seismic events induced by compaction (Table 12). Within this list of features, there are however multiple features that may be a proxy for the same physical mechanism, such as gas column height and gas saturation in the aquifer. We, therefore, apply a feature cross-correlation cut-off based on the average and distribution of Pearson correlation coefficients per grid cell, that groups highly cross-correlated features and selects one feature out of that group as representative for that group (Figure 24, and Appendix C for the spatial and spatiotemporal correlation analysis). For spatially varying time-invariant data, the correlation is calculated in all locations and averaged. For time-dependent data, the average coefficient for all locations and all timesteps from the spatiotemporal feature table is used. Only that feature is carried further into the ML pipeline. To remain consistent with earlier work, a cross-correlation threshold of 0.8 is used (Limbeck et al., 2018). The remaining features after cross-correlation and grouping that are used in the ML models are listed in (Table 17).

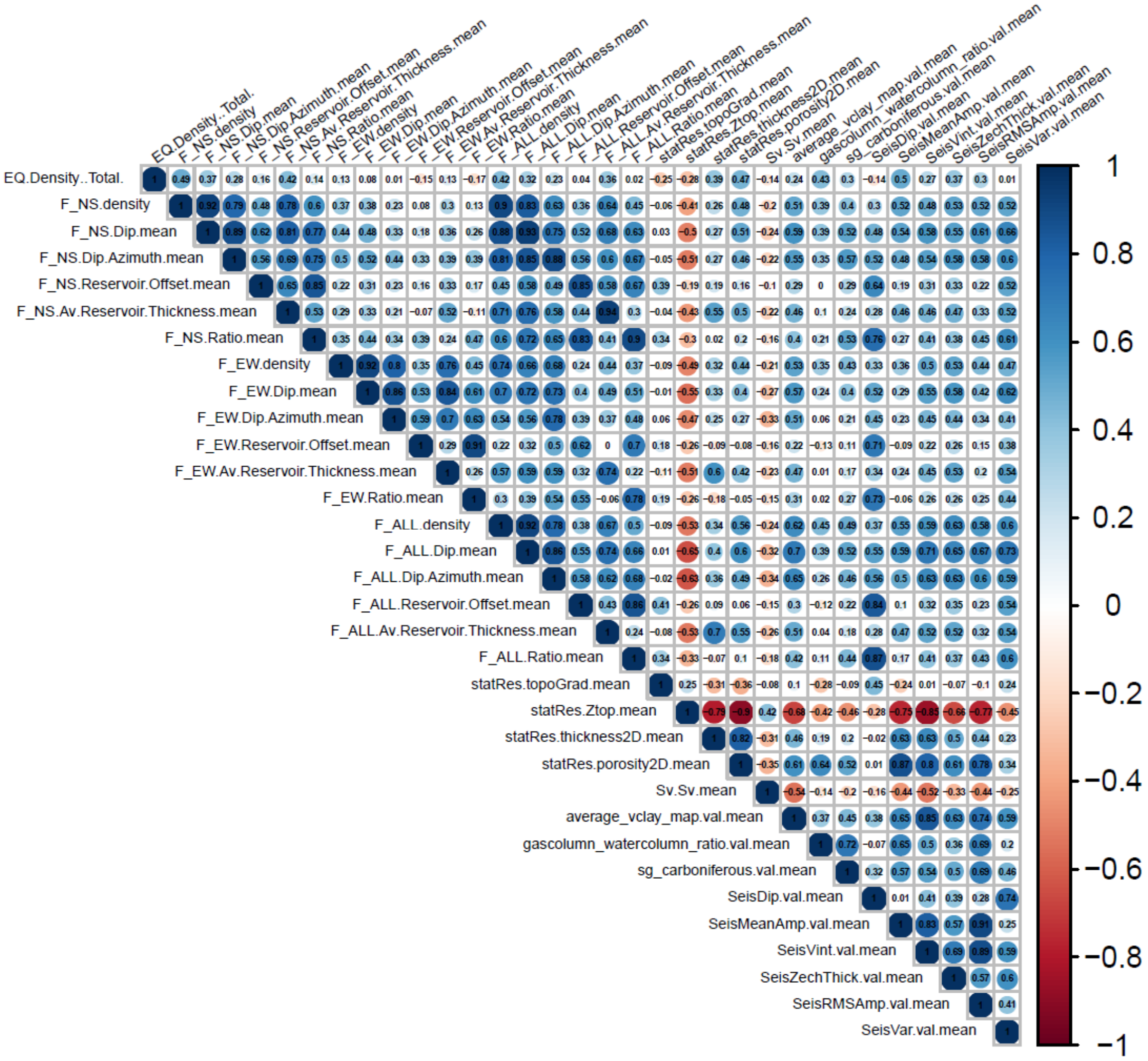


Figure 24: Correlation matrix for time-invariant spatial data (see Appendix C for the complete spatiotemporal cross-correlation matrix). Dark blue and red colours show strong correlations and anti-correlations respectively. Lighter colours and smaller circles show poor correlations between features. Label abbreviation legend is found in Table 12.

Table 17 List of downselected features following cross-correlation threshold filtering and grouping

FEATURE NAME	DESCRIPTION	MODEL LABEL
NNW-SSE fault density	The density of the orientation subset of faults with a strike within the range N160±45°/N340±45°	F_NS.Density

NNW-SSE fault offset	Vertical reservoir offset along faults (in meters), for the N160±45°/N340±45° striking group of faults	F_NS.Reservoir.Offset.mean
Reservoir thickness at location of NNW-SSE faults	Average reservoir thickness at the location of the N160±45°/N160±45° striking faults	F_NS.Av.Reservoir.Thickness.Mean
Variance volume attribute	Seismic volume attribute is capturing the variance around the depth of the top reservoir formation, as a proxy for structural deformation in the reservoir.	SeisVar.val.mean
Compressibility	Reservoir rock compressibility [MPa^{-1}] for calculating compaction and strain thickness from pressure changes.	statRes.cm
Top reservoir depth	The depth of the top reservoir surface in meters.	statRes.Ztop.mean
Top reservoir surface gradient	The average gradient at each location calculated from the gradient in the adjacent cells. Only considers average absolute gradient without orientation.	statRes.topoGrad.mean
Mean overburden stress	Overburden stress at the top reservoir level in Pascal (before production).	Sv.Sv.mean
Absolute reservoir pore pressure	Average reservoir pressure, based on the vertical weighted average	weighted.mean.P
Change in pore pressure over time	The first temporal difference of pressure	weighted.mean.dPdT
Rate of pore pressure change over time	The second temporal difference in pressure	weighted.mean.d2PdT2
Field-averaged pressure	Field-wide average reservoir pressure	weighted.mean.P.agg
Field-average pressure change	Field-wide averaged change in pressure	weighted.mean.dP.aggdT
Field-average rate of pressure change	Field-averaged second derivative of pore pressure	weighted.mean.d2P.aggdT2
Produced gas volume	Field-wide total volume of produced gas in a period of time (in m^3)	sum.Q.Gas.M3
Average production rate	Field-wide average production rate over a time period (in m^3)	sum.dQdT.Gas.M3
Variance in production rate	Field-wide variance production rate (m^3)	variance.dQdT.Gas.M3
Change in production rate	The second derivative of field-averaged production	sum.d2QdT2.Gas.M3
Variance in production rate change	The variance of the second derivative of field-averaged production	variance.d2QdT2.Gas.M3

Compaction	Amount of compaction within a time step (i.e., incremental compaction) in meters, using MoReS pressure, compressibility and reservoir thickness.	mean.C
x coordinate	X coordinate (in meters)	RD_X_center
Y coordinate	Y coordinate (in meters)	RD_Y_center

5.4. Meta-Parameter Choices

Meta-parameters are parameter settings of the machine learning models and include amongst others the exact target definition, feature selection thresholds, and integration choices. A selection of several of these parameters related to the forecast target has been discussed in the previous section. For many of the meta-parameters, there are potential ranges of values that could be used, such as different aggregation periods, for which the optimal value is not known up front. For that reason, a factorial approach has been set up to run each ML model with all possible combinations of meta-parameter ranges (Limbeck, et al., 2018).

The meta-parameters for the ML experiments are summarized with a short description for each parameter in Table 18. Most of these parameters are identical to those used in the temporal analysis (Limbeck et al., 2018), whereas the following additional meta-parameters are added for the spatial model extension:

- Bandwidth (in meters) for the kernel smoothing of spatial features;
- Resolution (in meters) of the grid cells used for generating x and y coordinate features;
- A number of spatial blocks to be generated for the x, y coordinates using k-means clustering (i.e., the number of blocks the field is divided into). These blocks are used for spatial cross-validation.

Even though the investigation of extensive ranges of meta-parameters is made possible by means of the factorial set-up of the ML experiments, the use of a wide range of parameters in this factorial approach rapidly results in hundreds of experiments that need to be run and evaluated. We optimize this process by using results from the temporal meta-parameter analysis (Limbeck et al., 2018) to guide down-sampling of the parameter space in the spatiotemporal analysis. Specifically, the following parameters are assigned single values or limited subsets of values instead of full ranges, based on earlier findings:

- No time-shifts are applied, as time-shifts for dynamic variables in the temporal analysis had no material effect on the estimated performance of the models, and time shifts are also not regarded in the PSHRA model. It could be reasonably argued perhaps, that time shifts might be present at specific locations in the reservoir. However, previous results suggest that this is not the case at the field aggregated level nor for selected regions.
- The M_{min} threshold, aggregation and time intervals have been studied extensively in other works (see discussion in section 4.3), so rather then again exploring all possible combinations of M_{min} thresholds for different periods, we focus on M_{min} of 1.5 from 1995-2016.
- Target quantity is the earthquake rate, to allow comparison with the previous temporal models and the PSHRA model, which have the same target.
- No aftershock processing is applied.
- A feature correlation threshold of 0.8 for feature down-selection is applied, to be consistent with the temporal models.

- No feature transformations such as principle component analysis are applied.
- Maximum number of lags is 2 (i.e., models investigate 0, 1 and 2 lag periods).
- Aggregation periods of 3, 6 and 12 months are used.
- Two values for the end of the time periods for model training are used: events up to and including December 2016 and events up to and including December 2012. Both end intervals provide separate holdout validation sets for 2017 and 2018 that are used to assess model performance. The 1995 – 2016 interval contains 268 historical events with a magnitude of 1.5 or higher, and the 1995 – 2012 interval contains 192 events.

Table 18 Meta-parameters for the experiments, with a brief description, the value range considered in the spatiotemporal analysis and the number of values that are used in the factorial approach for each parameter.

Meta parameter	Description	Value range	# Val.
ML Model (excl. baselines)	Type of machine learning model	RF, KNN, KSVM, GLMnet, GLM with top 5 significant variables	5
Target quantity	Forecast target, e.g., earthquake rate, count.	EQ rate	1
Gridsize	Resolution of the grid cells for spatial gridding to generate x, y, feature values.	1500 m	1
Time delay Production	Delay (number of time steps) in production data versus target quantity.	0	1
Min Magnitude	Lower bound for earthquake magnitudes to be used.	1.5	1
Time delay Pressure	Delay (number of time steps) in pressure data.	0	1
Number of spatial blocks	Number of blocks that the grid cells in the field are divided in using k-means clustering	5, 10	2
Time delay Compaction	Delay (number of time steps) in compaction data.	0	1
Kernel Smoothing bandwidth	Bandwidth in meters of the kernel smoothing applied to spatial predictor features.	2000, 2500, 3000, 3500, 4000	5
Max. nr. Lags	The maximum number of lags to be added to the time-series data.	2	1
Feature correlation threshold	The threshold above which features are defined as highly correlated. These features are then grouped, and one representative feature is used.	0.8	1

Feature transformations	Transformations applied to the features before use in modelling, e.g., principle component analysis.	None	1
Interval Length	Length of the period over which features are temporally aggregated.	3, 6, 12 months	3
Feature significance threshold	The minimum threshold for features to be significant.	0.4	1
Interval Start	Start of the time interval for model training	1995-01-01	1
Interval End	End of the time interval for model training	2012-12-31, 2016-12-31	2
Validation Strategy	Method for model validation	Spacetime walk-forward	1
Total			**300**

5.5. Experiment (down)selection

Based on the results from the event rate report, a subset of five models was identified that consistently showed better performance from the larger selection of machine learning models initially analysed. The current study is therefore limited to these five models: Random Forests (RF), more specifically the Ranger implementation, due its faster computation time while still remaining functionally equivalent (Wright and Ziegler, 2015), KSVM, KNN, GLMnet and GLM top, the latter being a GLM model (Hastie et al., 2009) that takes only the top five most significant features (see Appendix H for a short description of the ML models). In addition, several baselines are used for model comparison.

These baselines produce forecasts of the earthquake rate $p_{i,z}$ between time t_{i-1} *and* t_i for location bin z as follows:

1. **Depletion thickness based moving average (DepletionThickness MA)**:
   This baseline assumes that the activity rate at a given location is proportional to change in depletion thickness. Concretely, writing $D_{i,z}$ for depletion thickness, i.e. pressure times reservoir thickness, at time t_i and location z, obtained from the reservoir simulation model forecasts, the rate forecast is given as

$$p_{i,z} = \frac{D_{i,z} - D_{i-1,z}}{D_{i-1,z} - D_{i-a,z}} \sum_{j=i-a}^{i-1} \lambda_{j,z}$$

   where $\lambda_{j,z}$ is the observed earthquake count between time t_j and t_{j-1} in the spatial bin z, and a is the lookback parameter for the moving average.

2. **Strain thickness (compaction) based moving average (StrainThickness MA):**
   This baseline assumes that the activity rate at a given location is proportional to change in depletion thickness. Concretely, writing $S_{i,z}$ for strain thickness, i.e. vertical strain times

reservoir thickness, at time t_i and location z, obtained from the reservoir simulation model forecasts described in section, the rate forecast is given as

$$p_{i,z} = \frac{S_{i,z} - S_{i-1,z}}{S_{i-1,z} - S_{i-a,z}} \sum_{j=i-a}^{i-1} \lambda_{j,z}$$

where $\lambda_{j,z}$ is the observed earthquake count between time t_j and t_{j-1} in the spatial bin z, and a is the lookback parameter for the moving average.

The relative performance of the machine learning models and the baselines in terms of MAE, RMSLE and Mean Poisson Loss over the explored experimental space indicates that the Strain thickness MA baseline is performing better than the depletion thickness baseline (Figure 25 and Figure 26). Therefore, this will be the baseline against with the performance of the ML models is compared to. A detailed explanation of the error metrics and their application to model evaluation can be found in section 6 of this report.

Among the ML models we see comparable performance amongst most of the models with only the GLMnet model relatively underperforming in terms of MAE, RMSLE and Mean Poisson Loss with respect to the other ML models. Based on these error metrics, the RF, KSVM are the models that we will focus on in the results section.

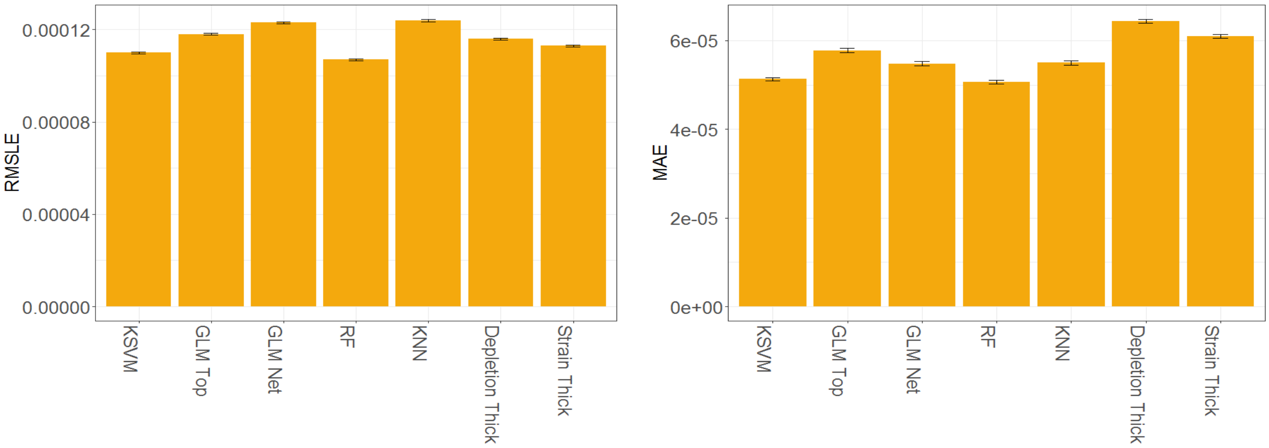


Figure 25: The RMSLE (left) and MAE (right) error metrics on a cell by cell basis per model. The errors are derived by comparing predictions for each cell with the actual earthquake rates in 3 months. The black bars denote standard error calculated with the Jackknife resampling technique as described in section 6.4.

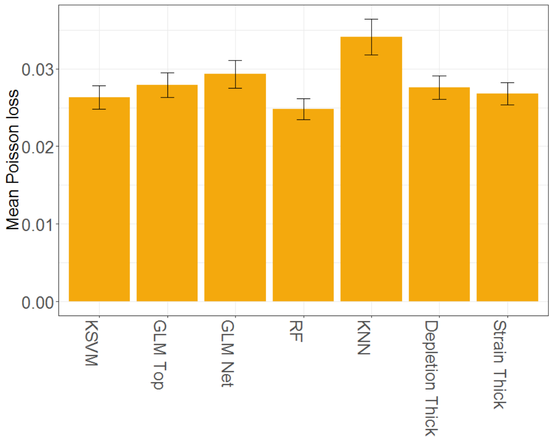


Figure 26: Comparative model performance in terms of the Mean Poisson loss aggregated over time. The figure is based on 3-month aggregation periods where the number of predicted earthquakes is compared with the actual earthquake count. The standard error denoted with the black bars is calculated with the Jackknife resampling technique as described in section 6.

Meta-parameter (down)selection

Based on the findings from the temporal analysis (Limbeck et al., 2018), a set of meta-parameters is used in the experiments. The only remaining parameters for which multiple values are being investigated in the experiments are the minimum magnitude limit of events (1.2 vs. 1.5), the bandwidth for kernel smoothing (2000 – 4000 m in steps of 500 m, Appendix J) and the aggregation period (3, 6 or 12 months).

For the minimum magnitude threshold, the MAE (mean average error) for experiments with a M_{min}= 1.2 is 0.0277 (se=0.005) whereas the MAE for experiments with M_{min}= 1.5 is 0.0173 (se=0.003). Since the MAE values were not normally distributed we proceed to test with a non-parametric test (Figure 27). According to the Wilcoxon signed-rank test, the errors are significantly different between experiments where the minimum magnitude was 1.2 and groups where the minimum magnitude was 1.5 (p-value < 2.2e-16). Thus, we chose to further only analyse experiments with a minimum magnitude of 1.5.

There is some indication that higher time-aggregations can lead to better performance when used on the reservoir grid where each cell has a dimension of 1.5km. We, therefore, use the shortest aggregation period (3 months) to obtain the highest temporal resolution, which is aligned with the aggregation period used in Limbeck et al. (2018).

The experiments with different bandwidths show that the RMSLE for experiments with a kernel smoothing bandwidth of 1500 m is significantly larger than the other two bandwidths and that there is a smaller difference in RMSLE between smoothing bandwidths of 3500 m and 5000 m. This is expected – if the smoothing bandwidth is large, the performance of all models improves but the models are less capable of making spatially accurate forecasts. The relatively large error in models with a smoothing bandwidth of 1500 m confirms that there is insufficient accuracy of the data at the reservoir level to make forecasts at this resolution. Models with a smoothing bandwidth of 5000 m have a slightly lower RMSLE on average compared to the 3500 m resolution but using models with a smoothing bandwidth of 5000 m decreases the forecasting power of these models, in terms of spatial resolution. Therefore, a smoothing bandwidth of 3500 m is used in the main experiment. This is also consistent with previous research (Bourne and Oates, 2015).

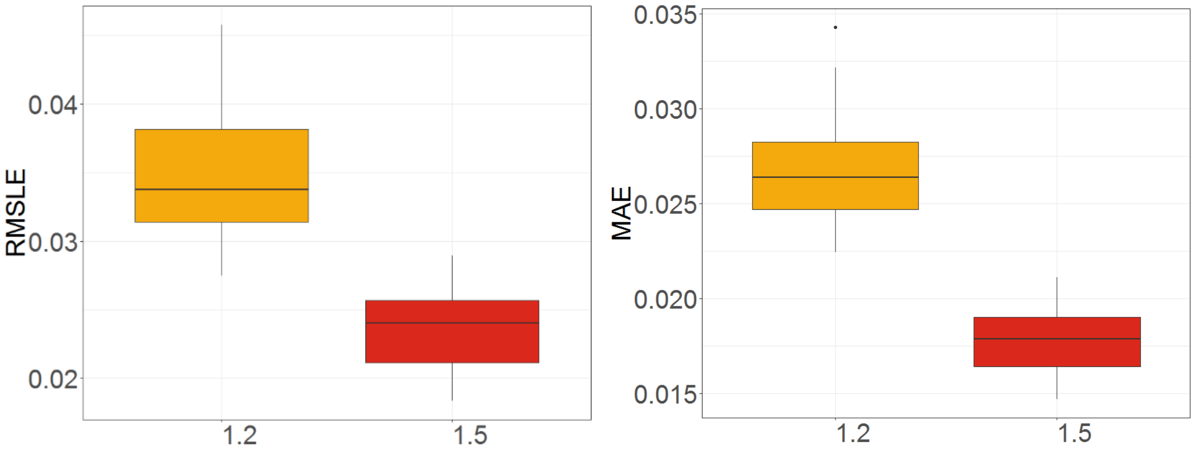


Figure 27: The boxplot illustrates the temporally aggregated median model performance in terms of RMSLE (left) and MAE (right) for magnitude 1.2 (red) and 1.5 (blue). Figures are based on 3-month aggregation periods, with a bandwidth of 3500. The boxplot depicts the performance metrics for all models.

6. Model Performance Evaluation

This section will expand on the resampling strategy used in the study, namely: a spatiotemporal walk-forward approach and the different metrics and measures used to test and validate the performance of the models. This section also covers the model comparison setup for validation and assessment of whether any of the machine learning models outperform the selected baseline(s).

6.1. Temporal: walk forward testing

When testing the forecasting power of a model, it is essential that the performance of the model is tested on data that has not been used to train the model. Due to overfitting to the given training data, model performance is in general better in-sample than for a hold-out set. For this reason, the data set needs to be partitioned into training and test sets, where the model performance is only estimated on the test sets. If a given data set has sufficiently many data points (the exact number will depend on several factors like the complexity of the problem, the complexity of the model and the signal to noise ratio in the data), stable estimates of model performance can be obtained using only few training and test splits.

In the case of a sufficiently large data set, the estimated model performance is insensitive to the chosen partition. For smaller data set sizes and complex problems, like the case considered in this study, the estimated model performance is more uncertain and depends to a certain extent on the chosen partition. To minimize the effect of the chosen partition on the computed error metric a common approach is to repeat the modelling experiments on many training and test partitions of the data to obtain more stable error estimates and to be able to bound the uncertainty introduced through the different partitioning schemes. If the forecasting target would be independently and identically distributed (i.i.d.) at different moments in time several resampling schemes like k-fold cross-validation or several different non-blocked flavours of the bootstrap, would be available, see chapter 7 in Hastie et al. (2009) for further details.

However, since we are dealing with time series data for which the i.i.d assumption, in general, does not hold; those techniques bear the risk of overestimating the forecasting performance of the models by leaking future information. Additionally, violation of the i.i.d. assumption can result in too small estimates of standard errors, which in turn could lead to Type 1 errors in hypothesis testing when testing two models for equivalence. The severity of the issues grows with increasing violation of i.i.d.-ness.

Therefore, we use a technique called Walk-Forward Testing, (see p. 548 in Kirkpatrick et al., 2013), which is commonly used for back testing algorithms when dealing with time series data as it arises, for instance, in the financial industries. Back testing of models is also frequent in other disciplines that are concerned with forecasts like meteorology and climatology but are referred to as hindcasting. Models are conditioned to historical data available at an initial moment in time where data of sufficient quality is available, then a forecast is created over a specific time interval, after which the model is reconditioned, and the procedure repeated. The quality of the model is assessed over the forecasting periods that have not been used to train the model. Depending on the actual application and the availability of the data the forecasting periods after which the models are updated can differ from hours (meteorology) to years or even decades (climatology). The methodology of walk forward evaluation honours the time series nature of the data, such that no data in the test set is younger than any data point in the training set. Note that this would not be true for ordinary k-fold cross-validation or conventional bootstrap resampling schemes.

Let $n > 0$ be the number of data points in our data set. We denote the data point at time instance i and location z by $d_{iz} = (x_{iz}, \lambda_{iz})$, where $x_{iz} \in \mathbb{R}^m$ are the m-covariate values that are available at time instance i and location z and λ_{iz} is the forecasting target at time instance i and location z (e.g., the earthquake rate or count). Without loss of generality, we assume that all categorical variables have been appropriately encoded as real numbers. Let $k \geq 0$ be the minimal number of samples that are required to train the (machine learning) model. Furthermore, let $1 \leq l \leq n$ be the forecast step size, i.e., the number of forecasts generated before the model is updated. Then, the Walk-Forward Testing approach works as follows (in pseudo code) for a given model f:

```
1. Let $i = k$
2. Let c = number of cells in the grid
3. Train model $f$ on $(d_1,\ldots,d_i)$, such that $f(x_{jz}) \approx \lambda_j$ for $1 \leq$ j $\leq$i for all c
4. Let $p_{i+1}=f(x_{j+1})$, …, $p_{i+1+l} = f(x_{j+1+l})$ be the forecast of the model trained in step
   2 for every c.
5. While $i \leq$ n−1− $l$ let $i = i + l$ for all c and go to 2.
```

An illustration of the approach is contained in the figure below. Each d_k block represents the data for all grid cells c available at time i.

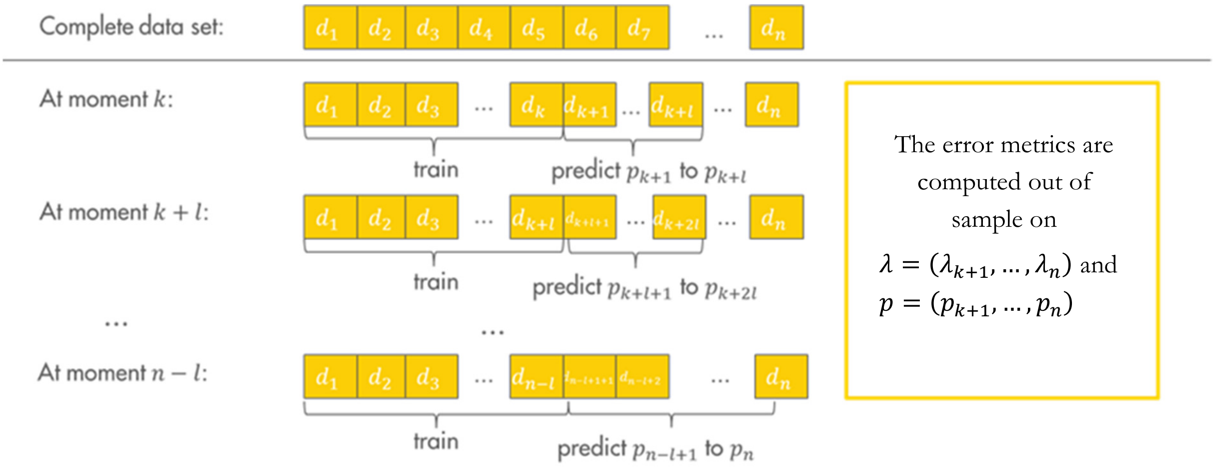

Complete data set:
d_1 d_2 d_3 d_4 d_5 d_6 d_7 … d_n
At moment k:
d_1 d_2 d_3 … d_k d_{k+1} … d_{k+l} … d_n
train
predict p_{k+1} to p_{k+l}
At moment $k + l$:
d_1 d_2 d_3 … d_{k+l} d_{k+l+1} … d_{k+2l} … d_n
train
predict p_{k+l+1} to p_{k+2l}
…
…
At moment $n - l$:
d_1 d_2 d_3 … d_{n-l} d_{n-l+1} d_{n-l+2} … d_n
train
predict p_{n-l+1} to p_n
The error metrics are computed out of sample on
$\lambda = (\lambda_{k+1}, \ldots, \lambda_n)$ and
$p = (p_{k+1}, \ldots, p_n)$

Figure 28: Diagram showing the step-ahead resampling approach. Each d_i block represents the data for all grid cells available at time i.

After the walk forward algorithm has terminated, we have a vector of forecasts for time instances $k + c$ to n, namely p_{k+c} to p_n. Those, paired with the true values λ_{k+c} λ_n can now be used to evaluate the performance of the algorithm f using one of the error metrics. We note that this approach implicitly assumes that models that perform better than others on short term (1 to 3 months) forecasts also do so on longer term forecasts. In theory long term (1 to 5 years) forecasts could be used for relative model performance evaluation, however, as uncertainty increases with time, longer-term forecasts are in general more difficult to differentiate from each other. Using short term relative forecast differences increases the differentiative power.

6.2. Model comparison for hypothesis testing

Model comparison is a key part of this study as it is at this stage that we strive to statistically quantify whether the machine learning models generated can beat the baselines that have been put in place as well as potentially comparing competing machine learning models should that be the case. This should be done in a way which is both able to rank models in terms of their performance and also to identify if one model is significantly better than another.

Error metrics and standard error

In this study, we use several error metrics that allow us to evaluate model performance and compare it. The table below shows the three main error metrics which were also used in the event rate report. Later in this chapter, we will introduce an additional error metric $\mathrm{L}(\mathcal{M}|\boldsymbol{\lambda})$ representing the likelihood of model $\mathcal{M}$ given data $\boldsymbol{\lambda}$ for spacetime evaluation and comparison using a likelihood ratio approach. In the table below let k be the minimum number of training points.

Table 19: Standard Performance metrics

Err. Metric	Formula	Properties
MAE	$\frac{1}{n-k}\sum_{i=k+1}^{n}\lvert\lambda_{i,z}-p_{i,z}\rvert$	• The result is in the unit of the original data • Uniform weighting of differences, thus less sensitive to outliers
RMSLE	$\sqrt{\frac{1}{n-k}\sum_{i=k+1}^{n}\left(\log\left(\lambda_{i,z}+1\right)-\log\left(p_{i,z}+1\right)\right)^2}$	• If $\lambda_{i,z}$ is large, deviations from p_i have less weight than if $\lambda_{i,z}$ is small . • Commonly used for count data • Applicable to $p_{i,z} \geq -1$ and $\lambda_{i,z} \geq -1$
MPL (Mean Poisson Loss)	$\frac{1}{n-k}\sum_{i=k+1}^{n}\left(p_{i,z}-\log\left(p_{i,z}\right)\lambda_{i,z}+\log\left(\lambda_{i,z}!\right)\right)$	• Error metric specific to count data • Hard to compute in FP-arithmetic for large values of $\lambda_{i,z}$ if implemented naively • Special handling for the case $p_{i,z} = 0$ required

The formulas in Table 19, will lead to an error per location (cell) at a given time interval, we then take the individual errors per cell and aggregate them temporally to get one average value of the error at the given time interval. And further take the average over all time intervals to get one value for the average performance of the model.

Furthermore, only having computed the values of the individual error functions for each experiment is not sufficient to assess if one method is significantly better than an alternative method with respect to a certain confidence level. For this reason, we also need to estimate the standard deviation/standard error that is associated with each error measure. While explicit formulas are available for some error measures like the MAE, we chose a general approach, that can be applied to most computable error measure to assess the standard error (SE) that is associated with it. To estimate the standard deviation of an error measure m, we make use of a technique called Jackknife resampling. The Jackknife estimate of variance is consistent for sample means and correlation coefficients, which covers the error metrics mentioned in Table 19, which are used in this study. Further details and references are contained for instance in (Efron and Stein, 1981).

Let $\lambda \in \mathbb{R}^n$ and $p \in \mathbb{R}^n$ again denote the true and forecasted values and let m be an error measure that takes the observed and the predicted EQ rate as inputs. Here m refers to the formulas introduced in Table 19. Let us further denote by $\lambda_{-i} \in \mathbb{R}^{n-1}$ and $p_{-i} \in \mathbb{R}^{n-1}$, respectively λ and p from which the i-th entry has been removed. I.e. $\lambda_{-i} = (\lambda_1, \lambda_2, \dots, \lambda_{i-1}, \lambda_{i+1} \dots, \lambda_n)$ and p_{-i} $(p_1, p_2, \dots, p_{i-1}, p_{i+1} \dots, p_n)$. Furthermore, λ and p here represent vectors where the data has been temporally aggregated such that the spatial data is aggregated over each time point or alternatively they represent the vector of all space-time observations. And where k is the minimum numbers of points after which the forecasts start. Then the standard error that is associated with m (error measure) for a given instance of λ and p is defined as SE_m which the standard error of error measure $m(\lambda, p)$.

$$SE_m = \sqrt{\frac{n-1-k}{n-k} \sum_{i=k+1}^{n} \left(m(\lambda_{-i}, p_{-i}) - \sum_{j=k+1}^{n} \frac{m(\lambda_{-j}, p_{-j})}{n-k} \right)^2}.$$

In the presence of significant auto-correlation in the series $m(\lambda, p)$ correlation correction needs to be applied to avoid underestimating uncertainties. Assuming stationarity, the adjusted formula to estimate the standard error is then given by

$$SE_m = \sqrt{\frac{1+\rho}{1-\rho} \frac{n-1-k}{n-k} \sum_{i=k+1}^{n} \left(m(\lambda_{-i}, p_{-i}) - \sum_{j=k+1}^{n} \frac{m(\lambda_{-j}, p_{-j})}{n-k} \right)^2},$$

where ρ is an estimate of the auto-correlation coefficient obtained for instance via the Praise Winsten estimation procedure, see Bence (1995). For mean-based error measures, this corresponds to the usual correlation adjustment of the sample error.

The use of the jackknife resampling to estimate the SE of the previously presented error metrics is first done at across space and time. The sample size is, therefore, s = b x d observations, where b is the number of time points (intervals) and d is the number of XY locations However we recognize that this can be problematic since locations are likely not independent from each other. Hence we estimate the SE in the case where the data is first temporally aggregated such that the value of individual cells are aggregated over the same time interval in which case the sample size is b, this is then directly comparable to what was done in the previous report (Limbeck et al., 2018). We note that we use and report the SE as calculated in the case where the data has been first temporally aggregated which avoids the issue of location points not being independent from each other.

Likelihood Ratio estimation

As an extension to the MAE, RMSLE and Mean Poisson Loss metrics we propose a method which takes inspiration from the R-test as outlined in Schorlemmer et al. (2007). The basis of this test is the widely used likelihood ratio which is defined as,

$$\mathrm{R}(\mathcal{M}_l, \mathcal{M}_0|\lambda) = \frac{\mathrm{L}(\mathcal{M}_i|\lambda)}{\mathrm{L}(\mathcal{M}_0|\lambda)}.$$

Where $\mathrm{L}(\mathcal{M}|\lambda)$ is the predictive likelihood of a given model $\mathcal{M}$ given the target value $\boldsymbol{\lambda} \in \mathbb{R}^n$ and where $\mathcal{M}$ is a model that makes one step ahead forecasts of the earthquake rate as described in section 6.1. Here $\mathrm{L}(\mathcal{M}|\lambda)$ can be seen as an additional performance metric together with the MAE, RMSLE and Mean Poisson Loss. In our case, we choose to use the Poisson likelihood with rate parameter given by the prediction of model $\mathcal{M}$. To ensure models are evaluated on their forecasting performance, λ contains the recorded count data from a hold-out set as outlined previously in the walk-forward forecast scheme. A value of greater than 1 indicates that the proposed model, $\mathcal{M}_l$ is performing better than the baseline model, $\mathcal{M}_0$. This can equivalently be written as a difference of log likelihoods,

$$r(\mathcal{M}_l, \mathcal{M}_0|\lambda) \ = \ell(\mathcal{M}_l|\lambda) - \ \ell(\mathcal{M}_0|\lambda).$$

For our application the likelihood function we use is the Poisson likelihood such that,

$$\ell(\mathcal{M}_l|\lambda) = \sum_{i,z} \lambda_{i,z}\, log\left(p_{i,z}^{(l)}\right) - p_{i,z}^{(l)} - log\left(\lambda_{i,z}!\right).$$

Where $p_{i,z}^{(l)} \in \mathbb{R}^n$ denotes the prediction of $\lambda_{i,z}$ given by model $\mathcal{M}_l$.

In classical statistics, this ratio is often assumed to have a chi-squared distribution however this assumes the models are nested and the number of degrees of freedom is known. These assumptions are not applicable to many of the models we are considering. Schorlemmer et al. (2007) judge the significance of the ratio by simulating earthquake catalogues from both models being considered and recalculating the ratio on this simulated data. This process is repeated many times and significance is judged by calculating the proportion of the simulated rations which fall below 1 or zero in the case of the log ratio. There has been some criticism of this method (see Gerstenberger et al., 2009; Rhoades et al., 2011; Bray and Schoenberg, 2013). Part of this criticism centres on the inconsistency of using two different simulation models which can mean a different model is favoured if the proposed model and baseline are switched. A further criticism is targeted at the underlying assumption that the proposed model is a sensible (essentially a good model) which is not necessarily the case.

Simulation Model

We aim to address the issues highlighted previously by introducing a simulation model $\mathcal{M}_s$. From this model, we repeatedly simulate new earthquake counts, $\tilde{\boldsymbol{\lambda}}$, and evaluate the log ratio, $r(\mathcal{M}_l, \mathcal{M}_0|\tilde{\boldsymbol{\lambda}})$, we then calculate the proportion of these simulations which are less than 0. Clearly the choice of $\mathcal{M}_s$ is important as this method will favour models which are most similar to $\mathcal{M}_s$. Ideally $\mathcal{M}_s$ should be the true earthquake generating model, since this is unknown we aim to choose a model which closely matches the observed events. We note that when estimating the simulation model, we can make use of all available data as we are not interested in evaluating the forecasting power of this model. For the same reason, we can allow the model to be unphysical as we will not be using it to make any predictions or extrapolations, we simply need it to accurately represent the natural variation in earthquake counts. The simulation model we choose is a Poisson process such that the simulated count at time point i and location z = (x,y), follows a Poisson distribution with rate parameter $\theta_{i,z}$. This rate is estimated by fitting a GAM of the form,

$$\log(\theta_{i,z}) \ = \ s_1(i) \ + \ s_2(z).$$

Where $s_1(i)$ and $s_2(z)$ are smoothly varying spline functions. We chose this model as it is different from any of the proposed machine learning models, and so should not unfairly favour any model class. The spline function is also flexible enough to closely match the spatial and temporal variations in the observed rate.

This model is fitted using the `gam` function from the `mgcv` package in `R` which uses the following algorithm We first define the general log likelihood of the model as,

$$\ell(\theta_{i,z}, g_1, g_2, \dots, g_M) = \log[f(X|\theta_{i,z}, g_1, g_2, \dots, g_M)],$$

where θ_{iz} are the rate parameters and $\{g_m\}$, is a set of smooth functions, in our case the spatial and temporal spline functions. These functions can in tern be defined by a set of basis functions, $b_{ij}(x)$ and coefficients, β_{ij} such that,

$$g_i(x) = \sum_j \beta_{ij} b_{ij}(x).$$

The coefficients are estimated by maximising a penalised likelihood function,

$$\hat{\beta} = arg\max_{\beta}\left\{l(\beta) - \frac{1}{2}\sum_{m}^{M} \omega_m \beta^T S^m \beta\right\}.$$

Here S^m is a matrix of fixed coefficients and $\{\omega_m\}$ is a set of smoothing parameters which control the smoothness of the spline function. These parameters are themselves estimated using the marginal log likelihood,

$$\nu(\omega) = \log \int f(y|\beta) f_\omega(\beta) d\beta.$$

In practice, this is replaced with a Laplace approximation. Full details of the fitting scheme and smoothness estimation can be found in (Wood et al., 2016).

Figure 29 and Figure 30 show the spatial and temporal forms of this model using 1000 simulated catalogues for which the mean count and 95% quantiles are calculated. These are aggregated over either time or space. Looking at these plots, we see visually that the model appears to give a close match to the spatial and temporal pattern of the earthquake occurrences. In the spatial plot, most of the real counts fall within the expected quantiles, although there are a few points higher than expected in the temporal plots, most notably February 2013. These points do appear to be unexpected outliers as the simulated mean appears to match the changes in the mean of the true counts.

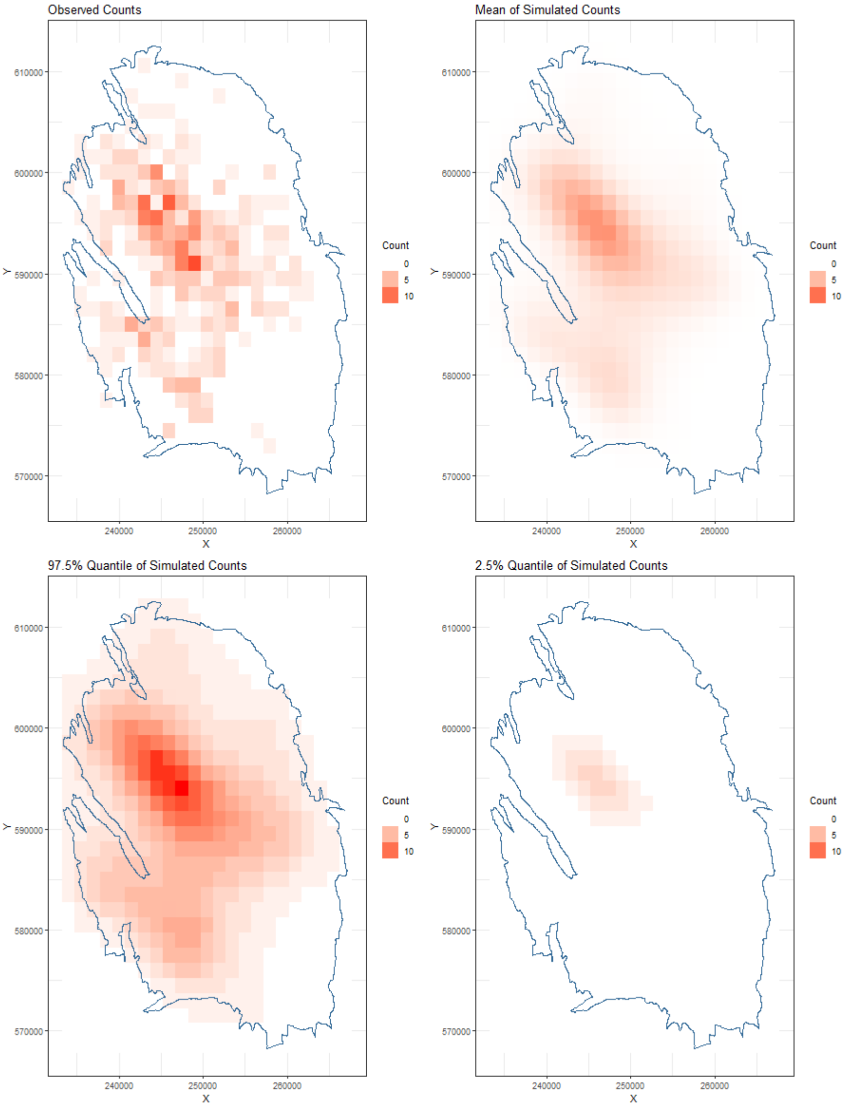


Figure 29: Spatial plots of earthquake counts aggregated through time. Moving clockwise from the top left, the plots show the recorded earthquake counts for M≥1.2, the mean count of 1000 simulations from $\mathcal{M}_S$, the 97.5% quantile of the simulated counts and the 2.5% quantile of the simulated counts. The gradient from white to red depicts the range of 0 – 10.

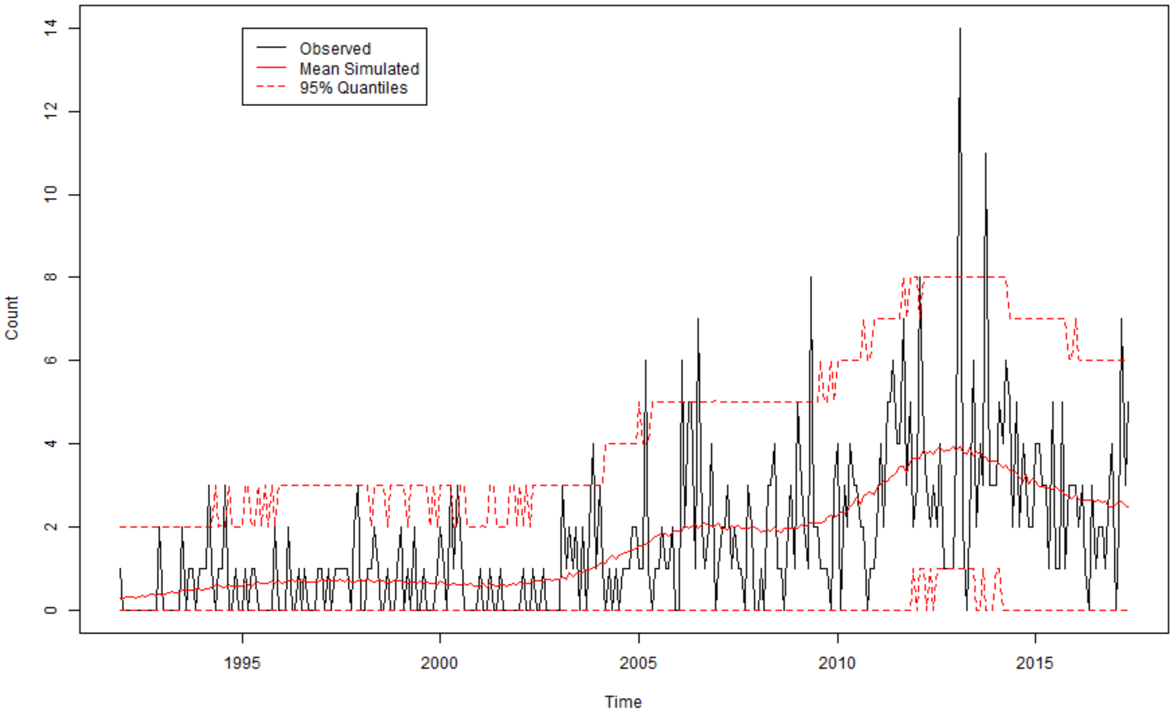


Figure 30: Time series of Earthquake Counts aggregated across the Groningen gas field. The black line shows the recorded counts; the solid red line shows the mean count of 1000 simulations from $\mathcal{M}_s$, the dashed red lines show the 97.5% and 2.5% quantiles of the simulated counts.

Using Likelihood Ratio and simulation model for model evaluation

In general, we would need to perform a simulation to calculate the percentage of log ratios which are above zero. In our case, the likelihood function we choose is the Poisson likelihood, as this is able to deal effectively with count data especially count data with a mean close to zero. We also draw our simulations from a Poisson distribution. Given these choices, we can find an analytical formula for the expected value and variance of $r(\mathcal{M}_i, \mathcal{M}_0|\tilde{\boldsymbol{\lambda}})$ which allows us to bypass having to actually perform the simulations. These are given by,

$$\mathrm{E}[r(\mathcal{M}_l, \mathcal{M}_0|\tilde{\lambda})] \;=\; \sum_{iz}\left[\theta_{i,z}\left\{log\left(p_{i,z}^{(l)} - p_{i,z}^{(0)}\right)\right\} - p_{i,z}^{(l)} + p_{i,z}^{(0)}\right].$$

$$\mathrm{var}[r(\mathcal{M}_l, \mathcal{M}_0|\tilde{\lambda})] = \sum_{i,z}\left[\theta_{i,z}\left\{log\left(p_{i,z}^{(l)} - p_{i,z}^{(0)}\right)\right\}^2\right].$$

Where θ_{iz} is the rate parameter used for the simulation model and $p_{iz}^{(l)}$ and $p_{iz}^{(0)}$ are the rates predicted by models $\mathcal{M}_l$ and $\mathcal{M}_0$. A derivation of these results can be found in Appendix I. If we then assume that, due to the central limit theorem, $r(\mathcal{M}_l, \mathcal{M}_0|\tilde{\lambda})$ is Normally distributed we can directly calculate the proportion of simulated log rations which are below zero as,

$$\mathrm{p} = 1 - \Phi\left(\frac{\mathrm{E}[r(\mathcal{M}_l, \mathcal{M}_0|\tilde{\lambda})]}{\sqrt{\mathrm{var}[r(\mathcal{M}_l, \mathcal{M}_0|\tilde{\lambda})]}}\right).$$

Where $\Phi(.)$ is the standard Normal CDF. We feel this assumption is valid as we are dealing with a large number of grid points and so the sum in the expression for $r(\mathcal{M}_l, \mathcal{M}_0|\tilde{\lambda})$ will contain a large number of terms. We then interpret this value as the p-value for the hypothesis test with the following null and alternative hypotheses,

- H$_0$: The model performance is identical, $\mathrm{E}[r(\mathcal{M}_l, \mathcal{M}_0|\boldsymbol{\lambda})] = 0$.
- H$_1$: Model $\mathcal{M}_i$ is performing better than the baseline, $\mathrm{E}[r(\mathcal{M}_l, \mathcal{M}_0|\boldsymbol{\lambda})] > 0$.

A value of < 0 would indicate that the model $\mathcal{M}_l$ is **not** performing better than the model $\mathcal{M}_0$ (the baseline).

6.3. Forecast uncertainty quantification

By default, our forecasts are one-time interval ahead. Consequently, the uncertainty estimates in the form of standard errors are by default only applicable for one-time interval ahead – i.e., not multiple time intervals ahead as required for long term seismicity forecasting. Given the non-parametric nature of most of our algorithms there is no analytical derivation from which we can obtain longer term uncertainty estimates, so we proceed to obtain such estimates using an empirical approach.

It is important to note that currently, we constrain the uncertainty quantification to be the aggregated uncertainty over all spatial points over a given time interval. This effectively means that currently, we do not estimate uncertainty at each spatial location but instead only at the level of the whole Groningen reservoir by aggregating the individual location cell values. A critical outcome of this is that the spatial variation that is present in the cells will be inherently encoded in the temporal aggregation at the reservoir level, making the confidence bands ostensibly larger than the ones presented in the temporal report. The difference with the previous approach being that we first estimate the errors at the cell level and then we aggregate these values in time rather than aggregating first and then estimating the error. We believe that in this way we are accounting for the spatial variability that is present in the forecasts while still being able to make use of the same code functionality as in Limbeck et al. (2018).

Let $h > 0$ be the number of time intervals points the historical data set, let $l \geq 1$ be the forecast step size (i.e., the number of forecasts that are generated before the model is retrained) and let k be the minimum number time interval points used for training the model. Furthermore, we denote the forecasts of a walk forward run with forecast step size l by $p_i^{(l)}$ where the forecast has been temporally aggregated over all locations at given time interval i for values of $i \in \{k+1, \ldots, h\}$. The associated forecasting errors are denoted by $\delta_i^{(l)} = m(\lambda_i, p_i^l)$ for a pointwise error measure m and λ_i the true value at time interval i aggregated over all locations at given time interval i for values of $i \in \{k+1, \ldots, h\}$. Since these estimates are all highly dependent on i, we are stabilizing the results by estimating the forecast uncertainty y time intervals ahead $\bar{\delta}^{(y)}$ as the 10th/90th percentiles of the set of all $\delta_i^{(l)}$ for which the time interval between (re)-training the model and the actual forecast is equal to b. To obtain the required $\delta_i^{(l)}$ the calculations are performed in a block-wise fashion for increasing forecast window sizes.

To quantify, e.g., the uncertainty $\delta_i^{(5)}$ of forecasting five steps ahead:

- We generate walk-forward runs calculating iteratively:
    - Five steps forward $\delta_i^{(5)}$ for values of $i \in \{6, \ldots, h\}$;

- Six steps forward $\delta_i^{(6)}$ for values of $i \in \{7, \dots, h\}$;
- Seven steps forward $\delta_i^{(7)}$ for values of $i \in \{8, \dots, h\}$;
- etc.

- From these runs we select:
  - $\delta_{k+5}^{(5)}, \delta_{k+10}^{(5)}, \delta_{k+15}^{(5)}, \dots$
  - $\delta_{k+5}^{(6)}, \delta_{k+11}^{(6)}, \delta_{k+17}^{(6)}, \dots$
  - $\delta_{k+5}^{(7)}, \delta_{k+12}^{(7)}, \delta_{k+19}^{(7)}, \dots$
  - etc.
- The 10th/90th percentiles of the elements listed above are used to obtain $\bar{\delta}^{(5)}$.

An illustrative example with $\bar{\delta}^{(3)}$ is contained in Figure 31.

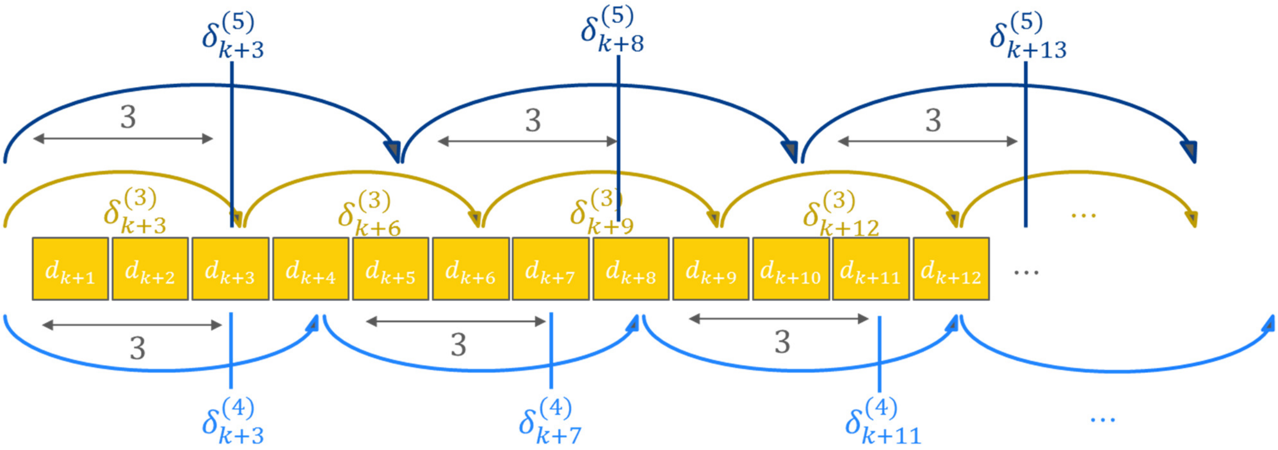


Figure 31: Illustrative example of how the uncertainty estimate for forecasting three (3) steps ahead $\bar{\delta}^{(3)}$ is derived from the 10th/90th percentiles of the set of all three step ahead forecasts.

Due to the variance of the estimation for the forecast errors, the empirical estimates can violate our theoretical assertion of a smooth, monotonically increasing error function with time. In view of that, we implemented an isotonic regression to ensure a monotonically increasing error. The lower confidence interval is bounded by 0 since we only consider non-negative forecasting targets. An alternative approach that could be considered in a future iteration would be to use the estimated standard errors instead of the bootstrapped percentiles which potentially exhibit high variance proportional to $1/(percentile\_density)^2$. Due to the large number of experiments that are necessary we also note that obtaining empirical uncertainty estimates as described above is computationally demanding. The computational demand is also one of the reasons why we provide these confidence bands for the temporally aggregated case rather than for the individual grid cells, however and as mentioned above we do attempt to encode the additional source of variability brought forward by space by estimating the error metric m on a cell by cell basis and then aggregating these values per time interval. We do note however and highlight this in the limitations of this study that a more refined method might be necessary to estimate confidence bands in this way given the new spacetime setup. Moreover, having a longer time aggregation could help by stabilizing the confidence band estimation and reducing spatially induced variability.

7. Evaluation of Machine Learning-based Spatiotemporal Seismicity Forecasts

Model performance in space and time is qualitatively and quantitatively analysed in this section, for a model with experiment parameters as defined in Table 20, comparing the machine learning models against the developed baselines using an alternative version of the R test (see section 6.2) and a simulation model. These models include the same extensions, i.e., convex hull analysis, ultimate state constraining and hold-out period training, as the temporal models described in section 2, and show approximately zero seismicity after the Groningen field has been shut in (Figure 32 and Table 21). The models are developed for four production forecasts which are based on different development scenarios:

- Reference Case model, from the Winningsplan 2016 (Nederlandse Aardolie Maatschappij, 2016c);
- The average production scenario announced by the Ministry of Economic Affairs and Climate in March 2018 (Ministry of Economic Affairs and Climate Policy, 2018);
- The production scenario for a warm winter, with lower than average production rates (Ministry of Economic Affairs and Climate Policy, 2018);
- The production scenario for a cold winter, with higher than average production rates (Ministry of Economic Affairs and Climate Policy, 2018).

Table 20: Experimental input parameters of selected experiments.

Parameter	**Value**
Seismicity data used for period:	01-01-1995 – 31-12-2016 01-01-1995 – 31-12-2012
Aggregation period length	3 months
Minimum magnitude	1.5
Bandwidth of spatial smoothing	3500 m
Time shifts	None
Ultimate states	Yes
Number of ultimate states repetitions	20

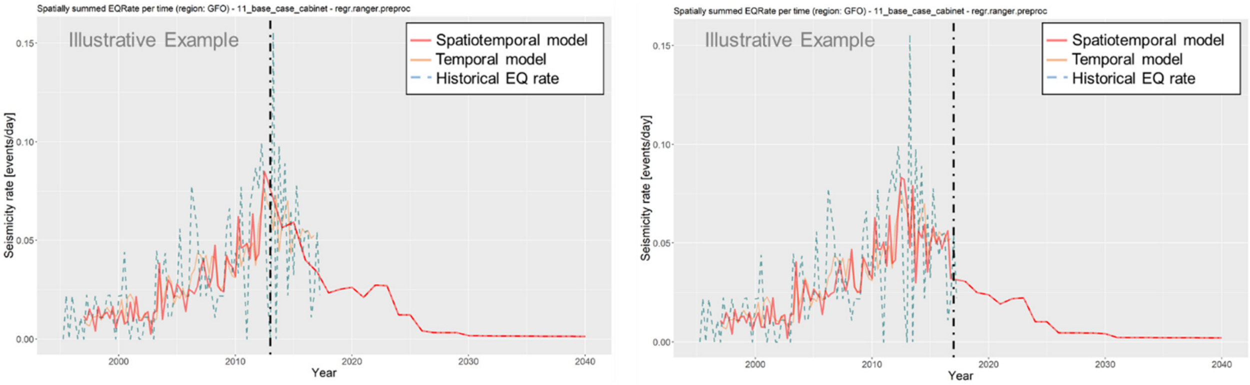


Figure 32: Illustrative examples of the impact of constraining spatiotemporal models using ultimate states for the RF models forecasting from 2013 (left) and 2017 (right) to 2040. Note that the seismicity rate approximately approaches zero towards 2040, though it is not exactly zero.

Table 21: MAE error metrics for the RF and KSVM experiments without (left) and with (right) constraining of the models to an ultimate post-shut-in state.

Without constraining		With constraining	
RF (1995-2012)	0.0147 (±0.0015)	**RF (1995-2012)**	0.0157 (±0.0018)
SVM (1995-2012)	0.0161 (±0.0019)	**SVM (1995-2012)**	0.0174 (±0.0019)

7.1. Quantitative Evaluation using Likelihood Ratio: Forecast Performance

The Log-Likelihood Ratio test as described in section 6.2 is used to compare the spatiotemporal performance of the ML models against the selected vertical strain thickness Moving Average (MA) baseline and against the simulation model described in section 6.2. The latter comparison is made for illustrative purposes to show which models best approach the simulation model although evidently, none of the models will beat the simulation model. However, their relative ranking against it might still provide useful information. The ultimate conditions mentioned in the previous section are implemented for all experiments discussed in the following sections.

Comparison of ML models versus baseline

The Random Forest and KSVM perform significantly better when space and time are considered than the selected baseline for the period 1995-2016 while only the Random Forest model performs significantly better than the vertical strain thickness MA on the 1995-2012 period. The other baseline, (Depletion Thickness MA) does not perform as well as the selected baseline which confirms that the comparison is made against the best of the two considered baselines. The "R VAL" indicates the value of the ratio, calculated using the observed earthquake counts, where higher positive numbers indicate better performance with respect to the baseline while negative values indicate that the performance is not as good as that of the baselines. The highest ratio values are for the Random Forest and KSVM

Table 22: Model test results for period 1995-2017, "R VAL" shows the ratio value calculated using the observed counts. The columns "E[R]", "VAR[R]" and "P-VALUE" are the expected value, variance, and p-value calculated using the formulas in Section 6.2. A higher positive R VAL in combination with a significant p-value represent models that perform better than the baseline model.

MODEL	R VAL	E[R]	VAR[R]	P-VALUE	Better then baseline?
RF	110.8	97.5	282.0	3.2E-09	Y
KSVM	32.1	38.4	254.0	8.1E-3	Y
KNN	-179. 7	-154.8	1489.1	1	N
GLM Net	-136.1	-126.1	377.0	1	N
GLM Top	-57.7	-57.7	141.4	1	N
Depletion Thickness MA	-44.1	-41.8	13.5	1	N

Table 23: Model test results for the time period 1995-2012, "R VAL" shows the ratio value A higher positive R VAL in combination with a significant p-value represent models that perform better than the baseline model

MODEL	R VAL	E[R]	VAR[R]	P-VALUE	Better then baseline?
RF	90.5	71.0	237.4	2.0E-06	Y
KSVM	16.7	10.8	212.9	2.3E-01	N
KNN	-152.1	-132.2	1208.2	1	N
GLM Net	-113.4	-103.3	311.3	1	N
GLM Top	-35.7	-40.9	108.7	1	N
Depletion Thickness MA	-32.1	-28.9	9.6	1	N

Comparison of ML models versus simulation model

As outlined in section 6, we have developed a single simulation model from which catalogues can be generated. This model uses all information to approximate the true earthquake generation model as closely as possible. The idea of using multiple catalogues is taken from the L-Test (CSEP, 2018) which is meant to account for the potential spatial uncertainty in the earthquake detection network as well as inherent uncertainty in the process itself.

Table 24 and Table 25 show the results of the likelihood ratio comparison between the ML models and the described simulation model. Based on the likelihood ratio values in this table, none of the models perform better than the simulation model. This is naturally expected since to be better than the simulation model one would need to be able to (almost) perfectly forecast seismicity. However, one valuable take away is that the relative performance ranking of the ML models is consistent with previous results in the sense that the Random Forest is the best performing followed by the KSVM.

Table 24: Model test results for the time period 1995-2016 for comparison with the simulation model. The "R VAL" shows the ratio value calculated using the observed counts, a negative value indicates that the model does not perform better than the simulation model, a value of 0 would mean that both perform equally well. "E[R]" is the expected value of the ratio and "VAR[R]" is the variance in the ratio. These are calculated using the formulas from Section 6.2.

MODEL	R VAL	E[R]	VAR[R]	P-VALUE
RF	-74.9	-91.3	231.1	1
KSVM	-153.6	-150.5	430.4	1
KNN	-365.4	-343.6	1718.3	1
GLM Net	-321. 9	-315.0	1118.8	1
GLM Top	-243.5	-246.5	731.3	1
Depletion Thickness MA	-229.8	-230.6	568.6	1
Strain Thickness MA	-185.8	-188.8	462.5	1

Table 25: Model test results for the time period 1995-2012 for comparison with the simulation model. The "R VAL" shows the ratio value calculated using the observed counts; a negative value indicates that the model does not perform better than the simulation model, a value of 0 would mean that both perform equally well. "E[R]" is the expected value of the ratio and "VAR[R]" is the variance in the ratio. These are calculated using the formulas from Section 6.2.

MODEL	R VAL	E[R]	VAR[R]	P-VALUE
RF	-54.9	-62.4	160.7	1
KSVM	-128.7	-122.5	366.5	1
KNN	-297.5	-265.5	1397.1	1
GLM Net	-258.8	-236.6	868.9	1
GLM Top	-181.1	-174.3	528.6	1
Depletion Start	-177.5	-162.2	429.3	1
Depletion MA	-145.4	-133.4	350.8	1

Performance on the validation period 2013-2017

We have also evaluated the period between 2013 and 2017 using the likelihood ratio test introduced in the previous section by comparing the performance of the ML models against the vertical strain thickness MA baseline. The Random Forest and KSVM beat this baseline over this validation period, but we do note that the ratio value is much lower than the ratio value for the periods 1995-2017 (Table 22) or 1995-2012 (Table 23). This perhaps highlights that this particular time interval (2013-2017) is one where the ML models forecast accuracy is penalized by the change in trend as it has already been hypothesized in the event rate report (Limbeck et al., 2018).

Table 26: Model test results for the validation period 2013-2017 for comparison with the Strain Thickness MA baseline. The "R VAL" shows the ratio value calculated using the observed counts; a positive value indicates that the model performs better than the baseline, a value of 0 would mean that both perform equally well. "E[R]" is the expected value of the ratio and "VAR[R]" is the variance in the ratio. These are calculated using the formulas from Section 6.2.

MODEL	R VAL	E[R]	VAR[R]	P-VALUE	Better then baseline?
RF	19.0	27.7	43.0	1.18E-05	Y
KSVM	15.3	27.4	41.7	1.07E-05	Y
KNN	-27.6	-22.6	280.9	0.911	N
GLM Net	-22.7	-22.9	65.7	0.998	N
GLM Top	-15.9	-16.9	30.4	0.999	N
Depletion Thickness MA	-12.0	-12.9	3.9	1	N

Error Metric Comparison

As can be seen from Table 27, the RF and KSVM models achieve the lowest errors across all error metrics. However, the RF model does not have a significantly lower mean Poisson loss error compared with the Strain Thickness model.

It is worthwhile to note that the scale of the error metrics here differs from what has been previously shown, in that the values in the table below represent the metrics estimated on a cell-by-cell basis and then averaged. The values in Table 9 in chapter 2 of this report, for example, show the error metrics calculated over the whole study region (temporally aggregated), and hence the value of the MAE is nominally larger and is therefore on a different scale.

In order to make a direct comparison between Table 27 and Table 9 one would need to either divide the values of Table 9 by the number of grid cells (556 in our case with a 1500 m resolution grid) or multiply the values in Table 27 below by the number of grid cells (556) to bring the MAE to the value of the whole reservoir.

Table 27: Mean Average Error (MAE), Root Mean Squared Logarithmic Error (RMSLE) and Mean Poisson loss error metrics on a cell by cell basis for each model together with the respective standard deviations for the period 1995-2017. The standard errors are calculated based on the Jackknife resampling method, as described in section 6.

	MAE	MAE SE	RMSLE	RMSLE SE	Mean Poisson loss	Mean Poisson loss SE
RF	5.1E-05	4.0E-07	10.7E-05	4.0E-07	2.48E-02	1.4E-03
KSVM	5.1E-05	4.1E-07	11.0E-05	4.1E-07	2.63E-02	1.5E-03
KNN	5.5E-05	4.7E-07	12.4E-05	4.7E-07	3.41E-02	2.3E-03
GLM Net	5.5E-05	4.7E-07	12.3E-05	4.77E-07	2.93E-02	1.8E-03
GLM Top	5.8E-05	4.4E-07	11.8E-05	4.4E-07	2.79E-02	1.6E-03
Depletion Thick	6.4E-05	4.1E-07	11.6E-05	4.1E-07	2.76E-02	1.5E-03
Strain Thick	6.1E-05	4.1E-07	11.3E-05	4.1E-07	2.68E-02	1.4E-03

7.2. Evaluation of Event Rate Forecasts

Aggregated temporal performance training and testing period

Spatial aggregation of the spatiotemporal model forecasts yields an earthquake rate that is similar to the temporal-only model on the historical step-forward training and testing period, as illustrated for the post-March 2018 Average production scenario for the KSVM model (Figure 33). The spatiotemporal model shows a declining trend in earthquake rate for the future period up to 2025.

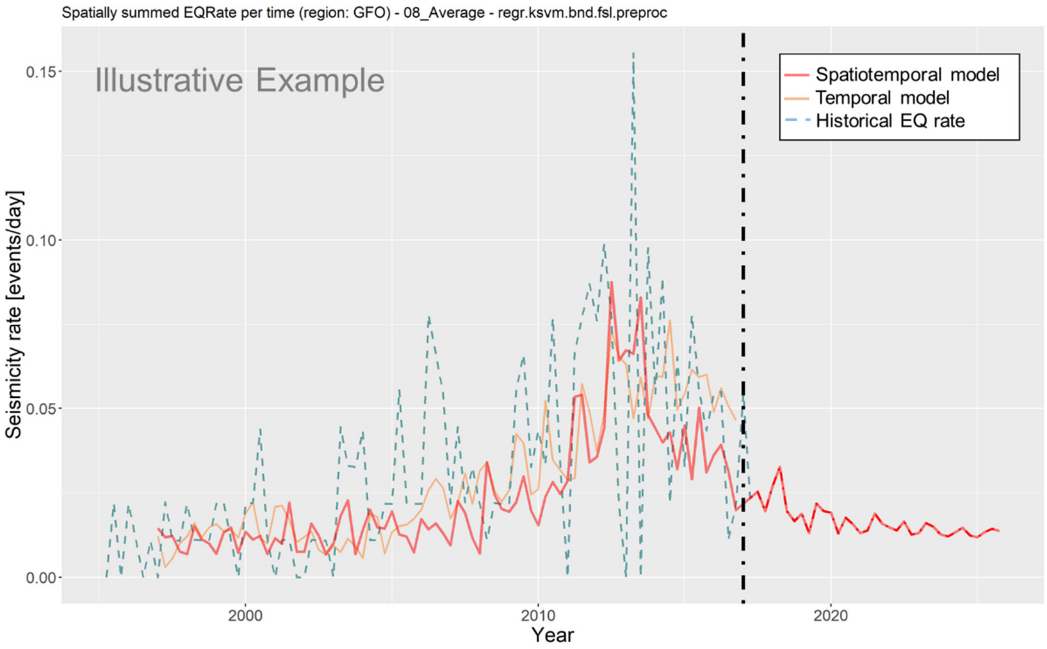


Figure 33: Illustrative example of a qualitative comparison of the temporal-only model forecasts (orange) with the spatially aggregated spatiotemporal model forecasts (red) and the historical data (dotted blue). The vertical dotted-dashed line is on December 31st, 2016, marking the end of the dataset used for training and testing the models. Left of the vertical line the algorithm is retrained after every forecast, right of the vertical line no retraining is done. The forecasts shown are for the post-March 2018 production scenario for the KSVM model with MoReS-based constraints based on 3-month aggregation periods with a 3500 m bandwidth for spatial smoothing over a 1500 m resolution grid.

Based on the variable importance of the spatiotemporal experiments, the similarity between the temporal-only models and the spatially aggregated spatiotemporal models is to be expected as the importance of the temporal features is consistently higher than that of the spatial features (Figure 34). Furthermore, the spatially aggregated performance of individual cells aggregated to the entire field should be comparable to the overall earthquake rates for the entire field.

The small differences we do observe are likely related to differences in the temporal feature selection between the temporal and spatiotemporal models: The spatiotemporal models consider pressure and compaction-related features for the individual cells, with only field-wide average production rates, whereas the temporal models consider additional features such as production, subsidence, and hydrocarbon column thickness.

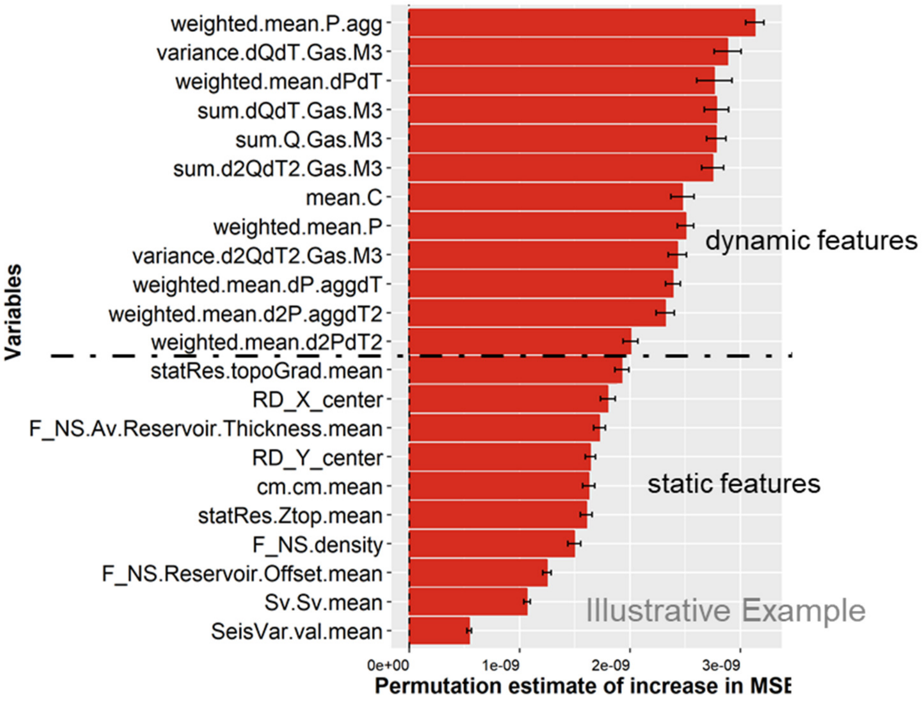


Figure 34: Variable importance for the spatiotemporal model shows that the relative importance of all dynamic features is higher than that of the spatial features.

Performance on the validation period 2013-2016

To assess long-term (1-5 years) model performance, we compare the forecasts against a fully hold-out set of events. These are events that have not been used for either training or testing during model development. to the hold-out set comprises events from the period between 1 January 2013 and 31 December 2016, which yields four years of recorded events against which forecasts can be compared. This period saw a stark change in production which lead to a decrease in seismicity, hence one further objective is to investigate whether the models can effectively capture this change in the seismic trend.

As shown in Figure 35, the KSVM model captures a downward trend over the 2013-2017 period. Furthermore, in terms of total (cumulative) seismicity forecasted over this period, the models forecast values that are within one standard deviation of the true value. Over the 2013-2017 period 76 earthquakes of magnitude 1.5 or higher where recorded. the KSVM (shown) forecasts 66 events (±100%).

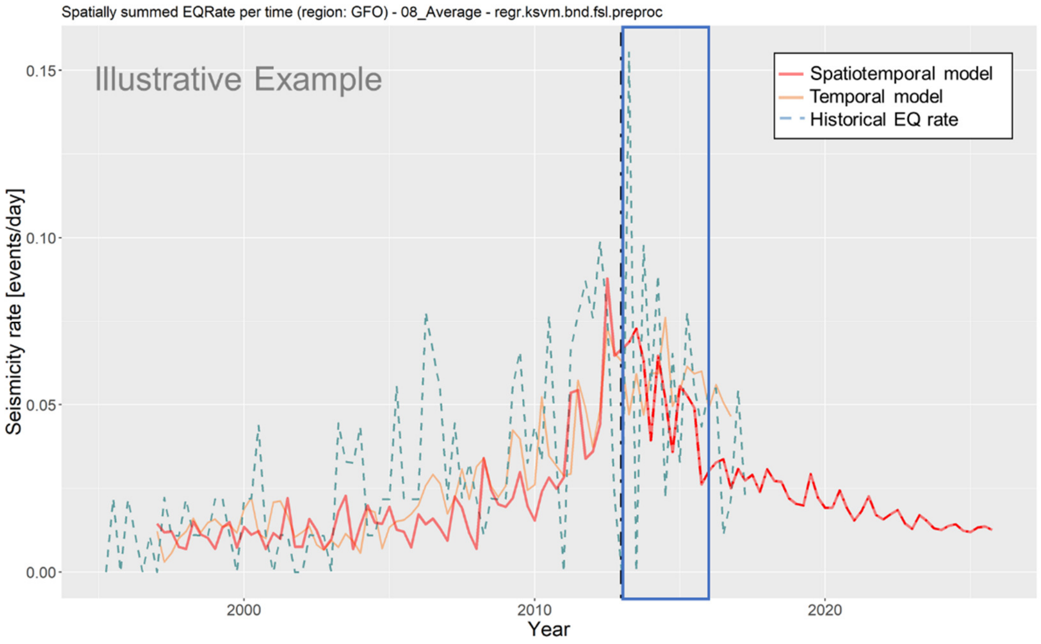


Figure 35: Illustrative example of a qualitative comparison of the temporal-only model forecasts (light orange) with the spatially aggregated spatiotemporal model forecasts (red) and the historical data (dotted blue). The validation period 2013-2016 is indicated by the blue box.

Performance of the models is comparable with only minor deviations. This is a further demonstration that machine learning models work best when they can be updated as new information comes in. The fact that the ML model can pick up on the change in trend - despite the fact that during the entire training period seismicity had an upward trend - offers support to the idea that the model is able to capture information about the underlying process.

Temporal forecasts for various production scenarios

Comparing the forecasts of different production scenarios for the forecasting period up to 2025, the post-March 2018 Average production scenario shows a declining seismicity event rate whereas the Winningsplan 2016 scenario shows a relatively steady rate (Figure 36). The relatively constant pressure decline of the Winningsplan 2016 scenario is consistent with the constant decrease in pressure over time for that scenario, compared to a decrease in depletion rate for the post-March 2018 scenario (Figure 37).

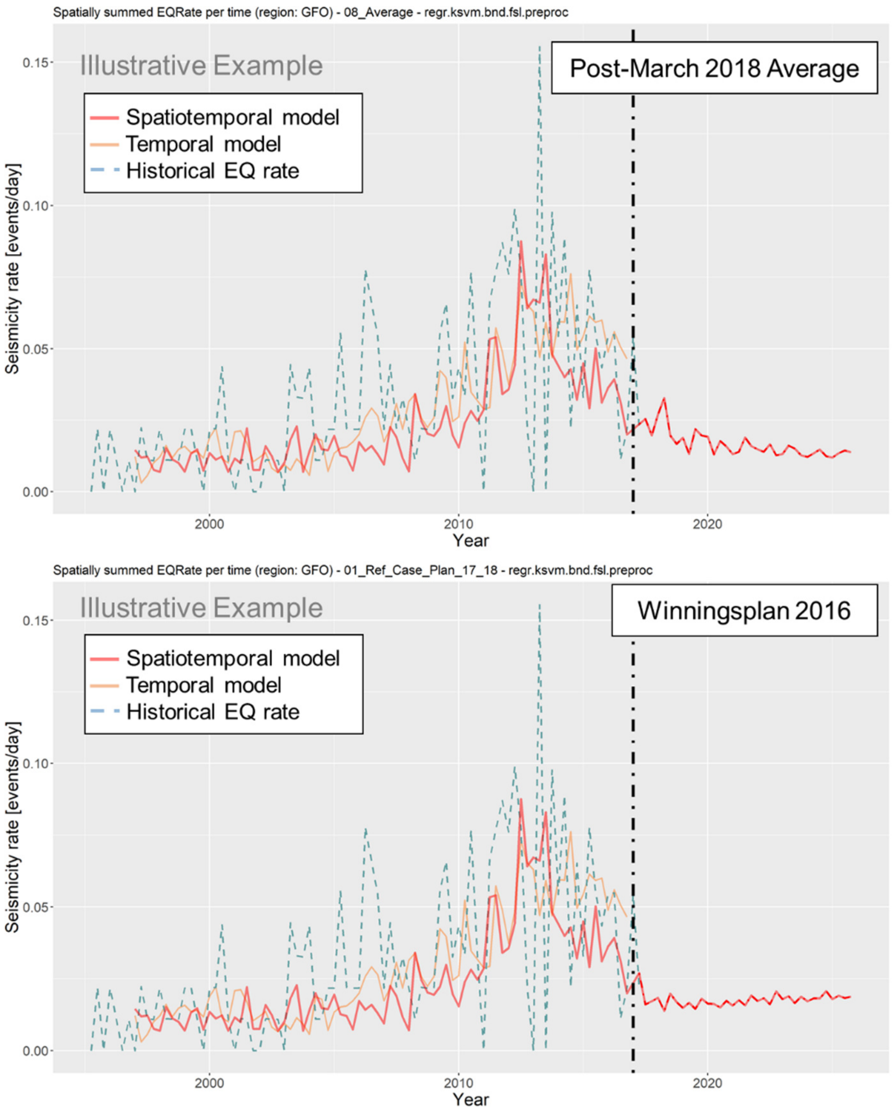


Figure 36: Illustrative example of seismicity event rate forecasts for the post-March 2018 Average (top) and the Winningsplan 2016 default (bottom) production scenario. Forecasts shown are for the same experimental setup as in Figure 33.

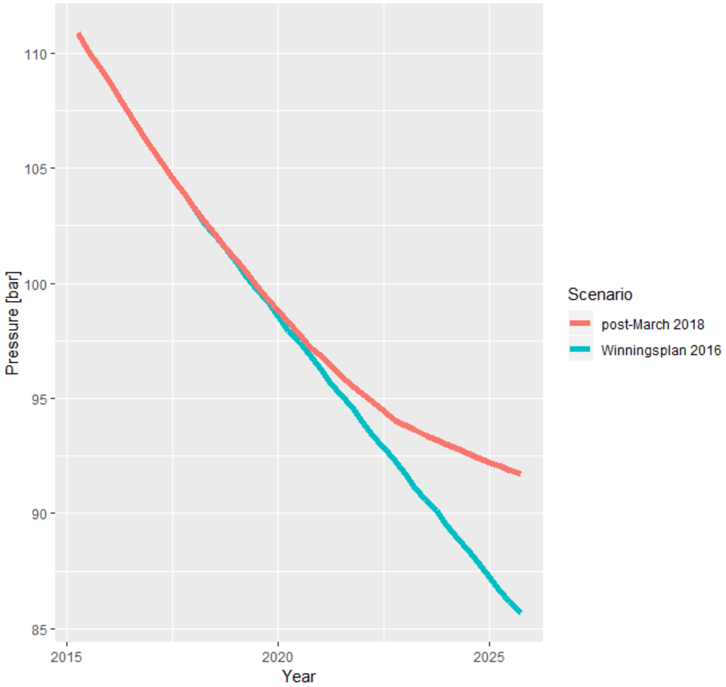


Figure 37: Average reservoir pressure in the entire Groningen field as a function of time for the post-March 2018 and Winningsplan 2016 scenarios shows a constant pressure decrease for the Winningsplan scenario versus a decreasing pressure decrease in the post-March 2018 scenario.

7.3. Evaluation of Spatiotemporal Performance

Here we show the results for the full spatiotemporal evolution of seismicity, first for the training and testing period, then for the validation period and subsequently for the two different production scenarios.

Spatiotemporal performance training and testing period

In general, the spatial distribution of forecasted earthquake rates shows a pattern similar to the true distribution of earthquakes. Figure 38 compares the forecasts of different models to the actual (spatially smoothed) events. Appendix D illustrates the number of earthquakes in each cell before smoothing, while Appendix E shows the number of earthquakes in each cell after smoothing. Qualitatively, the Random Forest and KSVM models make spatially more accurate forecasts compared to the depletion thickness MA and vertical strain thickness MA models by capturing the general area of activity more closely.

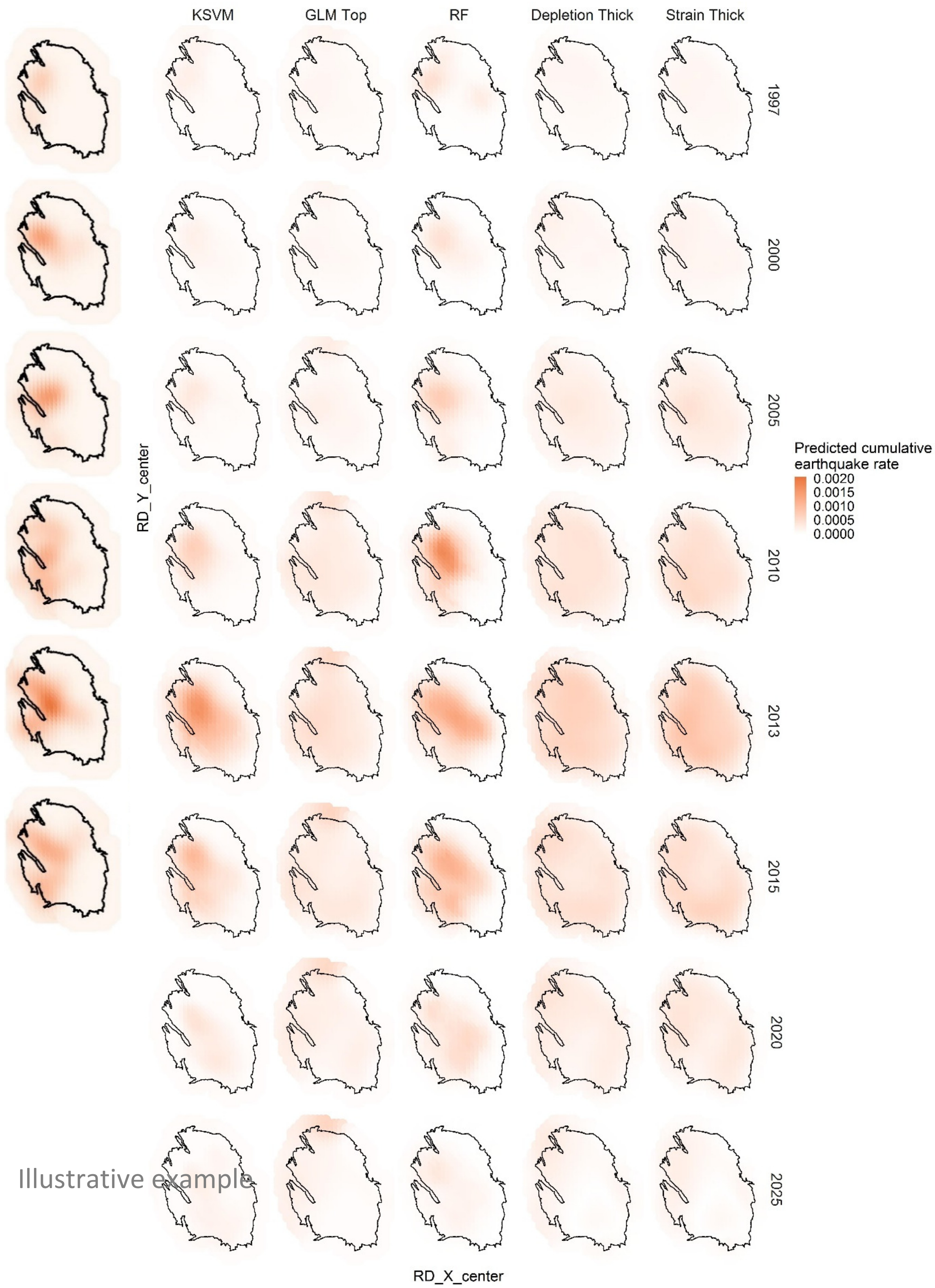


Figure 38: Model predictions (right) compared with the observed event rates(left) for the post March 2018 scenario. The colour range represents the cumulative earthquake rate per year. The full table including forecasts for all models between 2007 – 2016 is in Appendix F.

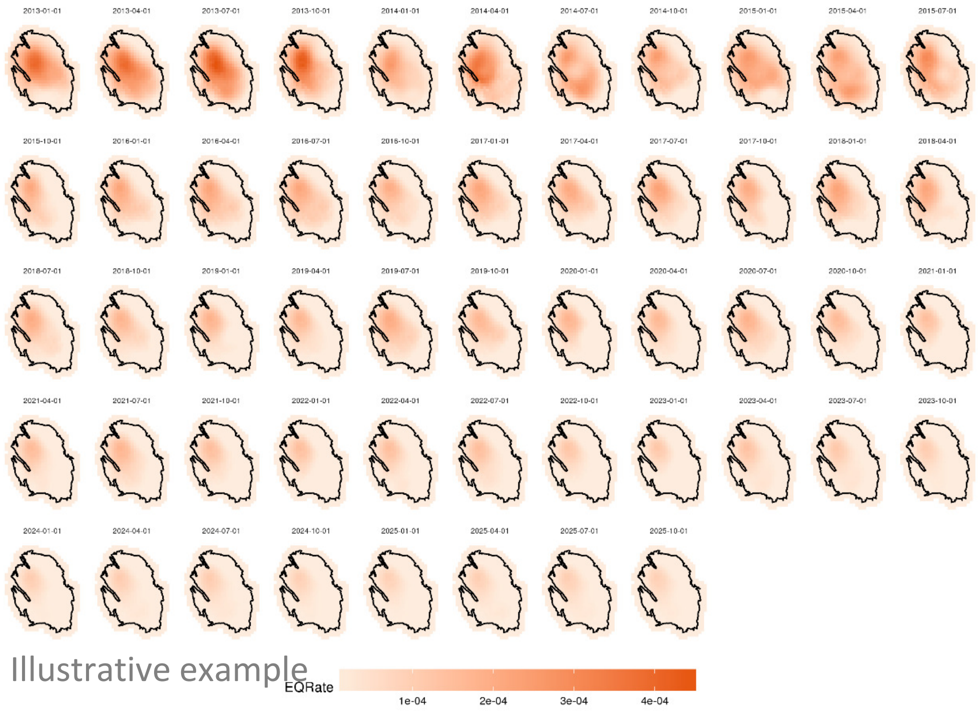


Figure 39: Example of forecasts of the spatial distribution of the daily earthquake rate (post March 2018 scenario, KSVM model) per 3-month time interval, from 1st of January 2013 (top left) to 31 December 2025 (bottom right).

Figure 39 above shows an illustrative example of the spatial density plots over time representing the forecasts for the KSVM model for the period 2013-2025. The model visually shows a decline in seismic density across the field over time, which matches the aggregated temporal performance plot in Figure 33 in terms of a declining rate over time for the post March 2018 scenario.

Moreover, to investigate some of the high level spatial differences between models, k-means clustering was used to divide the forecasts into five blocks based on grid coordinates (**Figure 40**). Although this division does not have a physics-based rationale, it can help to analyse performance differences between different parts of the field. The cumulative seismicity forecasts for various ML models and Depletion MA baselines are shown in Figure 41.

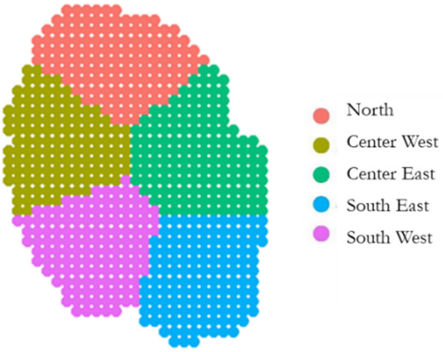


Figure 40 Spatial locations of blocks which are used to compare models' performance in different regions of the Groningen field.

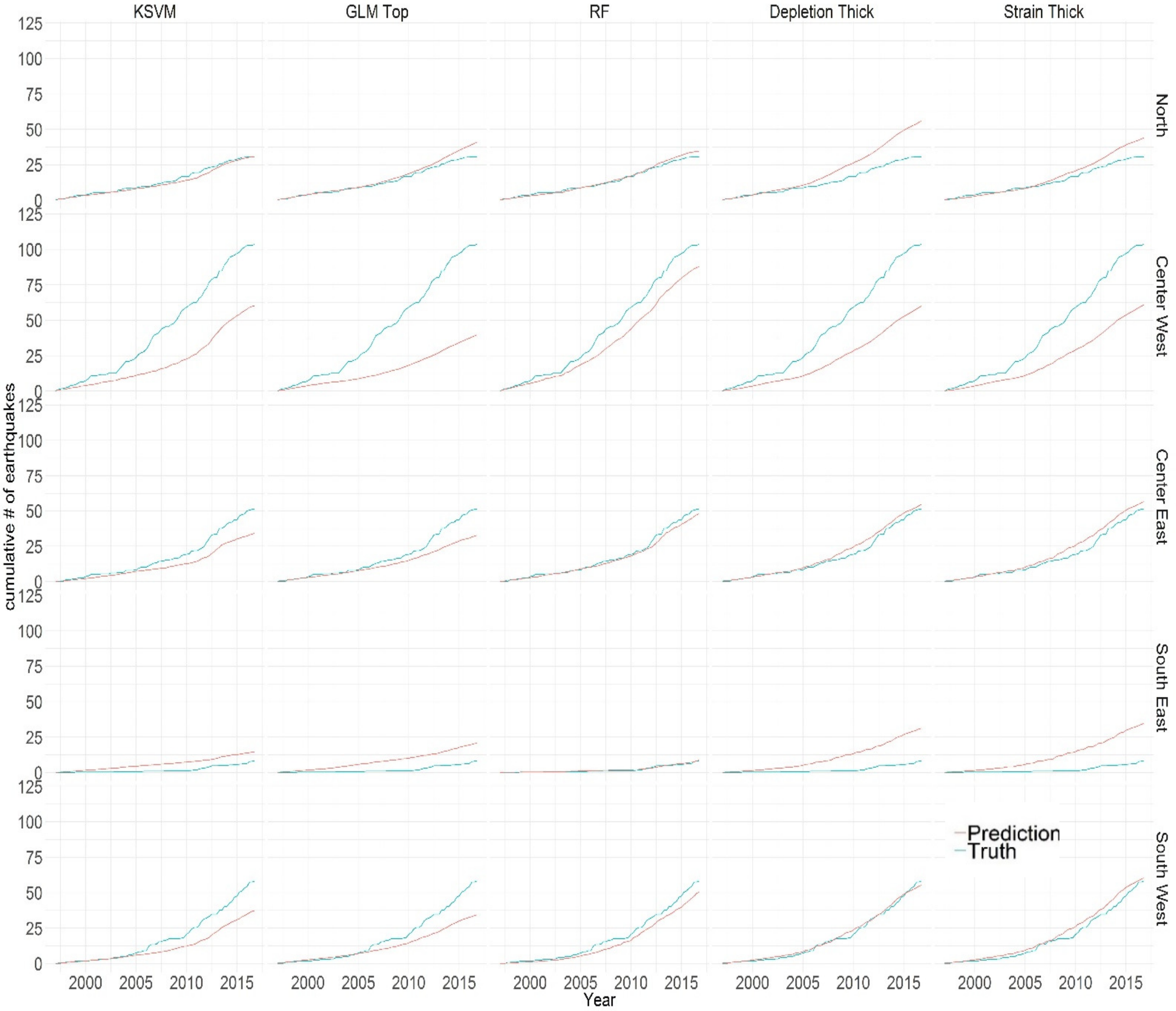


Figure 41: Cumulative number of earthquakes as predicted by the models (red line) and as in historical data (blue line). Numbers on the right side of the plot illustrate the number of blocks. Inset on the right shows the spatial distribution of the blocks 1-5.

We discuss in more detail the results of blocks in Centre West and South East, as the cumulative errors in these blocks are relatively large compared to the other blocks. As can be seen from Table 28, the Centre West block has had the most earthquakes throughout 1995 – 2016. In this block, all models underestimate the number of earthquakes, but the Random Forest model is closest to forecasting the actual number of earthquakes (88 out of 104 earthquakes; Table 28). However, all the evaluated models systematically underestimate the actual number of earthquakes in this block. The South East block has had the least earthquakes throughout 1995-2016. Furthermore, the underestimation of seismicity rate in the Centre West block and overestimation in the South East block might be the effect of patterns learned over the whole reservoir that do not translate appropriately to the discussed spatial area due to lack of spatial differentiation. New spatially dependent features that further differentiate spatial areas of the reservoir could be the key to improved spatial accuracy.

Table 28: Number of cumulative forecasted earthquakes for each model and per block. See Appendix G for additional results and a bar chart showing the number of earthquakes in each block for every year.

Region	Observed	RF	KSVM	GLM Top	Depletion Thick	Strain Thick
North	31	35	31	41	56	44
Centre West	104	88	60	40	60	61
Centre East	51	48	34	33	54	57
South East	8	9	15	21	31	35
South West	58	50	37	34	55	60
All	252	230	177	168	257	257

The table above shows that the number of events per block is significantly different, especially between the Centre West and Centre East blocks since the Central West block has more than 10 times the number of events than the Central East block. All models predict the highest number of earthquakes in the Centre West block and the lowest number of earthquakes in the Centre East block. However, as also can be seen from Figure 41, the RF model predicts the number of earthquakes most accurately in both blocks.

Spatiotemporal performance validation period

Figure 42 and Figure 43 show the model forecasts for the 2013-2017 period during training and testing and during the 5-year forecasts in the hold-out set. From a qualitative point of view, there are few differences between forecasts for the 2013-2017 period during training and testing and during the 5-year forecasts in the hold-out set. For the KSVM model, most importantly, it can capture the decline in seismicity rate over this period both when training and testing on short term forecasts (3 months) as well as when forecasting on the long term (5 years). Interestingly, even though RF, KSVM (only KSVM shown in this case) and the baselines (Depletion and Strain thickness MA) capture the decline in seismicity over this period, they do show spatial differences in their forecasted seismic rate. The KSVM model forecasts declining seismicity with activity mainly in the north-centre of the reservoir, moving slightly south over time while the baselines forecast seismicity mainly in the centre and moving towards the south of the reservoir.

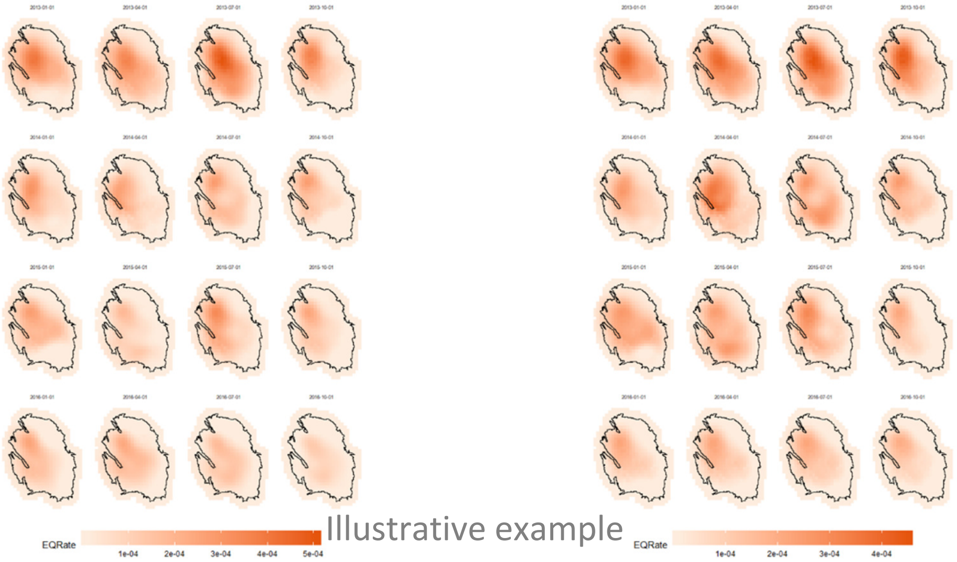


Figure 42: The KSVM model short term (3 months) forecast for the training/testing period 2013-2017 (Right) versus forecasts between 2013-2017 at 3-month intervals for the 5-year hold-out set forecast for the KSVM model trained/tested up to 2013 (Left)

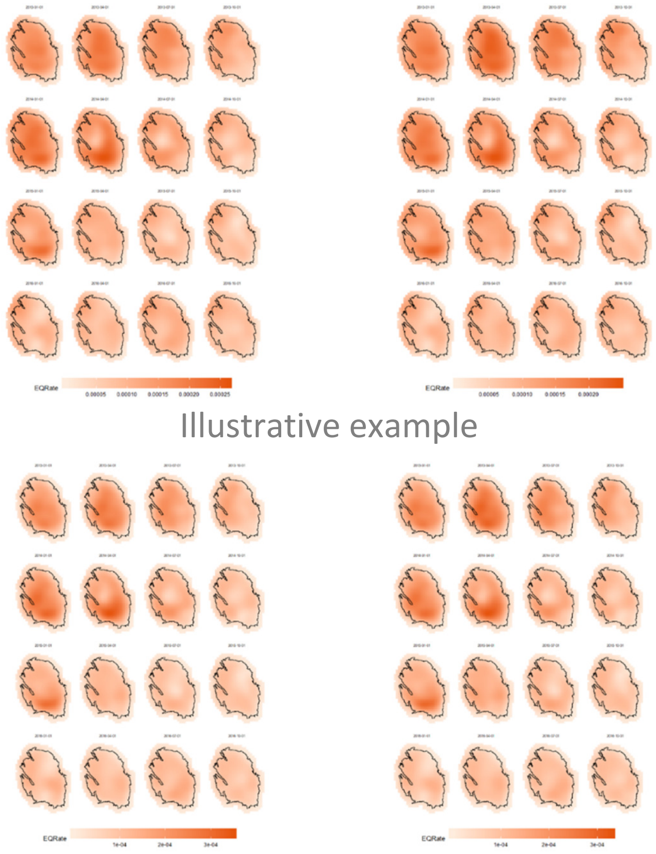


Figure 43: The Depletion Thickness (above) and Strain thickness (below) MA models short term (3-month) forecast for the training/testing period 2013-2017 (Right) versus 5-year hold-out set forecasts between 2013-2017 at 3-month intervals (Left)

Spatiotemporal forecasts for various production scenarios

The KSVM forecasts a declining trend in seismic activity for the post-March 2018 (Figure 44). This is in line with expectations since production under this scenario is scheduled to decline up to 2030 when the field will be shut-in. In sharp contrast is the forecast for the Winningsplan 2016 which assumed sustained production and for which the model density plots show a slightly higher density in seismic activity over time.

The qualitative assessment above indicates that the selected ML model can capture differences between these two production scenarios and produce forecasts that are broadly in agreement with expectations, i.e., a scenario with declining production such as the post-March-2018 seismicity is expected to decline while a sustained production scenario such as the Winningsplan 2016 seismicity is expected to continue at a similar rate.

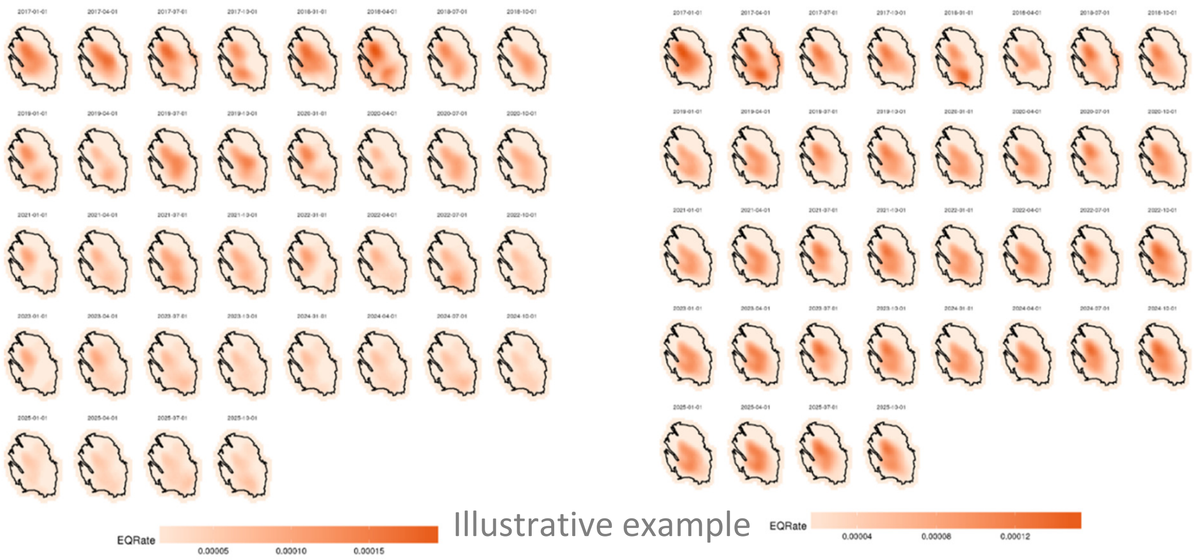


Figure 44: post March 2018 scenario forecasts (Left) versus Winningsplan 2016 scenario forecasts (right) for the KSVM model for min magnitude 1.5 for the period 2017-2025.

8. Conclusions & Recommendations

8.1. Conclusions

The aims of this study are to address the recommendations derived from the findings of Limbeck et al. (2018) and to extend their temporal analysis to a spatiotemporal machine learning approach that forecasts for each location in the field the seismicity rate.

The recommendations from Limbeck et al. (2018) relate predominantly to addressing a lack of strongly decreasing seismicity after the Groningen reservoir has been shut-in. We address this by constraining the models to a post shut-in steady state with zero seismicity rate, where all first and second derivatives of time-dependent input features are set to zero and where a range of absolute reservoir pressures is derived from multiple scenarios of the reservoir engineering flow model. As these pressure values are model-driven and not calibrated to data, we performed a sensitivity analysis on the pressure values showing that the ML models are insensitive to pressure values within the investigated pressure range. The information provided by the ultimate state data leads to forecasts that converge to an approximately zero seismicity event rate after the field has been long shut in. With these additions, the model performance, expressed in MAE, RMSLE and Mean Poisson Loss error metrics and Wilcoxon significance tests on the out-of-sample test set, has not changed significantly, based on our error metrics, when compared to the temporal models from Limbeck et al. (2018) that had not been taught the model-based post shut-in steady state constrains.

Second, the revised temporal model has been extended to a spatiotemporal model. The temporally aggregated performance of these spatiotemporal models is comparable to the performance of purely temporal models in the event rate report (Limbeck et al., 2018), which indicates that the addition of spatial dimension to the forecasts can be done without significant loss of temporal performance. Time-dependent features, such as reservoir pressure and compaction, continue to mostly drive model performance. The most significant spatial features appear to be the topographic gradient of the reservoir, reservoir thickness variations along the major faults and compressibility, which are features that from the view of the physical Coulomb strain model are likely to impact seismicity.

The spatiotemporal ML models are observed to capture spatial information to some degree, i.e., the models forecast the highest seismicity in the Loppersum area and the lowest seismicity in the south and towards the edges of the field, specifically in the southeast corner of the field. However, although there is a relative match in trends, the models systematically underestimate seismicity, meaning that extreme values in seismicity rates are not easily captured given the information and models available at this moment. This can be in part attributed to the ML model tendency to forecast towards to mean in the absence of clearly differentiated patterns and considerable uncertainty as is the case in this study. Note that similar under/over estimation patterns were already present in the previous event rate study.

Based on the analysed error metrics and especially on the results of the likelihood ratio comparison, the Random Forest model shows the best performance among the tested models, followed by KSVM. These two ML models significantly outperform the selected baselines which are based on a moving average of the depletion thickness and vertical strain thickness, for the study period 1995-2016.

Moreover, the results in section 7 of this report regarding the comparison between the post March 2018 and the Winningsplan 2016 scenario show that the ML models behave in both the temporal and spacetime setting according to physical expectations, meaning that the models forecast declining seismicity for the post March 2018 scenario under declining production conditions and a stable (slightly increasing) seismicity rate for the Winningsplan 2016 scenario which does not incorporate significant production cuts.

Finally, the ML models can capture the change in seismicity trend in the period 2013-2016 by forecasting a decrease in rate when compared against the hold-out set. However, we do acknowledge that the models struggle to capture the vertical spread and systematically underestimate the seismicity rate both in training and forecasting. These effects are similar in both the temporal and spatiotemporal models.

8.2. Limitations

These conclusions are subject to the following main limitations and potential improvement opportunities of the current methodology:

- The model accuracy and forecasting power remain limited by the small training dataset. The limitations of the small dataset are amplified in the out-of-sample-validation tests, which only have 192 events compared to the 268 events in the period up to 2016. This limitation can be addressed by extending the hold-out set to include data beyond 2016.
- Given a different production scenario where the ultimate pressure does not fall within the pressure range used in this study, the pipeline needs to be adjusted.
- Part of the spatiotemporal input data and the ultimate state points to which the ML models are constrained are themselves based on reservoir flow history matches and forecasts. Consequently, these points and thereby the long-term convergence of many ML models is only as accurate as the ultimate points themselves.
- Given the non-parametric nature of most of the ML algorithms, there is no analytical derivation from which to obtain longer-term uncertainty quantification. The conservative empirical approach from Limbeck et al. (2018) might be computationally too intensive and subjected to large variations in a spacetime setting. Especially if the assessment is done at the individual cell level, variations from one cell to another can be much larger than variations from one aggregated time period to the next which could make the confidence bands too large to be of any real use. Therefore, the temporal confidence band calculation method was not applied to the spatiotemporal approach.

8.3. Recommendations

- Investigate the alternative of forecasting cumulative counts instead of earthquake rate. This could help mitigate the perception that a steady pressure value would also lead to a steady seismicity rate.
- Include shear strain thickness, derived from topographic gradients, as a baseline instead of compaction, as shear strain thickness shows a better correlation to seismicity.
- Additional smoothing methods (e.g., anisotropic kernel smoothers) could be investigated further, as different smoothing methods may improve model performance.
- A more refined uncertainty quantification approach in both space and time is required. One solution could perhaps be to increase the time aggregation from 3 months to 1 year or more.
- We recommend doing an assessment of model performance by running without the XY coordinates as features. The location information might already mostly be encoded in the spatial features themselves. Conversely, it also makes sense to run using only XY as spatial features since this would tell us if the spatial maps contain more information than just spatial information.

A. Definition of non-extrapolating or extrapolating models used in this study

As introduced in section 1.3 and Limbeck et al. (2018), the ML models used in this study are divided into extrapolating and non-extrapolating groups. In our context, non-extrapolating models cannot forecast outside their target range of calibration, whereas extrapolating models theoretically are able to. That said, it is not trivial to access the quality of the forecasts done by models which are able to create plausible extrapolative forecasts. Hence, extrapolation outside of the training target range will likely diminish the quality of any model forecast, but this effect is likely to be amplified for models that are bounded to their training target range, e.g., RF and KNN. Therefore, in this study, we consider the following models as non-extrapolating: RF, KNN, and a Moving Average (the latter being a baseline forecasting the mean over an optimized fixed window found using our sample walk-forward validation). Given that KSVM, GLM Top (a generalized linear model trained using the 5 most significant features obtained from the variable significance analysis – see section 7 of Limbeck et al. (2018)) and the Depletion MA (a baseline that assumes that activity rate scales with depletion) are able to extrapolate - given the right internal parameter configuration - we call them extrapolating. We highlight that the results and conclusions drawn in this report are independent of this class division. This division in our context serves solely to differentiate models which were – in the Event Rate report by Limbeck et al. (2018) – able to forecast some decline in seismicity given previously unseen physical conditions (e.g., lower weighted mean pressure). Models which can forecast outside of their training target range forecasted a modest decline, though the default PSHRA event rate forecasts decline substantially faster after 2021. Models which cannot forecast outside of their training target range had difficulty with this future scenario as they forecasted unphysical behaviours, such as a stable or even increasing event rate with decreasing pressure.

B. Spatiotemporal data sources

For each data source from which features for the ML models were extracted, a summary of the data source origin and main uncertainties is given.

Seismic data

To capture the geometry and rock properties of the reservoir and overburden in 3-D, seismic reflection data is used. Seismic data provides a 3-D view of the geometry of the subsurface and the distribution of time-invariant rock properties, including reservoir depth, thickness and the distribution of faults.

The geometrical features are generally manually interpreted from seismic reflection data and serve as a basis for the geometry of the static and dynamic reservoir models, but as the manual interpretation and model construction introduce potential biases and errors, we opt to include features that are directly extracted from the seismic data as proxies for spatial trends in the reservoir. These features are referred to as attributes. Attributes are properties calculated from the 3-D seismic cube, e.g., maximum amplitude, that act as proxies for geological features, e.g., faults (Chopra and Marfurt, 2007). The choice for which attributes to consider from an extensive and ever-increasing collection of known attributes is driven by the aim to accurately quantify two of the main physical drivers in the Hazard and Risk Assessment models (Bourne and Oates, 2017):

1. Structural heterogeneities (faults, fracture corridors) in the reservoir that can create local pressure differences (i.e., the throw-thickness distribution);
2. Spatial variations in the overburden that may contribute to heterogeneities in the overburden stress, which is the principal stress component in the reservoir, and as such impacts the normal and shear stress acting on faults in the reservoir (i.e., the Coulomb stress distribution, see Dieterich, 2007; Zoback, 2007; Bourne and Oates, 2017).

We focus on attributes that are derived directly from the seismic data with minimal additional processing steps and with no or few subjective processing decisions.

The available seismic data is a 3-D seismic cube for the Groningen field that was created by NAM in 2015 (Wervelman, 2015). There is no individual single seismic survey that covers the entire Groningen field, but instead, there is a collection of legacy surveys that have been reprocessed and integrated into a dataset by CGG, referred to as "GreaterGroningen R-3089" (Figure 45). This dataset consists of 23 seismic surveys, that were mostly acquired in the late '80s and early '90s.

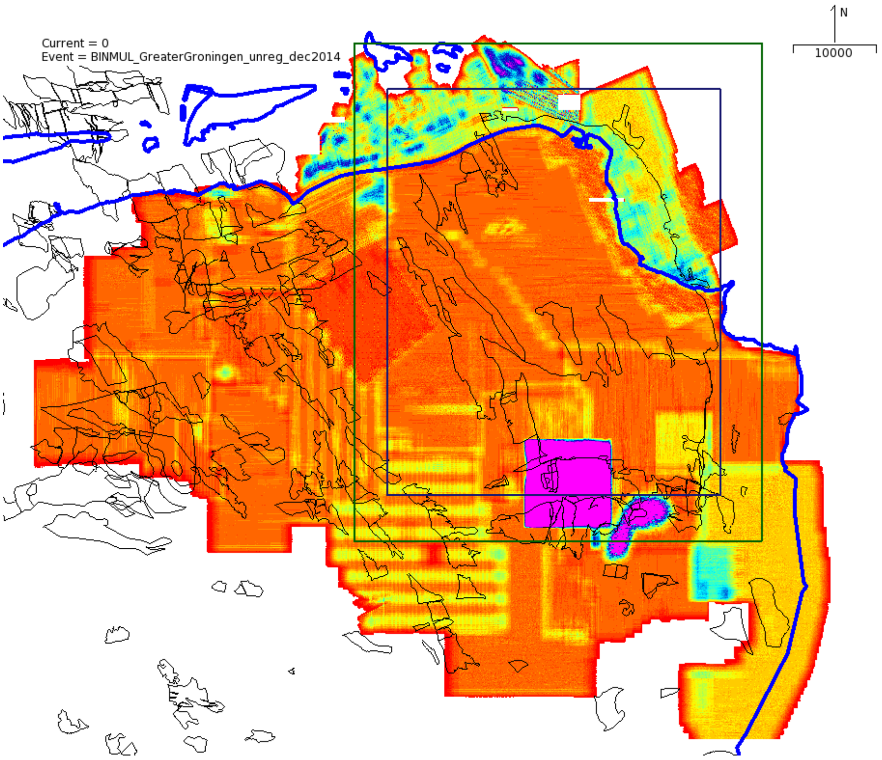


Figure 45: Northeast region of the Groningen province (coastline/border with Germany in blue). The colour-filled areas indicate coverage of the 'GreaterGroningen' seismic dataset. The different colours represent different legacy seismic datasets, and the colour intensity shows the data density of these datasets, where cold colours represent low density and warm colours high density. The thin black lines are field outlines. The Groningen Pre-SDM dataset is indicated by the blue rectangle (Wervelman, 2015)**.**

A dedicated dataset for the Groningen field was extracted from the Greater Groningen data, which includes 13 of the legacy surveys. This dataset is referred to as "Groningen Pre-SDM" (Pre-Stack Depth Migrated[7]). Full details on the processing of this dataset are given in (Wervelman, 2015), but below follows a brief summary focusing only on what is relevant for our application:

- The inline/crossline spacing, i.e., spatial data resolution, of all surveys is 25 meters.
- The seismic data was depth-converted using an interval velocity model, where average horizontal and vertical seismic velocities are used for each of the overburden formations. The depth-conversion and velocity model were calibrated with seismic-to-well ties.
- The final refined reservoir horizon interpretations were made on the pre-SDM stacked data:
    - The top Rotliegend reflector is easy to identify and map (i.e., low uncertainty). All attributes are mapped to this horizon.
    - The base reservoir reflector is more challenging to interpret and is guided by the assumption that the large faults displace top and base reservoir equally and that the reservoir thickness does not change significantly over short distances.

[7] Pre-SDM refers to the processing method. Seismic data in its raw form represents the subsurface travel times of acoustic waves from source to receiver, which are both located at the surface, and are typically some distance apart. The acoustic waves travel not only vertically but also laterally from source to receiver, creating an uncertainty in the exact location of the subsurface reflector. Through use of migration algorithms, source and receiver locations are 'mapped' to one location, increasing the accuracy of the geometrical constraints on the subsurface reflector. Migration can be done in time and depth, in different stages of the processing, where pre-stack depth migration is the most computationally intensive but also the most accurate.

    - The Top Zechstein Fm. is difficult to map, and as a result, uncertainties in the thickness of the Zechstein are relatively large. There are no uncertainty quantifications available for this deterministic interpretation, but to address the uncertainty, the interval velocity of the Zechstein Fm., which carries less uncertainty, is included as a second proxy for the Zechstein formation.
- The dominant wavelength is approximately 78 m, based on 50 Hz dominant frequency and velocity in the reservoir interval of 3900 m/s (Yilmaz, 2001; Romijn, 2017). Reservoir thickness varies between 100 m in the southeast and 300 m in the northeast, i.e., between approximately one and four wavelengths. The vertical resolution is a quarter of the dominant wavelength, i.e., 20 m. This resolution limit implies that to recognize faults on seismic, horizons need to be offset by at least 20 m.

The key uncertainties and limitations that need to be considered for the use of seismic attributes are:

- The horizontal resolution is limited to the spacing of the inlines and crosslines in the 3-D cubes, i.e., 25 x 25 m. The uncertainty in the horizontal resolution of the seismic data is negligible, as the pre-SDM process minimizes migration errors.
- The vertical resolution limit is approximately 20 m. Faults with an offset of less than 20 m cannot be detected, and faults with offsets close to this limit are likely undersampled.
- The uncertainties in the horizon interpretations are in general low, because of the complete coverage of the field by 3-D seismic, calibrated to a large dataset of wells. There are, however, relative differences in the quality of the interpretation of different horizons (Figure 46):
    - The top reservoir reflector is a bright easy-to-interpret reflector, i.e., minimal uncertainties.
    - The base reservoir reflector is less distinct. Interpretation of this reflector is guided by the assumption that the reservoir units are conformable. The base reservoir is not used directly, but only marks the lower boundary for the RMS and mean amplitude attributes.
    - The top Zechstein Fm. the reflector is not easily interpreted, because of the irregular geometry of the top of the salt, and the relatively small density contrast between top Zechstein and overlying formations. If we assume that the error margin is limited to the interpreter picking an underlying or overlying reflector as top Zechstein rather than the true top, the uncertainty in Zechstein thickness is +/- 90 m, based on the wavelength in the Zechstein Fm. (V_p in halite is 4400 m, with an approximate frequency of 50 Hz).
- Visual comparison and correlation analysis of the amplitude maps compared to the Zechstein thickness map indicate that part of the amplitude signal in the reservoir may be caused by variations in the overlying Zechstein instead of discontinuities within the reservoir.

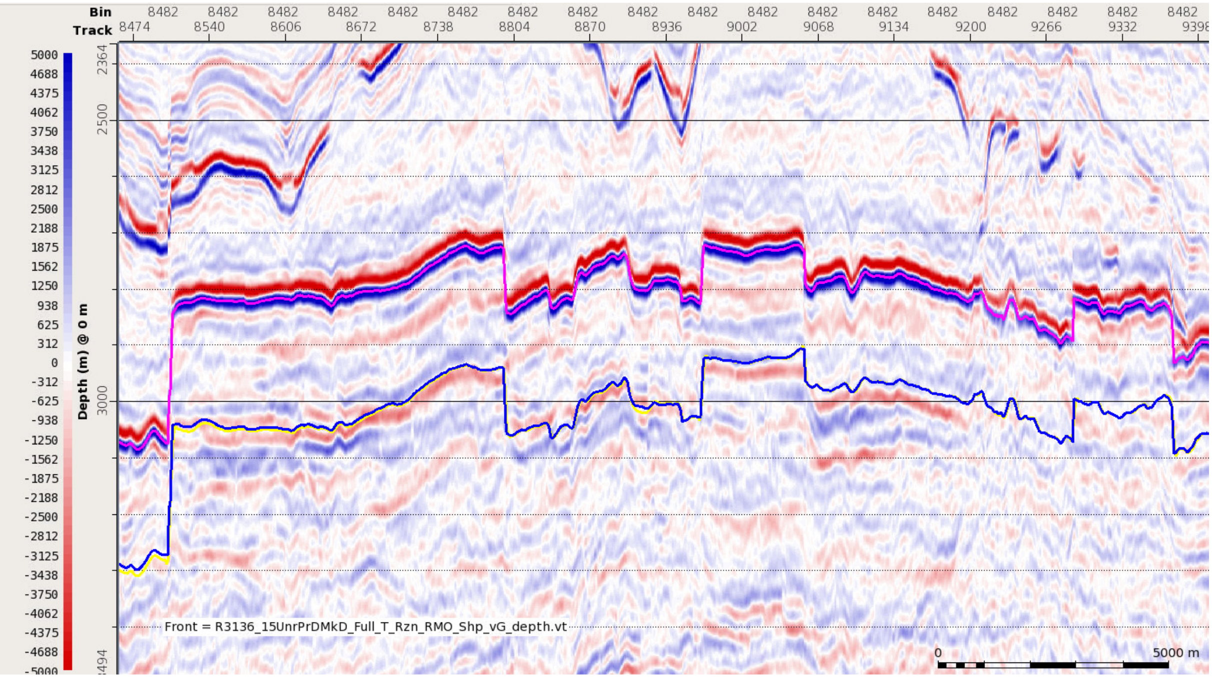


Figure 46: Arbitrary cross-section (in depth, with exaggerated vertical scale) showing the interpreted reflectors of the top reservoir (pink) and base reservoir (blue). The base reservoir reflector is difficult to interpret and therefore guided by the assumption that the reservoir thickness is conformable.

Regretted seismic attributes

Two additional seismic attributes were investigated as potential features for the ML models but were not further pursued based on a lack of correlation with seismicity.

RMS Amplitude

The RMS (Root Mean Square) amplitude is used as a proxy for heterogeneities within the reservoir. The RMS is calculated over the amplitudes within a vertical window, which is, in this case, the reservoir interval:

$$x_{\mathrm{RMS}} = \sqrt{\frac{\sum_n x_i^2}{n}}$$

The RMS amplitude map for the Groningen reservoir clearly shows lineaments representing faults in the reservoir, as well as patches of higher RMS amplitude values (Figure 47). These patches may represent trends in reservoir rock properties, e.g., porosity, though it should be emphasized that by definition, RMS gives relatively noisy results by taking the square of each amplitude value. For this reason, mean amplitude was preferred over RMS amplitude.

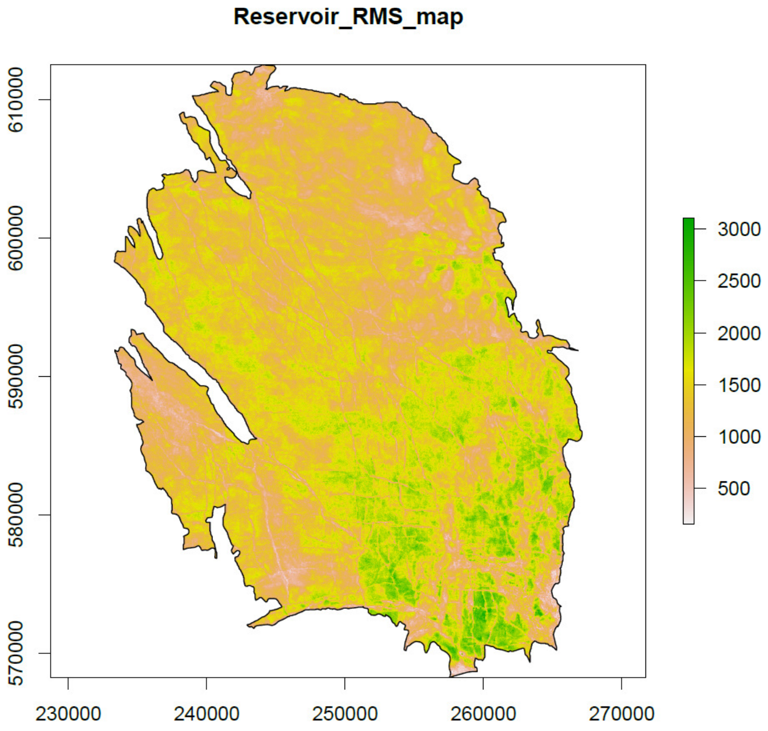


Figure 47: RMS amplitude over the reservoir interval, projected onto the top reservoir.

Ant tracking

Ant-Tracking is a proprietary algorithm in Schlumberger Petrel aimed at identifying faults. It is an edge detection method that is guided by settings for threshold (separating features from noise), and orientation, as often the dominant orientations of the faults are known. Ant-tracking is typically preceded by several pre-processing steps to reduce the amount of noise in the data (Kortekaas and Jaarsma, 2017):

1. Structurally oriented smoothing filter for random noise attenuation and discontinuity enhancement.
2. Variance (edge) detection applied to the smoothed cube.
3. Petrel Ant Tracking.

The Ant-Tracked result is a seismic volume attribute containing the ant-tracked traces. A section through this volume at the top reservoir level shows the lineaments that were identified by Ant-Tracking (Figure 48). The Ant-Track dataset has been used by EBN to re-interpret the faults in the Loppersum area, where they found that the location of Ant-Tracked faults was in better agreement with the calculated epicentre of the Huizinge tremor than the manually interpreted faults (Kortekaas and Jaarsma, 2017). However, a more exhaustive analysis of the location of Ant-Tracked faults versus all mapped seismic events has not been made in that study.

Ant-tracking has been considered as a proxy for faults that is less prone to interpretation artefacts compared to the Petrel fault model, but this feature has been regretted in favour of the variance attribute as ant-tracking tries to visualize individual faults, which results in a relatively scattered distribution without clear trends in orientation or density. The resulting map shows no correlation with seismicity, whereas the variance cube does highlight zones of increased fault density and an apparent organization of fault orientations, which correlate to some extent with seismicity.

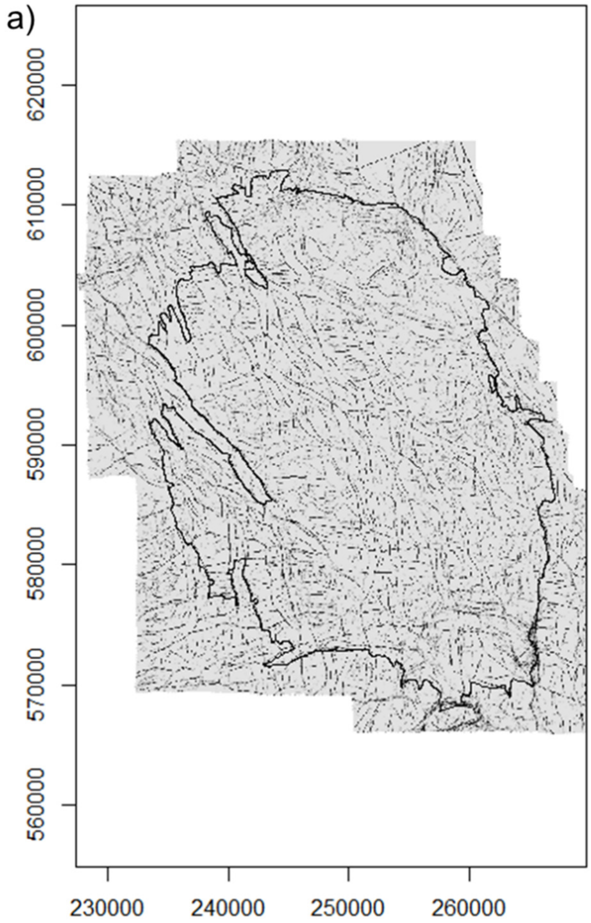


Figure 48: Ant-Track cube attribute results extracted for the top reservoir surface (Kortekaas and Jaarsma, 2017).

Overburden (Zechstein) data

Variations in overburden density are predominantly associated with the Zechstein Fm. From a geological point of view, the bulk of the Zechstein Fm. is made up of halite with variable thickness and a basal anhydrite layer directly above the reservoir, with a constant thickness of 50 m. The halite is the source of heterogeneity for both geometry and rock properties:

1. Geometry: the salt 'creeps', i.e., over geological timescales it behaves as a fluid. The base is spatially relatively consistent, thanks to the basal anhydrite, but the top Zechstein is highly variable and difficult to map on seismic (Figure 49).
2. Rock properties: The halite contains high-density anhydrite/carbonate floaters (Romijn, 2017), which are heterogeneously distributed in the halite: they can be absent, present as discontinuously present or continuously present. Floater thickness varies throughout the region. Generally, the floaters can be identified on seismic (Figure 50).

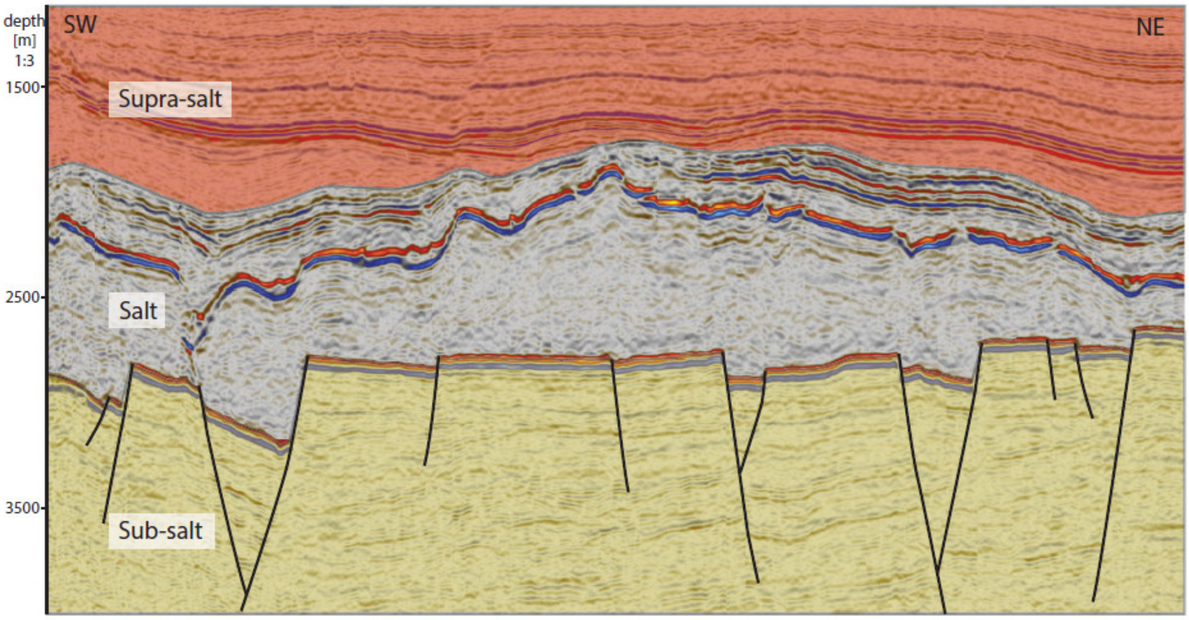


Figure 49: Arbitrary cross-section through the Groningen field, showing the variable thickness of the Zechstein Fm. (the part of the section that is not coloured). Image from Kettermann et al. (2017).

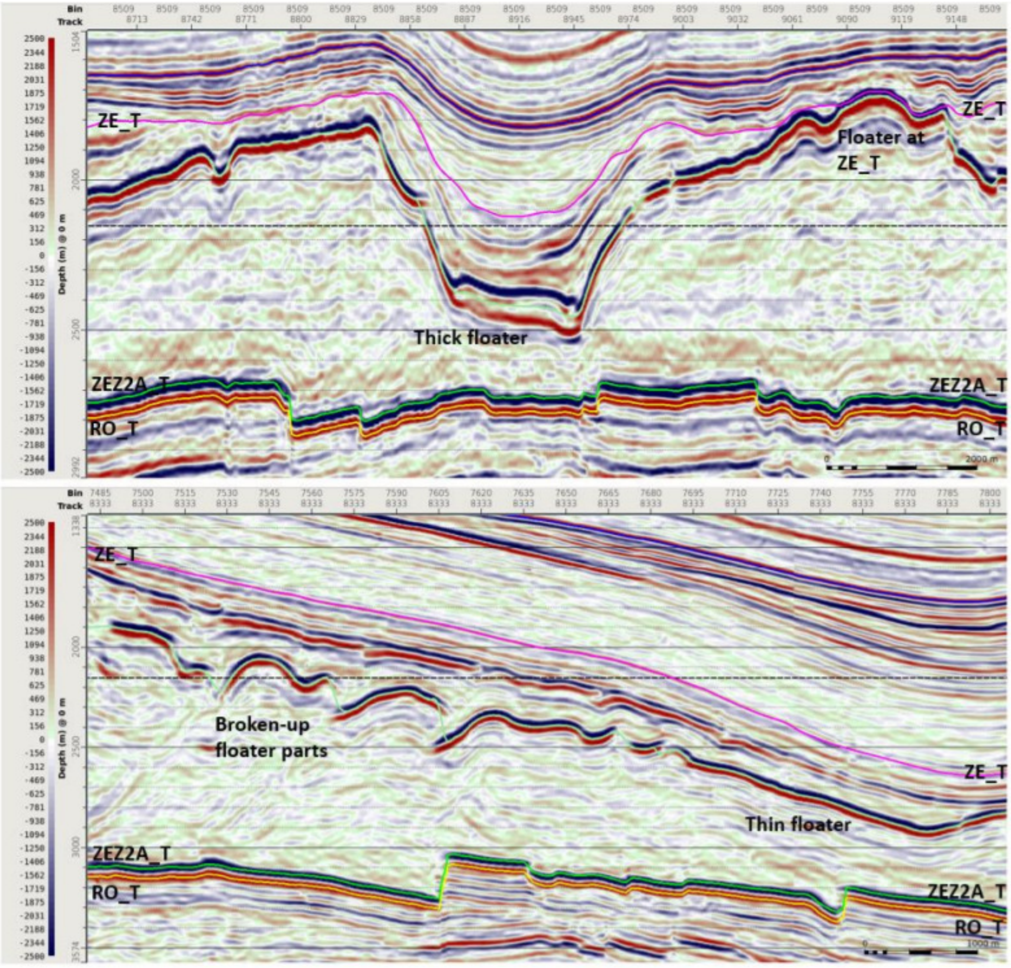


Figure 50: Seismic cross-sections of the Zechstein (top Zechstein in pink, base in green) with examples of the different floater geometries that are encountered (Romijn, 2017).

The anhydrite floaters in the Zechstein Fm. have been mapped as a separate internal unit, with zero thickness where the floaters are absent (Romijn, 2017). The anhydrite has a higher density and velocity than the halite. The combination of density contrast and variable presence of the floaters results in a spatially variable velocity and density model throughout the Zechstein.

Static reservoir model

The static model describes the reservoir geometry and the time-independent rock properties of the reservoir. The geometry is based on the depth-converted structural interpretation of seismic data and constrained by wells, and the rock properties are calculated from well data and interpolated between wells to generate a 3-D property distribution (Nederlandse Aardolie Maatschappij, 2016d). Although the static model carries more uncertainties compared to the data it is based on, the use of the model for feature generation has two main advantages:

- The model provides discrete, absolute measures of deterministic features, such as fault offset in meters or fault orientation in degrees, whereas seismic attributes only provide dimensionless proxies.
- The model provides an interpolation of rock properties measured from sparse well data, which is partly constrained by using the modelled properties in history-matched flow simulations.

The static model has been generated by NAM using the modelling software platform Petrel (Schlumberger) (Nederlandse Aardolie Maatschappij, 2016d). The lateral resolution of the model is 100 x 100 meters, with a variable vertical resolution down to 1 m thick cells.

The rock property model is based on interpolation between wells and upscaling of well data. Well log data used to calculate porosity typically has a resolution of 20 cm, whereas the vertical grid cell resolution in the model is at least 1 meter. The upscaling of logs to grid cells is done using arithmetic averaging, potentially averaging out porosity peaks. The Groningen reservoir has a relatively homogeneous geology with thick units of reservoir rock, so the effect of averaging on porosity is negligible.

The interpolation of well data between wells to populate a 3-D porosity model is driven by the acoustic impedance, which has a spatial resolution of 25 m (seismic resolution). Porosity data of each well is correlated to seismic velocity measurements at the well location (e.g., checkshot data), and using regression, a linear trend is derived from the crossplot. The same process is used for shale content and water saturation. The uncertainties associated with this process are small compared to porosity interpolation based on geological concepts, which was used in earlier models of the reservoir. Uncertainties in porosity are lowest at the well locations and increase with increasing distance from wells, but overall uncertainties are sufficiently low for the NAM to use only a single porosity model, instead of a distribution of models.

The gas saturation property in the carboniferous underlying the Rotliegendes reservoir carries a larger uncertainty, particularly in the northern part of the field, as it is based on interpolation of data from 25 wells using kriging rather than quantitative seismic interpretation.

The fault model is a deterministic interpretation made by geologists based on the identification of horizon discontinuities in seismic sections (pers. comm. A. Wood, 2018). Faults are not visible as discrete features on seismic but are only recognized by an offset in horizon reflectors. As the vertical resolution limit of seismic is typically between 20-30 meters, it follows that faults can only be reliably observed if there is at least 30 m vertical displacement along the fault, resulting in an undersampling of smaller faults (Figure 51).

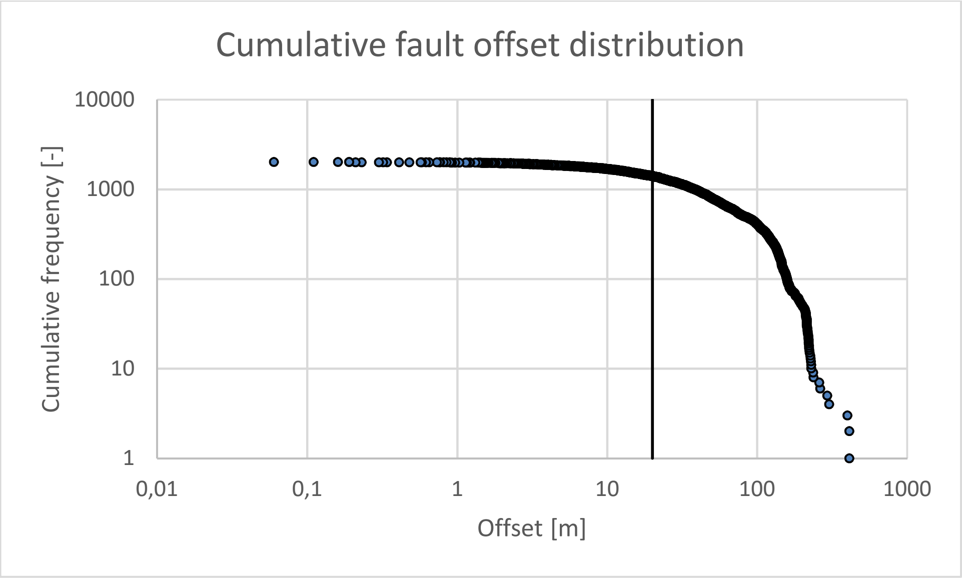


Figure 51: Cumulative distribution of fault offsets extracted from the geological reservoir model (points) and the vertical resolution limit of the seismic (line). Data from the reservoir geological model (NAM), processed by P. van den Bogert. The offsets are calculated from the displacement of modelled reservoir grid cells across faults, where each datapoint represents the offset of a fault in a single grid cell. These displacements approach zero towards the termination point of faults, explaining the sub-meter scale offsets.

A second limitation of deterministic fault interpretation is that faults are interpreted as discrete surfaces, whereas seismic sections do not show faults as discrete features but as deformation zones (Figure 52). The exact geometry of the fault, especially the dip angle, is uncertain, resulting in a potential error of several degrees in dip and in the order of 100 m in spatial position, but the geologist makes a discrete interpretation without quantification of these uncertainties.

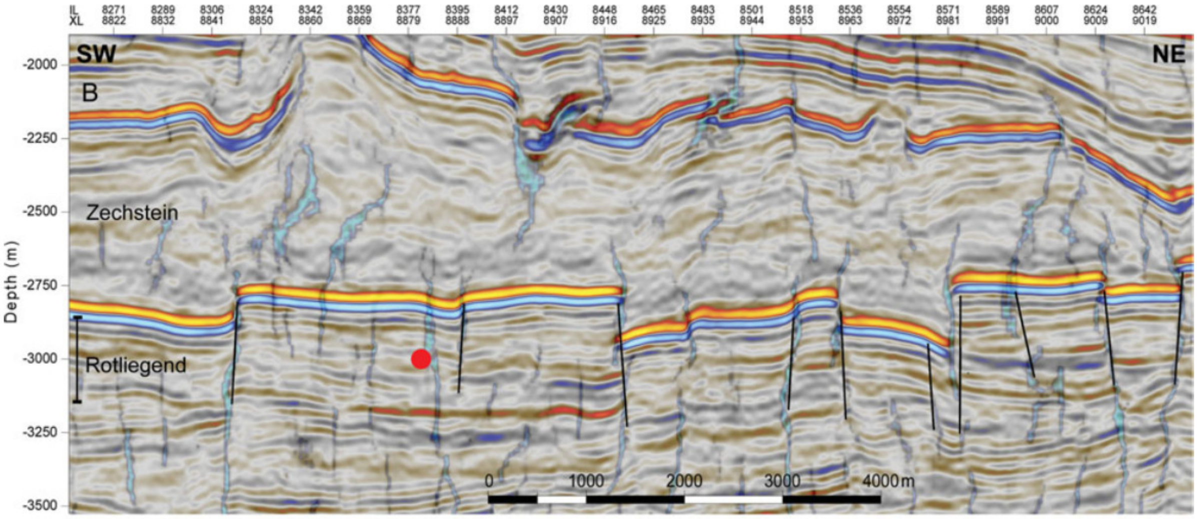


Figure 52: Arbitrary depth-converted cross-section through the seismic volume of the Groningen reservoir, showing several of the mapped faults (black lines) as well as the calculated epicentre of the Huizinge tremor (red). Note the exaggerated vertical scale. The shaded blue sub-vertical features indicate structural deformations based on the ant-tracking algorithm, and a fault interpretation driven by this attribute would yield an interpretation significantly different from the existing one (Kortekaas and Jaarsma, 2017).

The fault attributes, therefore, carry relatively large uncertainties associated with the deterministic, subjective interpretation of an intrinsically uncertain dataset, as the fault geometry (trajectory in 2-D) and throw are obtained from the deterministic reservoir geological model.

Geomechanical model

The geomechanical model from which the overburden stress used in this study is obtained is a Finite Element model that provides the 3-D stress distribution prior to and during production. We obtain the pre-production overburden stress from a geomechanical model constructed using the multiphysics solver COMSOL (pers. comm. R Wentinck, 2018). The model geometry meshes directly from the seismic interpretations of the reservoir and overburden horizons, and mechanical rock properties are assigned per geological formation. The rock property data is obtained from well logs. The overburden stress distribution is qualitatively calibrated to initial pressures from the reservoir flow model. The resolution of the model is variable, based on the curvature of the surfaces and the complexity of fault surfaces, but the resolution is typically higher than the geological model resolution. The overburden stress distribution is extracted from the geomechanical model for the top reservoir surface.

C. Complete feature crosscorrelation analysis

The features listed below are all features derived from a range of spatial and dynamic data sources, prior to any feature down-selection being applied.

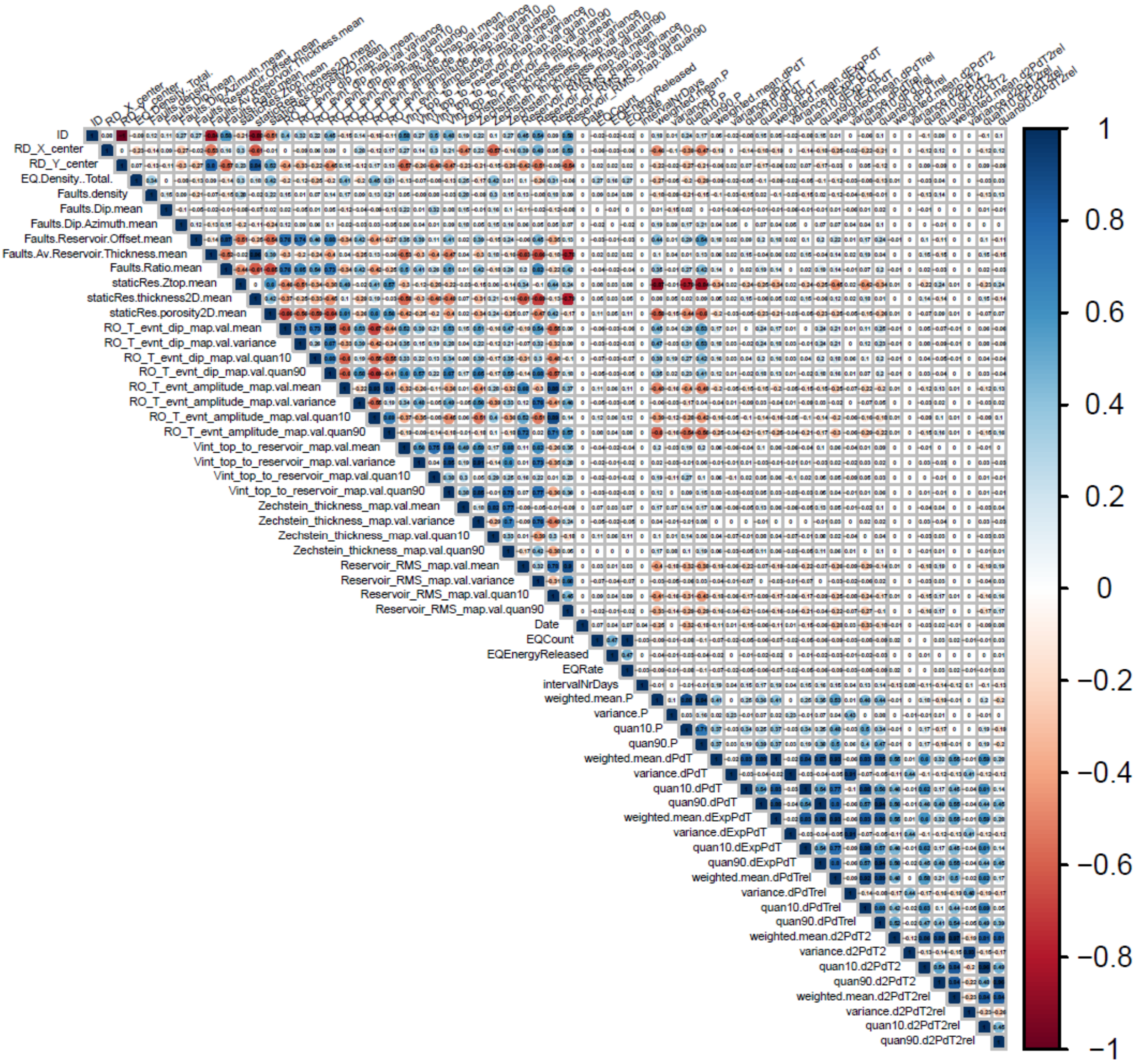


Figure 53: Feature correlation analysis for spatial and spatiotemporal features.

D. Earthquakes in 1997 – 2016

Figure 54 shows the number of earthquakes recorded in each cell. The figure illustrates the number of earthquakes before smoothing.

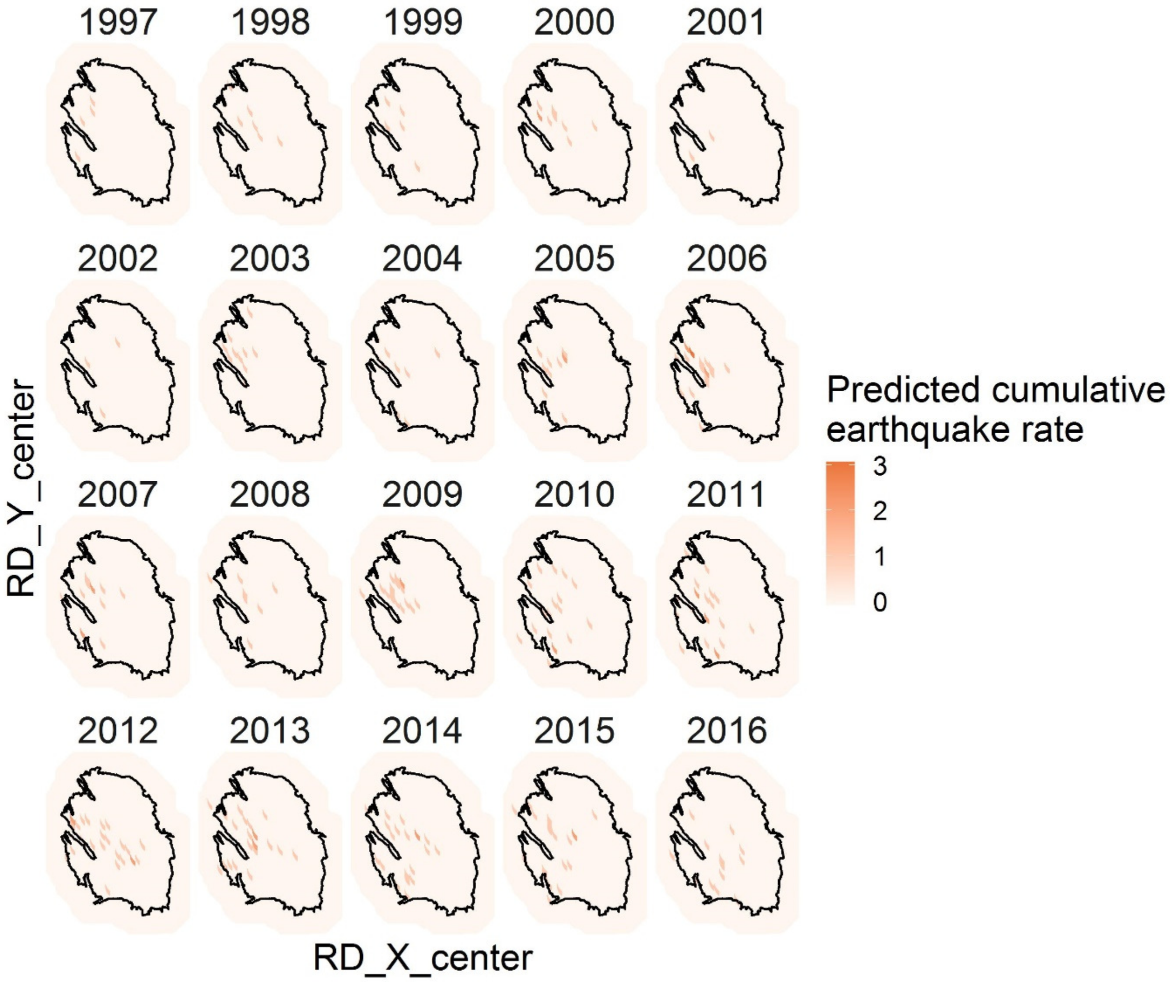


Figure 54: Number of recorded earthquakes with a magnitude of at least 1.5 between 1997 - 2016.

E. Smoothed earthquakes in 1997 – 2016

Figure 55 shows the number of earthquakes recorded in each cell. The figure illustrates the number of earthquakes after smoothing. A section of this chart is set to represent the number of observed earthquakes in Figure 38.

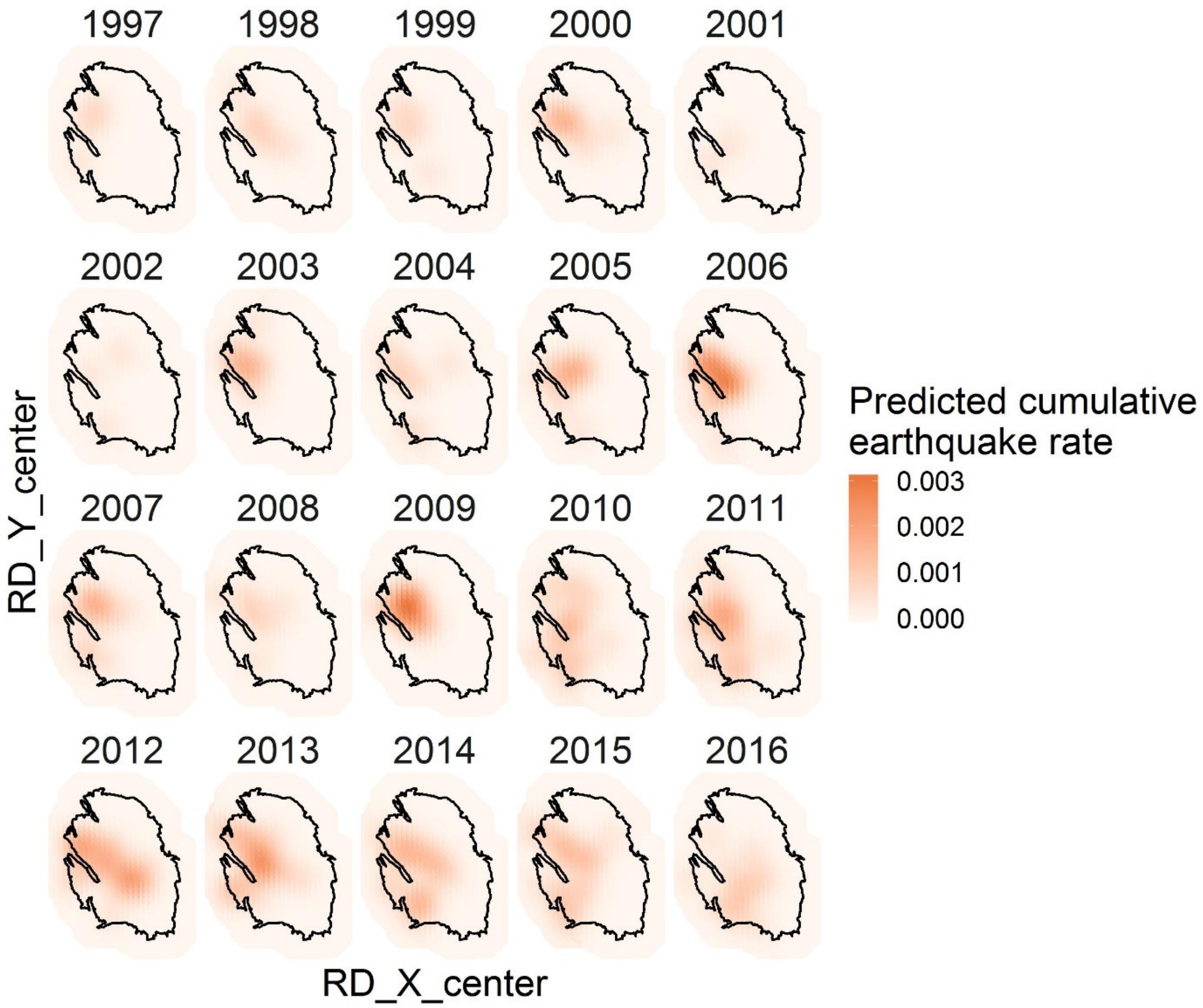


Figure 55: Number of earthquakes in each cell after smoothing.

F. Spatiotemporal seismicity forecasts

1997 – 2006

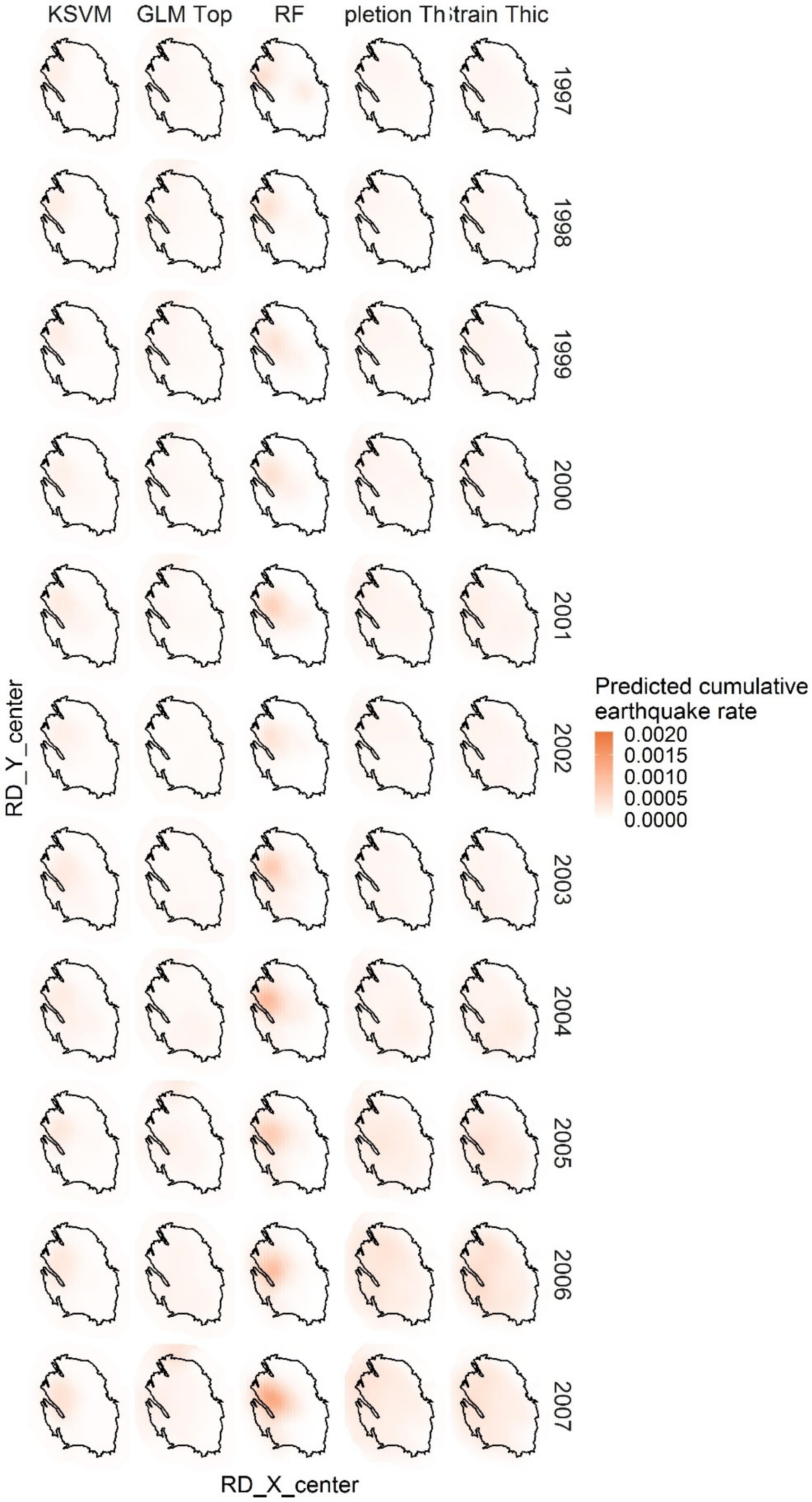


Figure 56: Predictions of the KSVM, GLM Top, Depletion MA, and RF models for 1998 - 2006.

2008 – 2016

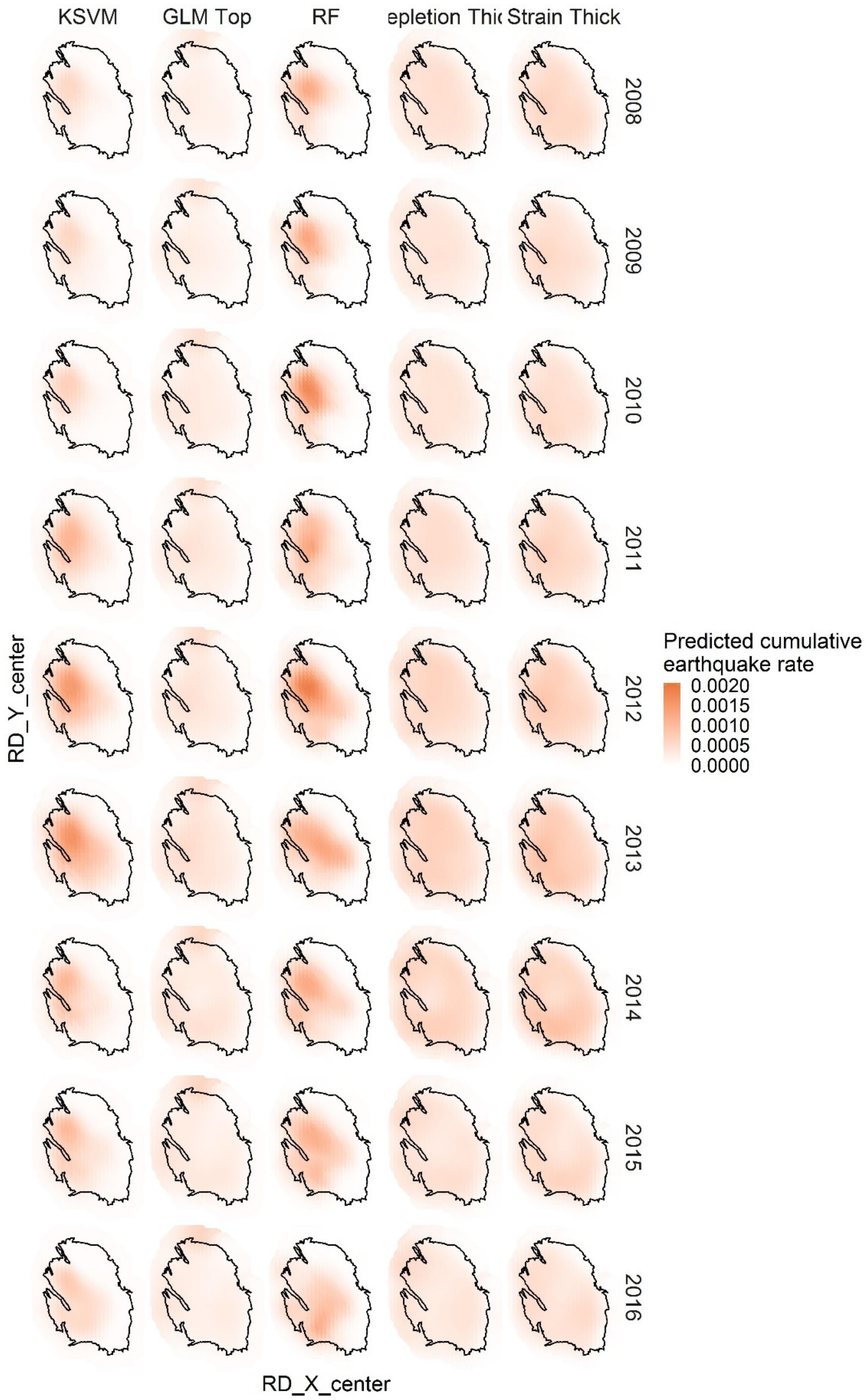


Figure 57: Predictions of the KSVM, GLM Top, Depletion MA and RF models for 2007 - 2016.

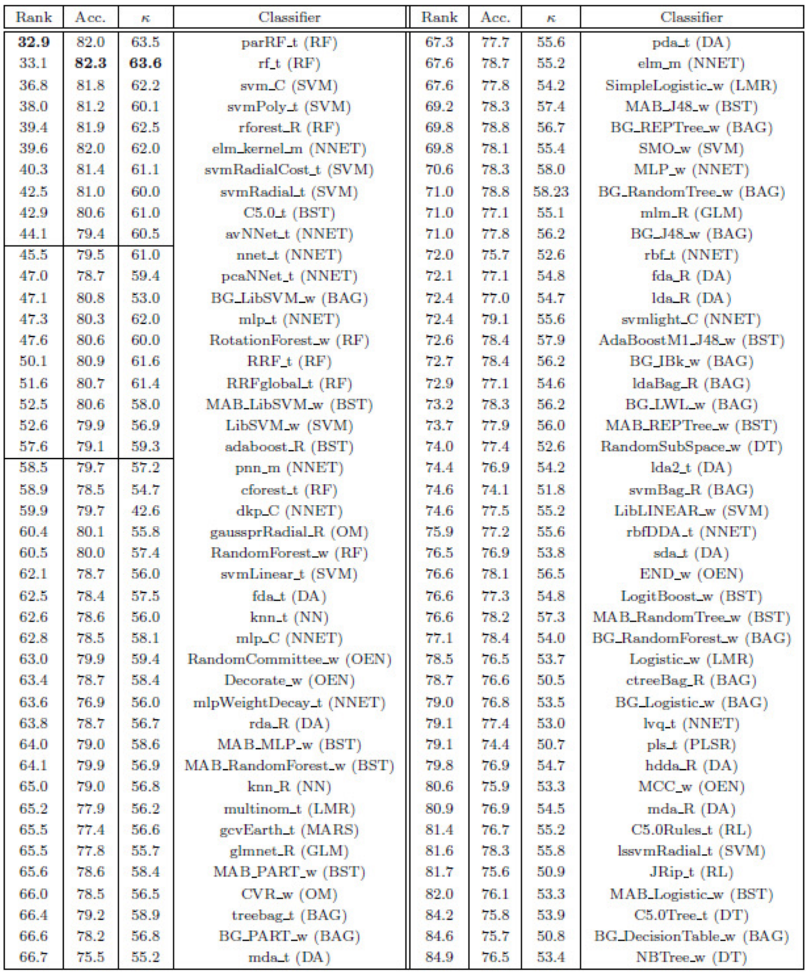
G. Number of forecasted earthquakes per block for each model

Table 29: Number of observed and smoothed earthquakes in each block together with the number of earthquakes forecasted by each model.

Region	Observed	RF	KSVM	KKNN	GLM Net	Training Mean	GLM Top	Depletion Thick	Strain Thick
North	31	35	31	30	23	17	41	56	44
Center West	104	88	60	85	23	16	40	60	61
Center East	51	48	34	47	23	18	33	54	57
South East	8	9	15	8	18	17	21	31	35
South West	58	50	37	45	23	17	34	55	60
All	252	230	177	215	111	85	168	257	257

H. Machine learning models

Given the comparative scarcity of research using datasets of the type we have in this study, we did not commit to a specific algorithm or algorithm family a priori. We opt to empirically test and rank several types of algorithms to determine which model families work well in the given context through a benchmarking study. The candidate algorithms are loosely based on the work of (Delgado, 2014), who have tested 179 different algorithms on datasets from the UCI Machine Learning repository (Bache & Lichman, 2013).



Rank	Acc.	κ	Classifier	Rank	Acc.	κ	Classifier
32.9	82.0	63.5	parRF_t (RF)	67.3	77.7	55.6	pda_t (DA)
33.1	**82.3**	**63.6**	rf_t (RF)	67.6	78.7	55.2	elm_m (NNET)
36.8	81.8	62.2	svm_C (SVM)	67.6	77.8	54.2	SimpleLogistic_w (LMR)
38.0	81.2	60.1	svmPoly_t (SVM)	69.2	78.3	57.4	MAB_J48_w (BST)
39.4	81.9	62.5	rforest_R (RF)	69.8	78.8	56.7	BG_REPTree_w (BAG)
39.6	82.0	62.0	elm_kernel_m (NNET)	69.8	78.1	55.4	SMO_w (SVM)
40.3	81.4	61.1	svmRadialCost_t (SVM)	70.6	78.3	58.0	MLP_w (NNET)
42.5	81.0	60.0	svmRadial_t (SVM)	71.0	78.8	58.23	BG_RandomTree_w (BAG)
42.9	80.6	61.0	C5.0_t (BST)	71.0	77.1	55.1	mlm_R (GLM)
44.1	79.4	60.5	avNNet_t (NNET)	71.0	77.8	56.2	BG_J48_w (BAG)
45.5	79.5	61.0	nnet_t (NNET)	72.0	75.7	52.6	rbf_t (NNET)
47.0	78.7	59.4	pcaNNet_t (NNET)	72.1	77.1	54.8	fda_R (DA)
47.1	80.8	53.0	BG_LibSVM_w (BAG)	72.4	77.0	54.7	lda_R (DA)
47.3	80.3	62.0	mlp_t (NNET)	72.4	79.1	55.6	svmlight_C (NNET)
47.6	80.6	60.0	RotationForest_w (RF)	72.6	78.4	57.9	AdaBoostM1_J48_w (BST)
50.1	80.9	61.6	RRF_t (RF)	72.7	78.4	56.2	BG_IBk_w (BAG)
51.6	80.7	61.4	RRFglobal_t (RF)	72.9	77.1	54.6	ldaBag_R (BAG)
52.5	80.6	58.0	MAB_LibSVM_w (BST)	73.2	78.3	56.2	BG_LWL_w (BAG)
52.6	79.9	56.9	LibSVM_w (SVM)	73.7	77.9	56.0	MAB_REPTree_w (BST)
57.6	79.1	59.3	adaboost_R (BST)	74.0	77.4	52.6	RandomSubSpace_w (DT)
58.5	79.7	57.2	pnn_m (NNET)	74.4	76.9	54.2	lda2_t (DA)
58.9	78.5	54.7	cforest_t (RF)	74.6	74.1	51.8	svmBag_R (BAG)
59.9	79.7	42.6	dkp_C (NNET)	74.6	77.5	55.2	LibLINEAR_w (SVM)
60.4	80.1	55.8	gaussprRadial_R (OM)	75.9	77.2	55.6	rbfDDA_t (NNET)
60.5	80.0	57.4	RandomForest_w (RF)	76.5	76.9	53.8	sda_t (DA)
62.1	78.7	56.0	svmLinear_t (SVM)	76.6	78.1	56.5	END_w (OEN)
62.5	78.4	57.5	fda_t (DA)	76.6	77.3	54.8	LogitBoost_w (BST)
62.6	78.6	56.0	knn_t (NN)	76.6	78.2	57.3	MAB_RandomTree_w (BST)
62.8	78.5	58.1	mlp_C (NNET)	77.1	78.4	54.0	BG_RandomForest_w (BAG)
63.0	79.9	59.4	RandomCommittee_w (OEN)	78.5	76.5	53.7	Logistic_w (LMR)
63.4	78.7	58.4	Decorate_w (OEN)	78.7	76.6	50.5	ctreeBag_R (BAG)
63.6	76.9	56.0	mlpWeightDecay_t (NNET)	79.0	76.8	53.5	BG_Logistic_w (BAG)
63.8	78.7	56.7	rda_R (DA)	79.1	77.4	53.0	lvq_t (NNET)
64.0	79.0	58.6	MAB_MLP_w (BST)	79.1	74.4	50.7	pls_t (PLSR)
64.1	79.9	56.9	MAB_RandomForest_w (BST)	79.8	76.9	54.7	hdda_R (DA)
65.0	79.0	56.8	knn_R (NN)	80.6	75.9	53.3	MCC_w (OEN)
65.2	77.9	56.2	multinom_t (LMR)	80.9	76.9	54.5	mda_R (DA)
65.5	77.4	56.6	gcvEarth_t (MARS)	81.4	76.7	55.2	C5.0Rules_t (RL)
65.5	77.8	55.7	glmnet_R (GLM)	81.6	78.3	55.8	lssvmRadial_t (SVM)
65.6	78.6	58.4	MAB_PART_w (BST)	81.7	75.6	50.9	JRip_t (RL)
66.0	78.5	56.5	CVR_w (OM)	82.0	76.1	53.3	MAB_Logistic_w (BST)
66.4	79.2	58.9	treebag_t (BAG)	84.2	75.8	53.9	C5.0Tree_t (DT)
66.6	78.2	56.8	BG_PART_w (BAG)	84.6	75.7	50.8	BG_DecisionTable_w (BAG)
66.7	75.5	55.2	mda_t (DA)	84.9	76.5	53.4	NBTree_w (DT)

Figure: Overview showing model rank on multiple datasets, reproduced from (Delgado M.F, 2014)

It is important to clarify however that the (Delgado, 2014) study focused on the use of algorithms for classification which differs from the setup of the seismicity study described in these pages. A more recent study by Makridakis (2018) focused on benchmarking machine learning models against classic statistical methods for the task of forecasting, which is in line with the usage of machine learning in this study. We have expanded and elaborated on the benchmarking study in a number of ways both in how the validation of the algorithms is concerned and in comparison with statistical baselines.

Generalized Linear Models

Generalized Linear Models (GLM) are an extension of "classical" ordinary least squares regression (OLS) (Nelder & Wedderburn, 1972). OLS tries to fit the parameter weights for the linear relationship between the features and the target. GLMs extend on this concept by allowing the target to exhibit error distributions that are not normally distributed. In this study, we use GLMS and two GLM variants: (i) GLMnet, a GLM with elastic net regularization and (ii) GLMtop, a GLM model that has been trained using the top 5 most significant features.

The GLMnet (elastic net) algorithm deals with the multicollinearity problem in the original feature space by applying dimensionality reduction. As the ratio between the number of fitted coefficients and number of observations increases, the estimates for the coefficients incur more variance. By applying the bias-variance trade-off, we can choose to introduce a controlled bias in our algorithm, to drastically reduce the variance of the estimates. We can do this by applying a technique called *regularization* (Friedman, Hastie, & Tibshiranie, 2010). Simply put, we can regularize the coefficients of the GLMs by applying a penalty for large components. Whereas a traditional GLM would seek to find the component weights such that a loss-function is minimized, regularized GLM allows for the optimization with regards to a loss function that is a weighted version of the sum of the absolute values of the coefficient weights (L1 norm), or the sum of the squared values of the coefficient weights (L2 norm). The latter technique is called *ridge regression*, while the former is referred to as *LASSO*. Elastic net regularization effectively combines both L1 and L2 types by using an additional alpha parameter to gauge the degree to which either should be implemented.

A GLM Top model with default hyperparameters has been trained only using a selection of the top 5 features. This model might have as an advantage that its performance is not potentially degraded by less well performing features. No regularization has been implemented in this case.

K-Nearest Neighbours

K-Nearest Neighbours (KNN) (Hu, Huang, Ke, & Tsai, 2016) is a simple, yet effective, machine learning algorithm that makes use of the distance (by some chosen metric) between different observations in the dataset. Intuitively, observations that are more similar to each other regarding a subset of the features (the predictors) are increasingly likely to be more similar regarding another subset of the features (the target variable). This idea is formalised in the KNN algorithm, where classifications or predictions for a new (unseen) instance are based on a committee or aggregation of *k* most similar examples.

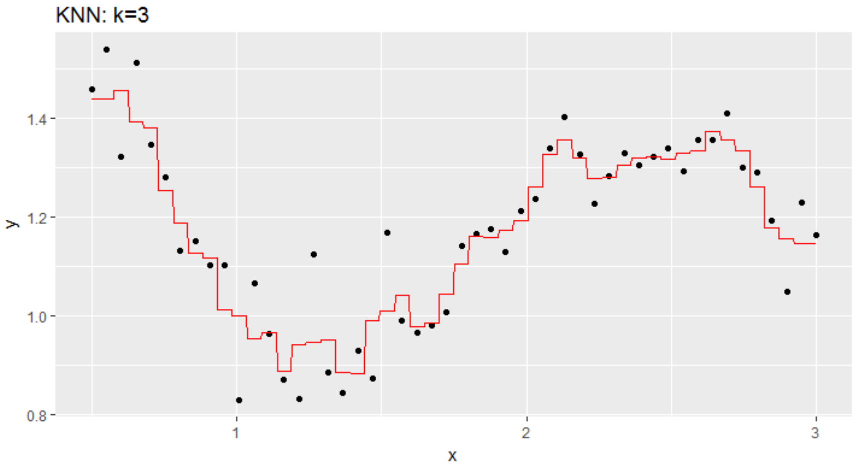


Figure: Example of KNN used for regression, showing the KNN predictions (red), the actual measurements (black dots). Adapted from (Kim, Kim, & Namkoong, 2016)

Random Forests

Random Forests (Breiman, Random Forests, 2001) have been used to outstanding results across a wide variety of tasks. Random Forests, at their core, represent an extension of decision tree algorithms using ensembles. Ensembles refer to a modelling technique where a decision or prediction is not produced by a single algorithm, but rather by a collection of them (Schapire, 1990). The use of this meta-modelling technique is not explicitly limited to Random Forests, but can be applied to any base algorithm, or even collection of diverse algorithms.

The units within a Random Forest are known as Decision trees (Breiman, Friedman, Olshen, & Stone, 1984), renowned for their simplicity, clarity, and speed. These trees perform a recursive partitioning of the data space with the objective of making the data partitions as homogeneous as possible. In this study, we use binary trees though other partitions are possible.

Decision trees do suffer from high variance. It is for this reason that we aim to decrease the variance of the solution by creating an ensemble of trees (Breiman, Friedman, Olshen, & Stone, 1984), rather than look at a single tree. Random Forests achieve this decorrelation in two ways: bootstrap sampling and restricting the set of candidates features to split on.

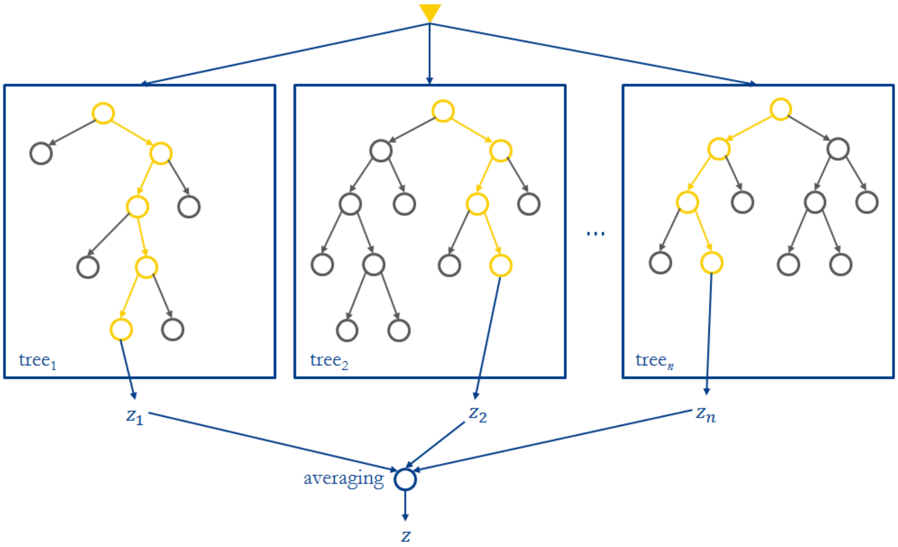


Figure: Illustrative example of how regression predictions of individual trees combine in a random forest through averaging of the prediction results of individual trees.

SVR/KSVM

SVR’s are non-probabilistic algorithms which can be considered extensions and generalizations of optimal separating hyperplanes that get defined by the data points closest to the decision boundary, which are referred to as support vectors (Cortes & Vapnik, 1995). SVR’s extend on the concept of optimal separating hyperplanes in two ways:

- By accommodating the case of overlapping classes.
- By allowing nonlinear decision boundaries in the original feature space by employing the kernel trick.

In SVR’s, the input is implicitly mapped onto an m-dimensional (where m can, in fact, be infinite, and in those cases not computable) feature space using some fixed (nonlinear) mapping (kernel trick), and then a linear model is constructed in this feature space (Friedman, Tibshirani, & Hastie, 2009). The main motivation is to seek and optimize the generalization bounds given for the regression. These bounds rely on defining the loss function that ignores errors situated within a certain distance of the true value. In other words, the goal is to find a function whose prediction deviates from the target value by an amount no more than ε.

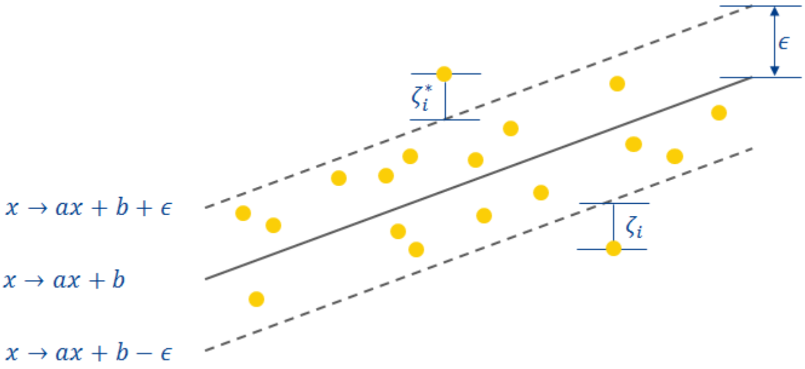


Figure: Example of KSVM regression. Points within the pink band, where the prediction error < epsilon, don't contribute to the total loss of the function. Outside of this band are the support points that determine the parameters of the functions.

I. Derivation of Expected Value and Variance for Log Likelihood Ratio

In this section, we detail the derivation of the equations given in Section 6 which relate to the expected value and variance of the log of the likelihood ratio. This ratio is calculated using realisations from the simulation model, $\mathcal{M}_S$. Under this model, we simulate a vector of counts, $\tilde{\boldsymbol{\lambda}}$, which contains values for each spacetime location. The elements of $\tilde{\boldsymbol{\lambda}}$ are sampled independently from Poisson distributions such that, $\tilde{\lambda}_{\boldsymbol{iz}} \boldsymbol{\sim Poisson}(\theta_{iz})$. Therefore $\mathrm{E}[\tilde{\lambda}_{\boldsymbol{iz}}] = \theta_{iz}$ and $\mathrm{var}[\tilde{\lambda}_{iz}] = \theta_{iz}$. The likelihood we are using is also the Poisson likelihood such that,

$$\ell(\mathcal{M}_l|\tilde{\boldsymbol{\lambda}}) = \sum_{\boldsymbol{iz}} \tilde{\lambda}_{\boldsymbol{iz}}\, log\left(p_{iz}^{(l)}\right) - p_{iz}^{(l)} - log(\tilde{\lambda}_{iz}!).$$

Therefore, the log of the likelihood ratio is,

$$\begin{aligned} r(\mathcal{M}_i, \mathcal{M}_0|\tilde{\boldsymbol{\lambda}}) &= \ell(\mathcal{M}_l|\tilde{\boldsymbol{\lambda}}) - \ell(\mathcal{M}_0|\tilde{\boldsymbol{\lambda}}), \\ &= \sum_{\boldsymbol{iz}} \left[\tilde{\lambda}_{\boldsymbol{iz}}\, log\left(p_{iz}^{(l)}\right) - p_{iz}^{(l)} - \tilde{\lambda}_{iz}\, log\left(p_{iz}^{(0)}\right) + p_{iz}^{(0)}\right], \\ &= \sum_{\boldsymbol{iz}} \left[\tilde{\lambda}_{\boldsymbol{iz}} \left\{log\left(p_{iz}^{(l)}\right) - log\left(p_{iz}^{(0)}\right)\right\} - p_{iz}^{(l)} + p_{iz}^{(0)}\right]. \end{aligned}$$

We first consider the expected value of the log ratio,

$$\begin{aligned} E[r(\mathcal{M}_i, \mathcal{M}_0|\tilde{\boldsymbol{\lambda}})] &= \sum_{\boldsymbol{iz}} \left[\boldsymbol{E}[\tilde{\lambda}_{iz}] \left\{log\left(p_{iz}^{(l)}\right) - log\left(p_{iz}^{(0)}\right)\right\} - p_{iz}^{(l)} + p_{iz}^{(0)}\right] \\ &= \sum_{iz} \left[\theta_{iz} \left\{log\left(p_{iz}^{(l)} - p_{iz}^{(0)}\right)\right\} - p_{iz}^{(l)} + p_{iz}^{(0)}\right]. \end{aligned}$$

We can then consider the variance of the log ratio. Here we make use of the fact that the elements of $\tilde{\boldsymbol{\lambda}}$ are independent.

$$\begin{aligned} \mathrm{var}[r(\mathcal{M}_i, \mathcal{M}_0|\tilde{\boldsymbol{\lambda}})] &= \mathrm{var}\left[\sum_{\boldsymbol{iz}} \left[\tilde{\lambda}_{iz} \left\{log\left(p_{iz}^{(l)}\right) - log\left(p_{iz}^{(0)}\right)\right\} - p_{iz}^{(l)} + p_{iz}^{(0)}\right]\right], \\ &= \sum_{iz} \left[\mathrm{var}[\tilde{\lambda}_{\boldsymbol{iz}}] \left\{log\left(p_{iz}^{(l)} - p_{iz}^{(0)}\right)\right\}^2\right], \\ &= \sum_{iz} \left[\theta_{iz} \left\{log\left(p_{iz}^{(l)} - p_{iz}^{(0)}\right)\right\}^2\right]. \end{aligned}$$

J. Varying bandwidth between 2000m – 4000m.

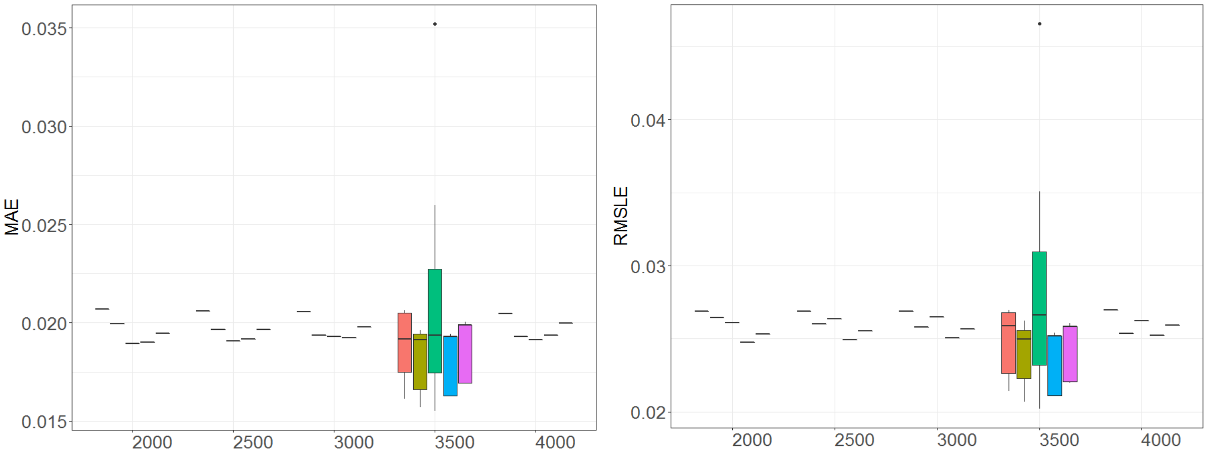


Figure 58 Change in MAE (right) and RMSLE (left) when varying predictor feature bandwidth between 2000m - 4000m in steps of 500 m.

Glossary

Physical quantities are denoted by the following variables:

- P Pressure in the reservoir (bar);
- S Subsidence (m);
- C Compaction (m);
- HCT Hydrocarbon Column Thickness (m);
- HCM Hydrocarbon Column Mass (kg)

Machine learning concepts used:

- **Target**: the (pre-processed) variable we like to predict. For example, the target used throughout this study is the rate of earthquakes per day of $M \geq 1.5$ per grid cell over a time interval.
- **Feature**: a (pre-processed) source data variable which might be a predictor for the target. Note that one source data variable might give rise to multiple features or the other way around. For example, in this study features are: the mean pressure in a cell per month, the mean pressure decreases in a cell per month, etc. Confusingly, sometimes features are also called covariates or plainly variables.
- **Covariates**: see features.
- **Machine Learning model**: a formula or association rule associating feature values to target values. Usually, machine learning models are complex and do not provide intuitive insights.
- **Hold-out set**: a portion of the data which is not used for training or testing of the ML models. Throughout the report, the period 2013-2016 is used as a hold-out set.

Special terminology relating to this study:

- **Non-extrapolating models**: These are models such as Random Forest (RF) or K-nearest neighbours (KNN). That will **not** be able to forecast a value for the Target variable that is outside of the range of values that have been observed for the Target during the training period. E.g., if the range for the number of earthquakes observed during training falls within 1 and 10, these models will only be able to forecast values for the Target within 1 and 10.
- **Extrapolating models**: These are models such as Support Vector Regression (KSVM) or generalized linear models (GLM) that could forecast a value for the Target variable that is outside of the range of values that have been observed for the Target during the training period. E.g., if the range for the number of earthquakes observed during training falls within 1 and 10, these models could, given the right input and choice of kernel/link function, forecast a value outside of this range such as 13 or 0.
- **Ultimate-state constrains:** These are additional values given as training features that relate to the definite state of the field as calculated by MoReS. E.g., in 2100, after production has long stopped, and the field has reached an ultimate state (leaving aside processes that operate on geological time scales), the expected weighted average pressure across the field will range between 94 bar (if production stops in 2019) and 37 bar (if production continues until maximum depletion). These extreme values act as definite states of the field.
- **Earthquake rate:** Number of earthquakes that take place within a period of 3 months

Mathematical notations and definitions used unless otherwise specified:

- $\mathbb{N}$ denotes the set of natural numbers;
- $\mathbb{R}$ denotes the set of real numbers;
- $[a, b]$ is the interval between a and b, including boundary values;
- (a, b) is the interval between a and b, excluding boundary values;
- m is the number of features/variables/covariates in a model;
- n is the number of data points that are available for each features/variables/covariate
- $x_i \in \mathbb{R}^m$ is the vector of features/variables/covariates of a model at time interval i;
- t_i is the target value at time interval i;
- $d_i = (x_i, t_i)$ denotes a data point d at time interval i, consisting of the m features/variables/covariates and the target;
- f denotes a model or association rule
- $p_i = f(x_i)$ is the prediction of model f based on features $x_i \in \mathbb{R}^m$ on time i

Geospatial coordinate systems mentioned:

- **RD**: the Netherlands triangulation system [Rijksdriehoekstelsel] is a coordinate system used at the national level in the European part of the Netherlands. It has two perpendicular coordinates x and y. For details and transformations to other coordinate systems see (Kadaster, 2018). The transformations to and from the Latitude/Longitude and the RD coordinate systems have been done using the "proj4" and "sp" R packages which have special functionalities for the transformation and application of different cartographic projections. For more details see the package descriptions (Urbanek, 2015).

Key abbreviations used throughout the text:

CV	Cross validation;
GFO	Groningen field Outline;
GLM	Generalized Linear Model, a machine learning model;
ICE	Individual Conditional Expectations;
I.i.d.	Independently and Identically Distributed;
KNN	K-Nearest Neighbours, a machine learning model;
MAE	Mean Absolute Error, an error metric;
ML	Machine Learning;
M_c	Magnitude of Completeness;
M_{min}	Minimum Magnitude;
MMP	Model and Meta Parameter combination;
MoReS	Dynamic Reservoir Model;
PSHRA	Probabilistic Seismic Hazard and Risk Assessment;
R	Statistical Computing Environment;
RF	Random Forest, a machine learning model;
RMSLE	Root Mean Square Logarithmic Error, an error metric;
SE	Standard Error;
(K)SVM	Support Vector Model, a machine learning model;
SVR	Support Vector Regression, a KSVM used for regression.

Tooling

In order to facilitate the setup of the forecasting experiments, we make use of the R package MLR which is described in more detail in (Bischl et al., 2016). By using the MLR package as our general framework for the benchmarking experiments, we avoid duplication of code and potential bugs in critical parts of the code related to estimating model performance. The MLR package provides a modular interface to the following common tasks in the context of machine learning workflows and benchmarking experiments:

- Common pre-processing routines as removal of constant or duplicated columns, and normalization of the covariate columns. MLR implements common scaling techniques like standardization but also provides access to more advanced pre-processing routines like those exposed through a wrapper for the pre-processing routines offered by the Caret package (Kuhn M., 2017). Those include Box-Cox transformations, Yeo-Johnson transformations and also a set of different imputation techniques.
- Feature selection, based on different criteria (e.g. correlation, RF variable importance, …).
- Definition of a sub-sampling strategy for the hold-out set prediction experiments. For instance, cross-validation and different versions of the bootstrap. Note that the walk forward resampling strategy is not implemented in MLR
≤ 2.11, which is why we had to implement it manually.
- Definition of several common error measures like MAE, RMSE, Kendall-Tau and
R_2
- A convenient interface to around 80 machine learning algorithms for both regression and classification. Those include (regularized) linear models, Support Vector Regression, tree-based methods and different flavours of neural networks.
- Support for automated parameter tuning of those algorithms using different tuning strategies like basic grid searches bot also more advanced gradient-based techniques.
- Setup and results reporting of benchmarking experiments for several techniques.

For feature selection and for the purpose of detecting "significant" features with respect to the selected prediction target, we make use of the Boruta package (Kursa, 2010). This record of significant and potentially significant variables is stored for every prediction experiment. It is a heuristic procedure that uses the variable importance measure calculated by implementations of the random forest algorithm. As such, non-linear effects and interactions between parameters are taken into account. In order to counter effects related to the multiplicity of noise variables, the algorithm is iterative in nature.

Acknowledgements

The authors are indebted to Jan van Elk (NAM) and Taco den Bezemer (NAM) for their strong and continued support for this study – without their enthusiasm and trust, this study would not be here. We are indebted to the lead reviewers for insightful discussions throughout this study, much of which has been incorporated in this study. In alphabetic order:

- Stijn Bierman (Shell P&T);
- Stephen Bourne (Shell P&T);
- Franz Király (University College London);

This study has been reviewed within a larger audience at successive stages of its maturity. We owe a big "Thank You" to the extended review team for their constructive feedback. In alphabetic order:

- Peter van den Bogert (Shell P&T);
- Xander Campman (Shell P&T);
- Pandu Devarakota (Shell P&T);
- Munish Goyal (IBM Services)
- Kees Hindriks (Shell P&T)
- Stephen Lord (IBM Services);
- Roger Yuan (Shell P&T);
- Rick Wentinck (Shell P&T);
- Mo Zhang (IBM Services).

This study builds on the data provided by specialists. We would like to thank (alphabetically):

- Hermann Baehr (NAM), for providing subsidence data;
- Stijn Bierman (Shell P&T), for sharing his subsidence and compaction interpolations;
- Leendert Geurtsen (NAM), Per Valvatne (NAM) and Assaf Mar-Or (NAM) for providing both the dynamic reservoir data from MoReS and the production forecasts;
- Gerard Joosten (Shell P&T) for his support in exporting the HCT, HCM and related properties from MoReS;
- Marloes Kortekaas (EBN) for providing the Ant-tracked seismic volume attribute used in several EBN publications;
- Remco Romijn (NAM), for providing exported seismic attributes from the Groningen 3-D seismic surveys, and for providing additional steer and documentation regarding attribute selection;
- Richard Vietje (NAM), for providing historical production data;
- Clements Visser (NAM), for providing the fault data and estimates for the Groningen field and the Petrel model with static reservoir attributes;
- Onno van der Wal (NAM), for providing compaction and subsidence data;
- Rick Wentinck (Shell P&T) for providing the overburden stress distribution from the COMSOL geomechanical model of the Groningen field;
- Alan Wood (Shell P&T), for sharing a recent version of the Petrel model.

Finally, we would like to thank Paul Gelderblom (Shell P&T), Maggie Wenham (Shell P&T), Harry van der Burg (Shell SITI), Mando Rotman (IBM Services), Jonito Douwes Dekker (IBM Services) and Phaedra Kortekaas (IBM Services) for their organizational support in realizing this study.

References

Asencio-Cortés, G., Martínez-Álvarez, F., Morales-Esteban, A., Reyes, J., 2016. A sensitivity study of seismicity indicators in supervised learning to improve earthquake prediction. Knowledge-Based Systems. https://doi.org/10.1016/j.knosys.2016.02.014

Avis, D., Bremner, D., Seidel, R., 1997. How good are convex hull algorithms? Computational Geometry: Theory and Applications. https://doi.org/10.1016/S0925-7721(96)00023-5

Balas, E., 2010. Disjunctive programming. 50 Years of Integer Programming 1958-2008: From the Early Years to the State-of-the-Art. https://doi.org/10.1007/978-3-540-68279-0_10

Barber, C.B., Dobkin, D.P., Huhdanpaa, H., 1996. The quickhull algorithm for convex hulls. ACM Transactions on Mathematical Software. https://doi.org/10.1145/235815.235821

Barton, C.A., Zoback, M.D., Moos, D., 1995. Fluid flow along potentially active faults in crystalline rock. Geology 23, 683. https://doi.org/10.1130/0091-7613(1995)023<0683:FFAPAF>2.3.CO;2

Bence, J.R., 1995. Analysis of short time series: Correcting for autocorrelation. Ecology. https://doi.org/10.2307/1941218

Bierman, S., Kraaijeveld, F., Bourne, S., 2015. Regularised direct inversion to compaction in the Groningen reservoir using measurements from optical leveling campaigns. Technical Report. Shell Global Solutions International.

Bourne, S.J., Oates, S.J., 2017. Extreme Threshold Failures Within a Heterogeneous Elastic Thin Sheet and the Spatial-Temporal Development of Induced Seismicity Within the Groningen Gas Field. Journal of Geophysical Research: Solid Earth 299–320. https://doi.org/10.1002/2017JB014356

Bourne, S.J., Oates, S.J., 2015. An activity rate model of induced seismicity within the Groningen Field (Part 2). 1–53.

Bray, A., Schoenberg, F.P., 2013. Assessment of Point Process Models for Earthquake Forecasting. Statistical Science. https://doi.org/10.1214/13-STS440

Buttinelli, M., Improta, L., Bagh, S., Chiarabba, C., Keranen, K., Savage, H., Abers, G., Cochran, E., Rubinstein, J.L., Ellsworth, W.L., Garr, A.M., Benz, H.M., Ellsworth, W.L., Evans, D., Healy, J.H., Rubey, W., Griggs, D., Raleigh, C.B., Raleigh, C.B., Healy, J.H., Bredehoeft, J.D., Zoback, M.D., Healy, J.H., Giardini, D., Clarke, H., Eisner, L., Styles, P., Turner, P., Edwards, B., Kraft, T., Cauzzi, C., Kästli, P., Wiemer, S., Stabile, T.A., Giocoli, A., Perrone, A., Piscitelli, S., Lapenna, V., Improta, L., Valoroso, L., Piccinini, D., Chiarabba, C., Patacca, E., Scandone, P., Boriani, A., Bonafede, M., Piccardo, G.B., Vai, G.B., Noguera, A.M., Rea, G., Mazzoli, S., Shiner, P., Beccacini, A., Mazzoli, S., Ferranti, L., Cello, G., Tondi, E., Micarelli, L., Mattioni, L., Maschio, L., Ferranti, L., Burrato, P., Improta, L., Gori, P. De, Chiarabba, C., Burrato, P., Valensise, G., Valoroso, L., Valoroso, L., Improta, L., Gori, P. De, Chiarabba, C., Candela, S., Mazzoli, S., Megna, A., Santini, S., Cucci, L., Pondrelli, S., Frepoli, A., Mariucci, M.T., Moro, M., Pastori, M., Trippetta, F., Collettini, C., Vinciguerra, S., Meredith, P., Horton, S., Goebel, T.H.W., Juanes, R., 2016. Inversion of inherited thrusts by wastewater injection induced seismicity at the Val d'Agri oilfield (Italy). Scientific Reports 6, 37165. https://doi.org/10.1038/srep37165

Candela, T., Wassing, B., Heege, J. ter, Buijze, L., 2018. How earthquakes are induced. Science 360, 598–600. https://doi.org/10.1126/science.aat2776

Carrasquilla, J., Melko, R.G., 2017. Machine learning phases of matter. Nature Physics. https://doi.org/10.1038/nphys4035

CGAL Project, 2018. CGAL User and Reference Manual.

Chandola, V., Banerjee, A., Kumar, V., 2009. Anomaly detection. ACM Computing Surveys 41, 1–58. https://doi.org/10.1145/1541880.1541882

Chopra, S., Marfurt, K.J., 2007. Volumetric curvature attributes for fault / fracture characterization. First Break 25, 35–46.

Chupeau, M., Bénichou, O., Majumdar, S.N., 2015. Convex hull of a Brownian motion in confinement. Physical Review E - Statistical, Nonlinear, and Soft Matter Physics. https://doi.org/10.1103/PhysRevE.91.050104

CSEP, 2018. Collaboratory for the STudy of Earthquake Predictability [WWW Document]. URL http://cseptesting.org/ (accessed 12.1.18).

Dershowitz, W.S., Herda, H.H., 1992. Interpretation of fracture spacing and intensity. The 33th US Symposium on Rock Mechanics. Balkema, Rotterdam, Sante Fe, New Mexico, 757–766.

DeVries, P.M.R., Viégas, F., Wattenberg, M., Meade, B.J., 2018. Deep learning of aftershock patterns following large earthquakes., Nature. https://doi.org/10.1038/s41586-018-0438-y

Dieterich, J.H., 2007. Applications of Rate- and State-Dependent Friction to Models of Fault Slip and Earthquake Occurrence. Treatise on Geophysics. Elsevier, 107–129. https://doi.org/10.1016/B978-044452748-6.00065-1

Dost, B., Goutbeek, F., van Eck, T., Kraaijpoel, D., 2012. Monitoring induced seismicity in the North of the Netherlands: status report 2010. De Bilt, Netherlands.

Dost, B., Haak, H.W., 2002. A comprehensive description of the KNMI seismological instrumentation. De Bilt, Netherlands.

Dost, B., Ruigrok, E., Spetzler, J., 2017. Development of seismicity and probabilistic hazard assessment for the Groningen gas field. Geologie En Mijnbouw/Netherlands Journal of Geosciences. https://doi.org/10.1017/njg.2017.20

Efron, B., 1965. The Convex Hull of a Random Set of Points. Biometrika. https://doi.org/10.1093/biomet/52.3-4.331

Efron, B., Stein, C., 1981. The Jackknife Estimate of Variance. The Annals of Statistics. https://doi.org/10.1214/aos/1176345462

Gärtner, B., Schönherr, S., 2000. An efficient, exact, and generic quadratic programming solver for geometric optimization. Proceedings of the Sixteenth Annual Symposium on Computational Geometry - SCG ’00. ACM Press, New York, New York, USA, 110–118. https://doi.org/10.1145/336154.336191

Gerstenberger, M., Rhoades, D., Stirlin, M., Brownrigg, R., Christophersen, A., 2009. Continued Development of the New Zealand Earthquake Forecast Testing Centre.

Getz, W.M., Wilmers, C.C., 2004. A local nearest-neighbor convex-hull construction of home ranges and utilization distributions. Ecography. https://doi.org/10.1111/j.0906-7590.2004.03835.x

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning (2nd edition), Elements. https://doi.org/10.1007/978-0-387-84858-7

Jordan, M.I., Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects., Science. https://doi.org/10.1126/science.aaa8415

Karpatne, A., Atluri, G., Faghmous, J.H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., Kumar, V., 2017. Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data. IEEE Transactions on Knowledge and Data Engineering 29, 2318–2331. https://doi.org/10.1109/TKDE.2017.2720168

Kettermann, M., Abe, S., Raith, A.F., De Jager, J., Urai, J.L., 2017. The effect of salt in dilatant faults on rates and magnitudes of induced seismicity - First results building on the geological setting of the Groningen Rotliegend reservoirs. Geologie En Mijnbouw/Netherlands Journal of Geosciences 96, s87–s104. https://doi.org/10.1017/njg.2017.19

Kirkpatrick, J.D., Bezerra, F.H.R., Shipton, Z.K., Do Nascimento, A.F., Pytharouli, S.I., Lunn, R.J., Soden, A.M., 2013. Scale-dependent influence of pre-existing basement shear zones on rift faulting: a case study from NE Brazil. Journal of the Geological Society 170, 237–247. https://doi.org/10.1144/jgs2012-043

Kneale, C., Brown, S.D., 2018. Small moving window calibration models for soft sensing processes with limited history. Chemometrics and Intelligent Laboratory Systems. https://doi.org/10.1016/j.chemolab.2018.10.007

Kortekaas, M., Jaarsma, B., 2017. Improved definition of faults in the Groningen field using seismic attributes. Geologie En Mijnbouw/Netherlands Journal of Geosciences 96, s71–s85. https://doi.org/10.1017/njg.2017.24

Kraaijpoel, D., Caccavale, M., van Eck, T., Dost, B., 2015. PSHA for seismicity induced by gas extraction in the Groningen Field. Schatzalp Workshop Induced Seismicity. KNMI.

Langley, P., 1988. Machine Learning as an Experimental Science., Machine Learning. https://doi.org/10.1023/A:1022623814640

Last, M., Rabinowitz, N., Leonard, G., 2016. Predicting the maximum earthquake magnitude from seismic data in Israel and its neighboring countries. PLoS ONE 11, 1–16. https://doi.org/10.1371/journal.pone.0146101

Limbeck, J., Lanz, F., Barbaro, E., Harris, C., Bisdom, K., Park, T., Oosterbosch, W., Jamali-Rad, H., Nevenzeel, K., 2018. Evaluation of a Machine Learning methodology to forecast induced seismicity event rates within the Groningen Field. Assen, Netherlands.

Liu, S., Yamada, M., Collier, N., Sugiyama, M., 2013. Change-point detection in time-series data by relative density-ratio estimation. Neural Networks. https://doi.org/10.1016/j.neunet.2013.01.012

Majumdar, S.N., Comtet, A., Randon-Furling, J., 2010. Random Convex Hulls and Extreme Value Statistics. Journal of Statistical Physics. https://doi.org/10.1007/s10955-009-9905-z

Markou, M., Singh, S., 2003. Novelty detection: A review - Part 1: Statistical approaches., Signal Processing. https://doi.org/10.1016/j.sigpro.2003.07.018

Marriott, F.H.C. (Francis H.C., Kendall, M.G. (Maurice G., International Statistical Institute., 1990. A dictionary of statistical terms, Fifth edition. ed. Published for the International Statistical Institute by Longman Scientific & Technical, Burnt Mill Harlow Essex England ;New York.

Marzocchi, W., Zechar, J.D., 2011. Earthquake Forecasting and Earthquake Prediction: Different Approaches for Obtaining the Best Model. Seismological Research Letters. https://doi.org/10.1785/gssrl.82.3.442

Melnikov, A.A., Nautrup, H.P., Krenn, M., Dunjko, V., Tiersch, M., Zeilinger, A., Briegel, H.J., 2018. Active learning machine learns to create new quantum experiments. Proceedings of the National Academy of Sciences 115, 1221–1226. https://doi.org/10.1073/PNAS.1714936115

Ministry of Economic Affairs and Climate Policy, 2018. Kamerbrief over gaswinning Groningen.

Nederlandse Aardolie Maatschappij, 2017. Optimisation of the Production Distribution over the Groningen field to reduce Seismicity. Assen, Netherlands.

Nederlandse Aardolie Maatschappij, 2016a. Study and Data Acquisition Plan Induced Seismicity in Groningen - Update Post-Winningsplan 2016. Assen, Netherlands.

Nederlandse Aardolie Maatschappij, 2016b. Gaswinning Groningen Meet- en Regelprotocol Aardbevingen. Assen, Netherlands.

Nederlandse Aardolie Maatschappij, 2016c. Winningsplan Groningen Gasveld 2016. Assen, Netherlands.

Nederlandse Aardolie Maatschappij, 2016d. Technical Addendum to the Winningsplan Groningen 2016.

Panakkat, A., Adeli, H., 2009. Recurrent neural network for approximate earthquake time and location prediction using multiple seismicity indicators. Computer-Aided Civil and Infrastructure Engineering. https://doi.org/10.1111/j.1467-8667.2009.00595.x

Pathak, J., Hunt, B., Girvan, M., Lu, Z., Ott, E., 2018. Model-Free Prediction of Large Spatiotemporally Chaotic Systems from Data: A Reservoir Computing Approach. Physical Review Letters. https://doi.org/10.1103/PhysRevLett.120.024102

Perol, T., Gharbi, M., Denolle, M., 2017. Convolutional Neural Network for Earthquake Detection and Location. Science Advances. https://doi.org/10.1126/sciadv.1700578

Pimentel, M.A.F., Clifton, D.A., Clifton, L., Tarassenko, L., 2014. A review of novelty detection., Signal Processing. https://doi.org/10.1016/j.sigpro.2013.12.026

Ramirez, J., Meyer, F.G., 2011. Machine learning for seismic signal processing: Seismic phase classification on a manifold. Proceedings - 10th International Conference on Machine Learning and Applications, ICMLA 2011. https://doi.org/10.1109/ICMLA.2011.91

Rhoades, D.A., Schorlemmer, D., Gerstenberger, M.C., Christophersen, A., Zechar, J.D., Imoto, M., 2011. Efficient testing of earthquake forecasting models. Acta Geophysica. https://doi.org/10.2478/s11600-011-0013-5

Romijn, R., 2017. Groningen Velocity Model 2017 - Groningen full elastic velocity model September 2017.

Rouet-Leduc, B., Hulbert, C., Lubbers, N., Barros, K., Humphreys, C.J., Johnson, P.A., 2017. Machine Learning Predicts Laboratory Earthquakes. Geophysical Research Letters 44, 9276–9282. https://doi.org/10.1002/2017GL074677

Schorlemmer, D., Gerstenberger, M.C., Wiemer, S., Jackson, D.D., Rhoades, D.A., 2007. Earthquake Likelihood Model Testing. Seismological Research Letters 78, 17–29. https://doi.org/10.1785/gssrl.78.1.17

Spetzler, J., Dost, B., 2017. Hypocentre estimation of induced earthquakes in Groningen. Geophysical Journal International. https://doi.org/10.1093/gji/ggx020

Stout, M., Bacardit, J., Hirst, J.D., Krasnogor, N., 2008. Prediction of recursive convex hull class assignments for protein residues. Bioinformatics. https://doi.org/10.1093/bioinformatics/btn050

TNO, 2017. Groningen Gas field [WWW Document], NLOG. URL https://www.nlog.nl/en/groningen-gasfield (accessed 12.19.17).

Urbanek, S., 2015. A simple interface to the PROJ.4 cartographic projections library [WWW Document]. URL http://www.rforge.net/proj4/ (accessed 12.1.17).

Van De Weygaert, R., Vegter, G., Edelsbrunner, H., Jones, B.J.T., Pranav, P., Park, C., Hellwing, W.A., Eldering, B., Kruithof, N., Bos, E.G.P., Hidding, J., Feldbrugge, J., Ten Have, E., Van Engelen, M., Caroli, M., Teillaud, M., 2011. Alpha, Betti and the Megaparsec universe: On the topology of the cosmic web. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/978-3-642-25249-5_3

van Gent, H.W., Back, S., Urai, J.L., Kukla, P.A., Reicherter, K., 2009. Paleostresses of the Groningen area, the Netherlands-Results of a seismic based structural reconstruction. Tectonophysics. https://doi.org/10.1016/j.tecto.2008.09.038

Van Oeveren, H., Valvatne, P., Geurtsen, L., Van Elk, J., 2017. History match of the Groningen field dynamic reservoir model to subsidence data and conventional subsurface data. Geologie En Mijnbouw/Netherlands Journal of Geosciences.

https://doi.org/10.1017/njg.2017.26

Wagner, N., Rondinelli, J.M., 2016. Theory-Guided Machine Learning in Materials Science. Frontiers in Materials 3. https://doi.org/10.3389/fmats.2016.00028

Wervelman, R., 2015. Groningen Pre-SDM - external report.

Wolsey, L.A., Yaman, H., 2018. Convex hull results for the warehouse problem. Discrete Optimization. https://doi.org/10.1016/j.disopt.2018.06.002

Wood, S.N., Pya, N., Säfken, B., 2016. Smoothing Parameter and Model Selection for General Smooth Models. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.2016.1180986

Wright, M.N., Ziegler, A., 2015. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. https://doi.org/10.18637/jss.v077.i01

Yilmaz, Ö., 2001. Seismic Data Analysis. Society of Exploration Geophysicists. https://doi.org/10.1190/1.9781560801580

Zoback, M.D., 2007. Reservoir Geomechanics, Reservoir Geomechanics. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511586477

Bibliographic information

Title	Evaluation of a Machine Learning methodology for spatio-temporal induced seismicity forecasts within the Groningen field
Author(s)	F. Lanz (IBM Services) K. Bisdom (GSNL-PTX/S/RM) E. Barbaro (IBM Services) J. Limbeck (GSNL-PTX/D/S) T. Park (GSNL-PTX/D/S) C. Harris (SUKEP-UPO/W/T) K. Nevenzeel (IBM Services)
Keywords	Seismicity, earthquakes, forecasting, event rate, activity rate, spatio-temporal, analytics, data science, machine learning, NAM, Groningen
Date of Issue	March 2019
Period of Work	June 2017 to February 2019
US Export Control	Non US - Non Controlled
Issuing Company	Nederlandse Aardolie Maatschappij B.V. Upstream International Schepersmaat 2 9405 TA Assen The Netherlands