Computational Geosciences (2021) 25:529–551
https://doi.org/10.1007/s10596-020-10023-0

ORIGINAL PAPER

Using machine learning for model benchmarking and forecasting of depletion-induced seismicity in the Groningen gas field

Jan Limbeck[1] · Kevin Bisdom[1] · Fabian Lanz[2] · Timothy Park[1] · Eduardo Barbaro[2] · Stephen Bourne[1] · Franz Kiraly[1,3,4] · Stijn Bierman[1] · Chris Harris[1] · Keimpe Nevenzeel[1,2] · Taco den Bezemer[5] · Jan van Elk[5]

Received: 7 February 2020 / Accepted: 5 December 2020 / Published online: 3 January 2021

Abstract
The Groningen gas field in the Netherlands is experiencing induced seismicity as a result of ongoing depletion. The physical mechanisms that control seismicity have been studied through rock mechanical experiments and combined physical-statistical models to support development of a framework to forecast induced-seismicity risks. To investigate whether machine learning techniques such as Random Forests and Support Vector Machines bring new insights into forecasts of induced seismicity rates in space and time, a pipeline is designed that extends time-series analysis methods to a spatiotemporal framework with a factorial setup, which allows probing a large parameter space of plausible modelling assumptions, followed by a statistical meta-analysis to account for the intrinsic uncertainties in subsurface data and to ensure statistical significance and robustness of results. The pipeline includes model validation using e.g. likelihood ratio tests against average depletion thickness and strain thickness baselines to establish whether the models have statistically significant forecasting power. The methodology is applied to forecast seismicity for two distinctly different gas production scenarios. Results show that seismicity forecasts generated using Support Vector Machines significantly outperform beforementioned baselines. Forecasts from the method hint at decreasing seismicity rates within the next 5 years, in a conservative production scenario, and no such decrease in a higher depletion scenario, although due to the small effective sample size no statistically solid statement of this kind can be made. The presented approach can be used to make forecasts beyond the investigated 5-years period, although this requires addition of limited physics-based constraints to avoid unphysical forecasts.

Keywords Seismicity forecasting · Groningen gas field · Machine learning · Model benchmarking · Depletion-induced seismicity · Geomechanics · Earthquakes

1 Introduction

1.1 Context and background: Gas production induced seismicity

The Groningen field is the largest gas field in Europe and one of the largest gas fields in the world, with approximately 2900 billion m3 gas originally in place [1, 2]. Production from this field contributed significantly to the Dutch economy in the past 60 years [3]. Production of hydrocarbons can potentially lead to induced seismicity [4, 5]. In the depleting Groningen field, the high-porosity faulted sandstone reservoir has experienced several decimetres of compaction [6, 7]. Following depletion, induced seismic events have started occurring

✉ Kevin Bisdom
kevin.bisdom@shell.com

1 Shell Global Solutions International B.V, Grasweg 31, 1031 HW Amsterdam, The Netherlands

2 IBM Services Netherlands, Johan Huizingalaan 765, 1066 VH Amsterdam, the Netherlands

3 University College London, Gower Street, London WC1E 6BT, UK

4 The Alan Turing Institute, 96 Euston Rd, Kings Cross, London NW1 2DB, UK

5 Nederlandse Aardolie Maatschappij, Schepersmaat 2, 9405 TA Assen, The Netherlands

within the field boundary (Fig. 1) [1]. From 2005 onwards, the frequency of induced seismicity per volume of gas produced increased with further depletion, with the largest event (M3.6) occurring in 2012 [1].

With as prime aim to assess the hazard and risk resulting from induced seismicity, an international research program developed an integrated Probabilistic Seismic Hazard and Risk Assessment (PSHRA) [1]. The PSHRA are annually submitted by the field operator NAM to the Dutch government as input into the decision by the Minister of Economic Affairs and Climate Policy for this decision on gas production levels. A key element of the PSHRA is a seismological model, which should forecast the temporal and spatial probability densities of earthquakes within the Groningen natural gas reservoir, conditional on future production plans [2].

Forecasting of induced seismicity requires a detailed understanding of both the physical mechanisms governing depletion-induced seismicity, as well as reservoir properties in time and space. On physical mechanisms, research from lab to field experiments and from experimental to complex numerical models has provided insights into the mechanisms behind induced seismicity in general [8–12], in particular depletion-induced seismicity [13–19] and injection-induced seismicity [20, 21]. On reservoir properties, data acquisition efforts [22–24] can improve data quantity and quality, but as properties can only be measured directly in wells (which gives very limited spatial coverage) or using seismic reflection data (which needs to be interpreted, and has limits in applicability in regions with induced seismicity), these datasets carry large intrinsic uncertainties.

Consequently, current state-of-the-art models for forecasting induced seismicity try to address these uncertainties by combining physical and statistical model elements. An example for depletion-induced seismicity is the PSHRA statistical-physics seismological model for the Groningen gas field. This model is based on three components: (1) The mechanics of poroelastic deformations to describe pore pressure depletion; (2) The statistics of extreme thresholds combined with Coulomb failure stresses to describe fault reactivations; and (3) the statistics of a heterogeneous Poisson Point Process to describe the probability of induced earthquakes as they vary in space and time [1, 2, 16]. Similar physical-statistical approaches have been used in studies of injection-induced seismicity [21].

1.2 Induced seismicity forecasting using machine learning

The physical mechanisms described by the Groningen PSHRA statistical-physics seismological model show superior forecast performance compared to a broad set of potential alternative physical mechanisms [16]. This provides strong support that the physical mechanisms as described by the PSHRA model are relevant mechanisms for seismicity forecasting in the hazard and risk assessment but does not rule out that additional, yet unknown physical mechanisms could play a role as well. Given the importance of the seismological model for the PSHRA, it was decided to explore additional seismological models built using alternative methods. Alternative statistics-physics based approaches were thought to have similar strengths and weaknesses as the current seismological model, but physics-based modelling requires making narrow choices in an uncertain area of physics. Machine Learning (ML) offered an approach utilizing another paradigm to look for alternative mathematical formulations

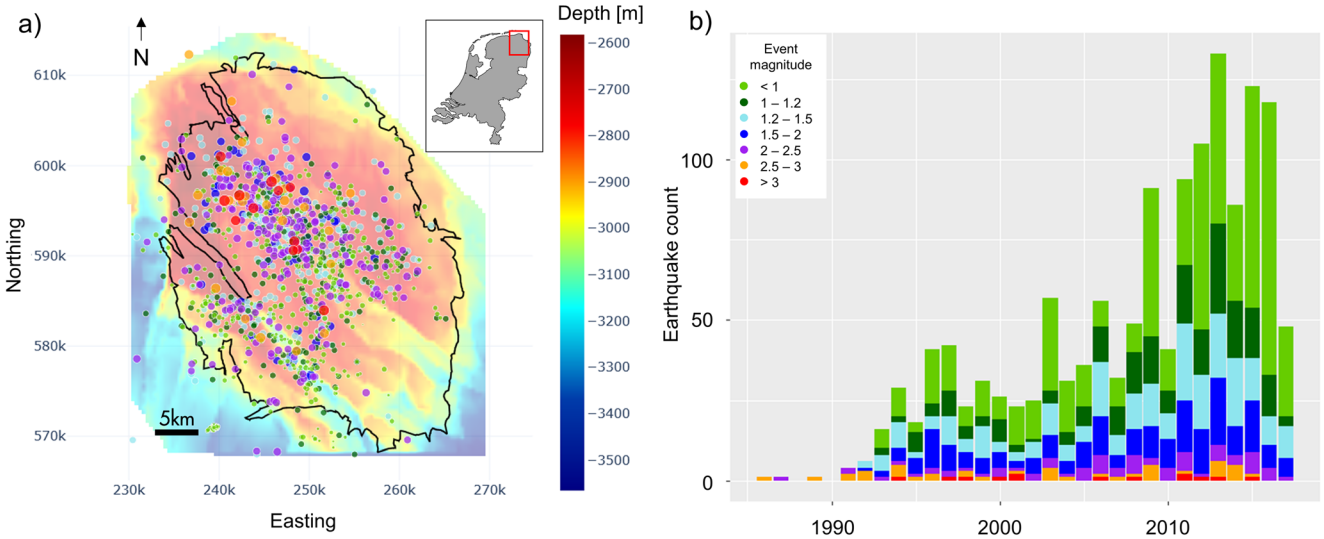

Fig. 1 Overview of seismicity between 1986 and 2017 in map view (left) and time view (right) with the different colours representing the different magnitude bins. The black outline shows the approximate field boundary. The background map shows the depth of the top reservoir surface in meters. The histogram only includes events observed within the black outline of the field

that utilize the available field data, that may outperform the physics-based models and provides the potential to be used in an operational setting. In particular, ML models have the ability to perform well in situations where underlying processes are not (fully) understood [25] and/or are complex [26], since ML models can infer non-linear relationships directly from data instead of requiring them to be prescribed in the modelling process. This approach is inspired by recent applications of ML to the field of geophysics [27, 28].

Combining ML with available physical knowledge has proven to accurately forecast the behaviour of a large spatiotemporal chaotic physical system where the mechanical description of the dynamics is limited [29], showing the potential for ML to *complement* physical and statistical seismicity modelling efforts [30]. In the context of seismicity analysis, ML has been applied to (i) earthquake identification [31, 32], (ii) catalogue-based seismicity forecasting [33] and (iii) model parameter inference (e.g. the Gutenberg-Richter b-value) [34–36]. Although these examples demonstrate the complementary value of ML methods, the use of complex ML models potentially introduces the risk of overfitting or providing a 'black box' solution to a physical mechanism that can be explained instead with physics-based models, as observed by [37, 38]. However, physical processes such as depletion-induced seismicity for which physical models exist [16] provide an opportunity to quantify the complementary value of ML methods [39].

We aim to build a pipeline to apply a set of ML methods (referred to as ML pipeline) on a broad set of physical parameters including but extending beyond the parameters used in the PSHRA model. Parameters include absolute pore pressure, compaction, average production rate, fault density and surface gradient, see Appendix Table 7 for a full overview. We did not commit to a specific ML algorithm or algorithm family a priori but instead selected a set of algorithms particularly suited for small and/or sparse datasets, taking into account a combination of ranking studies [40, 41]: a regularized GLM (Generalized Linear Model), a regularized GLM with the five most important features ("GLM top"), K-Nearest Neighbours (KNN), Random Forests (RF) and Support Vector Machines (SVM). These ML methods will be used to forecast the temporal and spatial induced seismicity rates within the Groningen gas field, conditional on two future production plans which result in varying degrees of field depletion, above a minimum magnitude of 1.5. For hazard and risk assessment magnitude information is required as well. In principle, the approach presented here could be extended to forecast seismicity rates over the dimensions of time, space and magnitude – however that is out of scope of the current work.

1.3 The machine learning pipeline developed in this study

To achieve our aim stated above, an ML pipeline is designed that extends time-series analysis methods to a spatiotemporal framework with a factorial setup [42, 43]. This pipeline allows probing a large parameter space of plausible modelling assumptions and meta-parameter choices, followed by a statistical meta-analysis to account for the intrinsic uncertainties in subsurface data and to ensure statistical significance and robustness of results. Here, meta-parameters are parameters defining the experimental setup, e.g. the spatial and temporal bin sizes, time delays, different interval start and end dates. A semi-automated implementation enables testing of hypotheses whether a specific parameter increases predictive performance in a statistically significant way.

We present the ML pipeline developed in the following order (Fig. 2):

- Data input, including the data sources used, aggregation approach, target and predictor feature generation, is discussed in Section 2.
- A prediction strategy, including choices for ML models, baselines and the tuning approach, is detailed in Section 3.
- The model evaluation and comparison approach, including confidence interval quantification, hypotheses testing, and the simulation model approach used are explained in Section 4.
- Meta-analysis based on a Random Forest model is described in Section 3.7.

The pipeline is implemented in the statistical programming language R, relying partly on off-shelf methodology available in the MLR package [44–48]. Functionality has been augmented by custom implementations of baselines (Section 3), error metrics and support for spatiotemporal model validation (Section 4).

2 Data preparation: Pre-processing and variable selection

The first part of the integrated ML pipeline consists of collation and pre-processing of various data sources to a 3D array of observations, indexed by longitude/latitude bins (referred to as lon/lat) and temporal bins (i.e. two spatial dimensions and one temporal dimension). The raw data comprises physical and geological properties and subsurface simulation model outputs at different spatial and temporal granularities, and a catalogue of seismic events. As off-the-shelf functionality of machine learning toolboxes such as the MLR toolbox used in this study usually assume tabular format, the raw data is converted to a single table in long tabular format, indexed by the

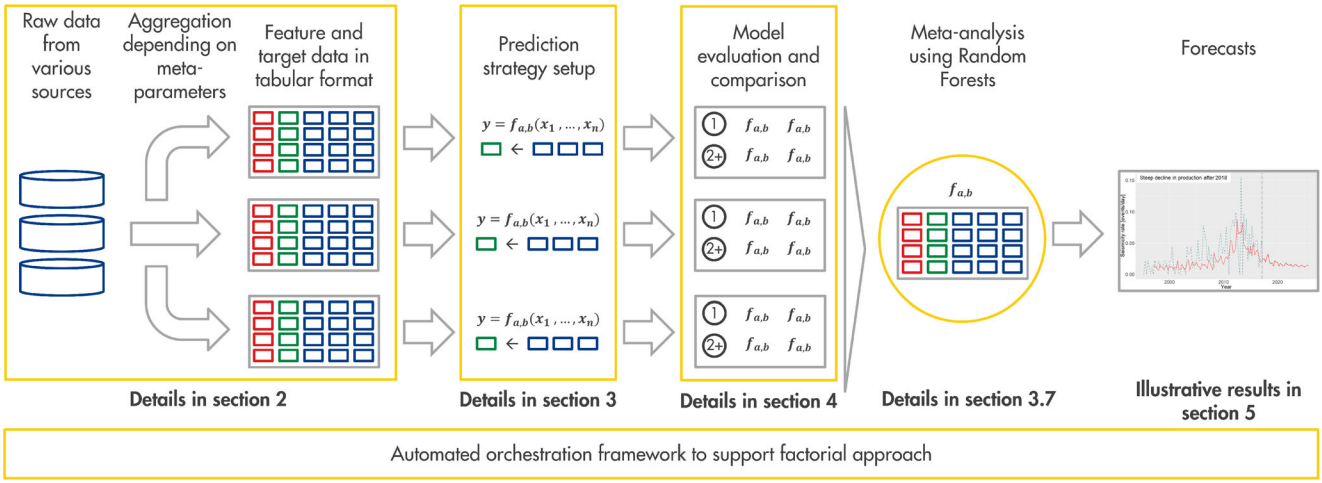

Fig. 2 Illustrative overview of the machine learning pipeline described in this article. Components are explained in subsequent Sections 2–4

lon/lat/time bins, by a combination of binning and interpolation techniques.

Section 2.1 describes the various raw data sources; Sections 2.2–2.3 detail pre-processing of the predictor features and the target feature and the final pre-processed data set in long format. With this data set complete, Section 2.4 explains the mathematical notation used in this study, which allows us to describe the subsequent variable selection process in Section 2.5. Physical considerations suggest that on the long term, decades after production has ceased, the system should come to rest. Section 2.6 describes the addition of 'ultimate states' to the input data, which are meant to encode these considerations.

The entire data processing pipeline is subject to certain parameter choices which we consider to be meta-parameters, these will be chosen subject to sensitivity analysis of the entire pipeline as further discussed in Section 3.

2.1 Input data and subsurface models

The data selection follows practical limitations of data availability, as well as statements in existing literature (e.g. [5, 16, 18]; Fig. 3) regarding relevance to seismicity and seismicity rates.

Our raw data includes both direct observations as well as outputs from subsurface models. Direct observations are obtained from wells, which provide local petrophysical characterization of rock properties, reservoir pore pressure measurements, gas production rates, and reflection seismic data. These data have also been used to build models of the reservoir structure and rock properties (time-invariant) and fluid flow distribution (time-dependent), making use of physics-based numerical solvers to interpolate between observations in space and time [49]. These may be leveraged for reservoir production history, or forecasts of reservoir properties for given future gas production scenarios. Compared to most subsurface reservoirs, production forecasts of the Groningen gas field have narrow uncertainty bands due to the full field coverage of 3-D seismic reflection data and the long history of reservoir surveillance data (production, pressure, subsidence, seismicity).

We would like to stress one key limitation of reservoir models as a data source: These models are frequently updated with data from the wells, by a process which is external to our modelling pipeline. While we carry out walk forward cross-validation we are unable to automatically history match the reservoir models such that they are precisely constrained to the data that would be available when we start forecasting in each temporal cross-validation fold. A complete walk-forward instantiation of the data preprocessing – in-principled required for fair evaluation – is thus not possible, and information leakage, or subsequent over-optimism in model performance evaluation, cannot be fully ruled out.

Nonetheless, geoscientific experts' confidence in the physical models tends to be high, therefore for all practical purposes the chosen setup of this study is considered to be a good proxy to the real-world situation of prospective model application.

Table 1 provides a summary of the different models and data sources leveraged in generation of the pre-processed data, including raw data resolution, and steps for interpolation or binning to obtain the tabular data set. The complete list of predictor features extracted from these sources is listed in Appendix Table 7, with a brief motivation as to why each predictor feature is potentially relevant for seismicity modelling.

The machine learning models will be used to generate induced-seismicity forecasts, based on reservoir model forecasts, which in turn depend on gas production scenarios. We consider two gas production scenarios with forecasts derived from the reservoir flow models to illustrate the influence of changing gas production on our seismicity forecasts [49, 51] (Fig. 4):

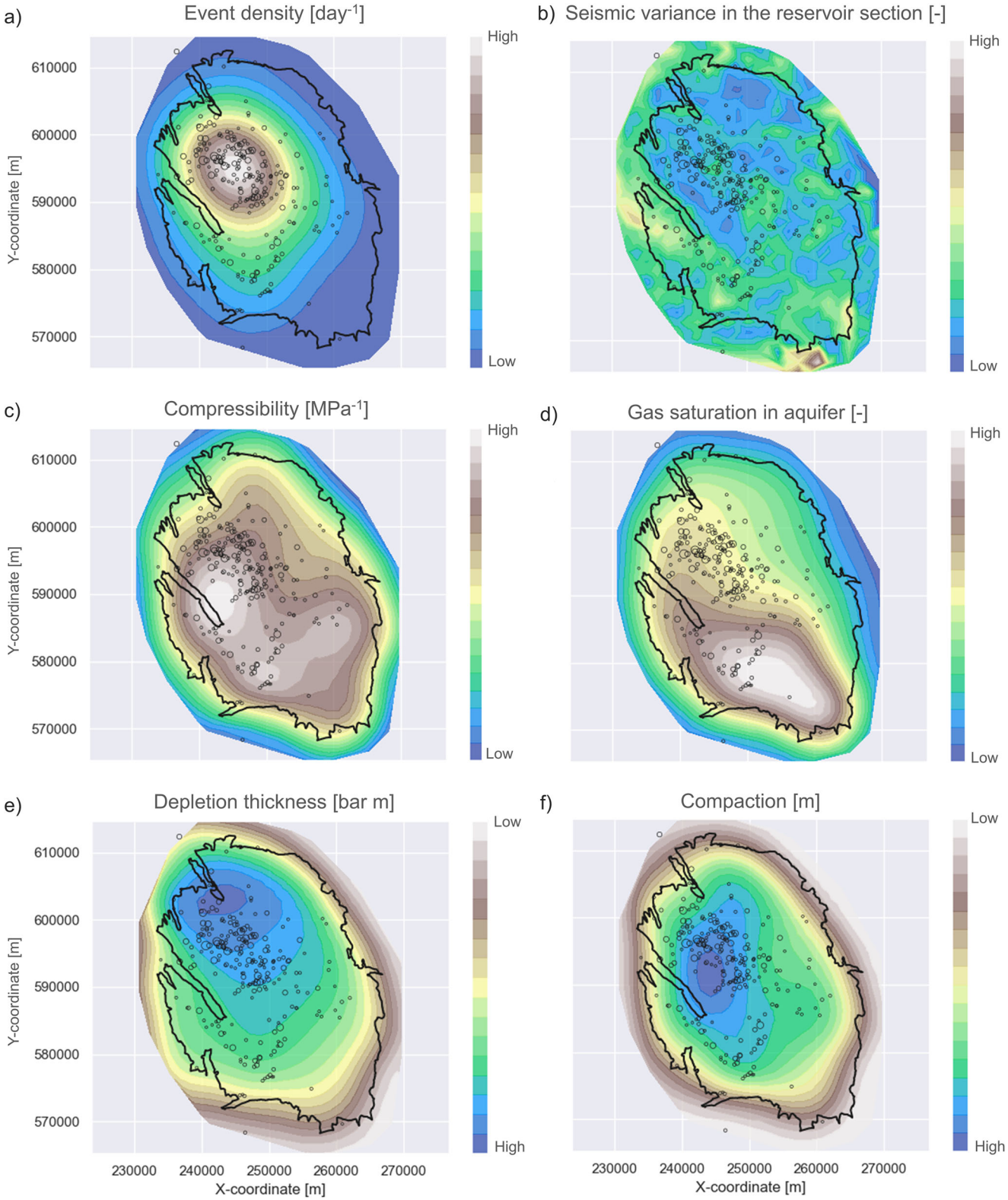

Fig. 3 Distribution of the target feature (a, seismic events) and examples of predictor features constructed for the Groningen gas field (b-f). The open circles show the distribution of seismic events between January 1st, 1995 – December 31st, 2016 with circle size scaled by magnitude. The depletion thickness and compaction are cumulative over the period of interest. Note that our models only consider data within the field outline

Table 1 Summary of available data sources for spatiotemporal feature extraction

Data source	Origin	Raw format	Resolution	Tabularization strategy
Fault model from the reservoir geological model [23]	Reservoir geological grid (i.e., regular grid at 100x100m horizontal 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 structural 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.	100 m × 100 m	Averaging for fault attributes (e.g., orientation, thickness) for multiple points within a cell. Fault density is calculated as the sum of fault points within a cell.
Seismic reflection data [23]	3-D seismic reflection volume attributes calculated from seismic reflection amplitudes around the top reservoir (surface attributes) or from the reservoir interval (volume attributes).	2-D point set with x, y data, and attribute values.	25 m × 25 m	Averaging of grid points within a cell.
Reservoir geological model (well data interpolated using acoustic impedance) [23]	Reservoir 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 predictor features using (cell) volume weighted averaging in the vertical direction.	Regular gridded 2-D point set with x, y centre coordinates per cell and attribute values.	100 m × 100 m	Averaging of grid points within a cell. The topographic gradient in each cell is calculated from the average surface gradient of the 8 neighbouring cells.
Compressibility inverted from subsidence data [50]	Compressibility model inverted from geodetic (subsidence) data.	2-D regular point set with x, y and compressibility values.	500 m × 500 m	Averaging of grid points within a cell.
3-D Geomechanical Finite Element reservoir model	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)	Averaging of grid points within a cell.
Reservoir flow model [49, 51]	Model geometry and static properties are upscaled from the geological model. The fluid model obtained from Pressure, Volume and Temperature (PVT) data, and pressure, saturation, temperature and flow are modelled based on calibration to measurements at wells.	3-D point set with cell centre points, cell volume, and dynamic attributes.	Variable, an average of 650 m. The temporal resolution is monthly.	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.

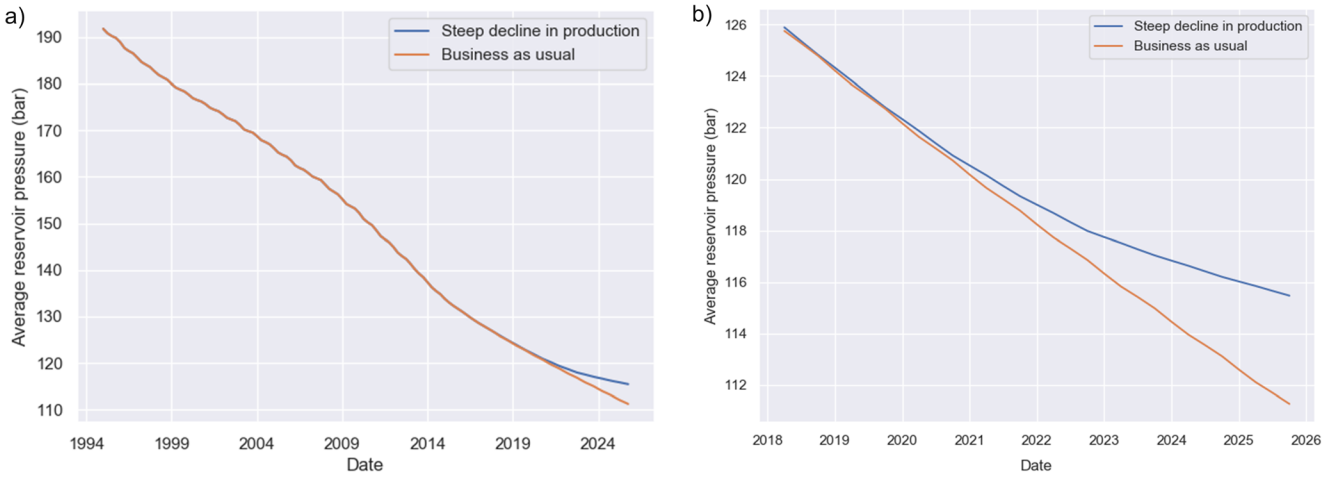

Fig. 4 Average reservoir pressure in the entire Groningen field as a function of time for two production forecasts derived from the reservoir simulation model, scenario 1 (orange) and 2 (blue). Left: Entire period of interest from 1994 up to 2025; Right: Zoomed-in version showing the difference in forecasted depletion between the two scenarios

1) The field is depleted to a relatively low pore pressure state by producing at a near-constant rate that is comparable to the historical rate of the recent past [24].
2) Production is rapidly decreased compared to the recent historical state and the field is shut in by 2030, at a depletion pressure that is higher than in the other scenario [52].

2.2 Binning and interpolation for data tabularization

The different data sources described in Section 2.1 have varying spatiotemporal resolutions. For application of machine learning toolboxes, the data are brought into tabular format by binning and interpolation. For all data sets, we follow the same high-level process to create data indexed by a rectangular grid with given spatial and temporal resolutions. The following steps are carried out for each individual predictor feature:

1) If the feature possesses a depth index, cell volume-weighted averaging over this depth index is carried out to remove the depth index entirely.
2) Second, the feature is binned according to the rectangular grid, and bin-averaged by longitude and latitude, for fixed time index.
3) Third, if the feature possesses a time index, temporal aggregation or averaging is applied.
4) Finally, for discrete features such as the fault network geometry, feature values are replaced by an isotropic Gaussian kernel interpolate for each fixed time index. The kernel bandwidth is considered a meta-parameter of the full modelling pipeline which is later optimized and investigated in the meta-parameter sensitivity analysis.

Furthermore, time lagged variants of time-indexed features (1, 2, 3, 6, 12 months) are included in the pre-processed dataset by adding each time-legged feature for each lag period as a separate feature to the feature matrix.

2.3 Target feature definition and prediction task

The pre-processed dataset also contains the target feature that the ML pipeline has to predict, namely the number of induced seismic events within a given cell within a time bin, at Richter magnitude 1.5 or above.

The choice for this Richter magnitude lower bound is to prevent modelling artefacts arising from changes in the sensitivity of the Groningen monitoring system associated with upgrades of the monitoring system over time. This follows argumentation and mirrors decisions taken in [53, 54]. In the final pre-processed data, 265 events are observed in the period between 1st of January 1995 and 31st of December 2016.

For reference in Section 4, we note that the probability distribution of these events in space and time is assumed to follow a Poisson Point Process, where we do not take into account aftershocks.

The resulting pre-processed dataset used for subsequent analysis is a single data table in long format, i.e. all spatiotemporal information is stored for each location for each temporal bin.

Most features are not direct observations but obtained from subsurface models. In Appendix Table 7 we list which features were obtained from the different subsurface models. The number of seismic events above the magnitude threshold is a direct observation. For simulated data at historical time points (December 2016 and earlier), the table is populated with history-matched data. Whenever the same predictor features are used in forecasting induced seismicity for January 2017 or

later, predictor features from the respective simulation model forecasts are used.

2.4 Mathematical notation for pre-processed data

For reference in later sections, we introduce a mathematical notation for the data set as described in Section 2.3. We will refer to the combined lat/lon cell locations on the grid by the symbols s_i, $i = 1, \ldots, B$, in no particular order, and not separating latitude and longitude notationally (this will not be of importance in occurring formulae). Historical cell times are noted using the symbols t_j, $j = 1, \ldots, T$, where indexing is in chronological order, from January 1995 (j=1) to December 2016 ($j = T$). The vector of predictor features in a given cell, is defined at a location s and time t by the symbol $x(s, t)$, with event counts (as in Section 2.3) noted by the symbol $y(s, t)$. The row vector $x(s, t)$ takes values in R^d, i.e., there are d variables, and $y(s, t)$ takes values in the integers N. There are in total $B \cdot T$ cells indexed by s_i, t_j, $i = 1 \ldots B$, $j = 1 \ldots T$. For each such cell there is one predictor feature row vector $x(s_i, t_j)$ and one target feature observation $y(s_i, t_j)$. We will also consider the total number of events per time bin, $z(t) := \sum_{i=1}^{B} y(s_i, t)$, of which there are T instances $z(t_1), \ldots, z(t_T)$.

2.5 Predictor feature pre-selection algorithm

The last data preparation step in the pipeline is one of variable selection to down-sample the initial 63 unique features derived from models and data, multiplied by several factors depending on the model time lag parameters as described in Sections 2.2–2.4. Predictor features are selected in two ways described in further detail below:

(i) Based on stated expert relevance, and common use in physics-based models found in literature [5, 16, 18].
(ii) Based on an unsupervised approach to limit correlation between different predictor features. Predictor features are removed stochastically based on a 90% absolute (Pearson) correlation threshold that aims to remove features that are essentially copies of each other.

The variable selection based on expert relevance is guided by choices in the Coulomb stress model for heterogeneous thin sheet reservoirs [55]. This model relates the likelihood of seismic fault slip to the fault geometry, friction properties, initial stress state and poroelastic stress changes due to reservoir pressure depletion [16]. Examples of direct representations of Coulomb model features are pore pressure changes and fault geometry features from the reservoir models. Examples of potential proxies are reservoir rock shale volume as a proxy of fault friction behaviour, or seismic variance attributes as a proxy of fault density. The complete list of considered variables is included in Appendix Table 7.

Algorithmic variable selection based on correlation thresholding (ii) is based on the following unsupervised dimension reduction algorithm:

1. Between all prediction features, we compute sample correlations. More precisely, correlations $C_{k\ell}$ are obtained as independent sample correlations between the paired samples $x(s_i, t_j)_k$, $i = 1 \ldots B, j = 1 \ldots T$ and $x(s_i, t_j)_\ell$, $i = 1 \ldots B, j = 1 \ldots T$, where pairing is by the joint indices i, j.
2. Prediction features that have at least an absolute correlation above 0.9 with another prediction feature are "marked". That is, we compute the indicator set $S := k : |C_{k\ell}| > 0.9$ for some ℓ.
3. If S is empty, terminate; otherwise, select an element of S uniformly at random, and delete the corresponding column in the data $x(., .)$. Then, go to 1.

A few remarks are to be made about the above algorithm: First, it is heuristic and stochastic. The criterion for variable selection is not based on performance of the entire pipeline; Second, the correlation is computed as the independent sample correlation, but the samples $x(s_i, t_j)$, $i = 1 \ldots B, j = 1 \ldots T$ are not independent samples, but correlated spatially and temporally, thus the correlations are not necessarily a measure for statistical dependence of the features.

Despite these limitations, we argue that this heuristic, automated procedure is sensible with regard to its primary aim: removal of predictor features that are essentially duplicates and copies – rather than, say, as a general modelling principle.

2.6 Augmentation by artificial data points ('ultimate states')

To ensure that physically plausible predictions are made at feature prediction combinations mimicking shut-in states of the reservoir, several artificial data points obtained from the reservoir simulation forecasts mimicking this state are added to the time series:

1) The reservoir flow simulation model that is used to generate forecasts of the dynamic reservoir properties (e.g. pressure) as described in Section 2.1 is run for an extended simulation time, that covers the period during which gas is produced from the reservoir as well as an additional 50 years period after production is ceased. This 50-years period is sufficiently long to allow the reservoir model to reach a pressure equilibrium after production has ceased. The reservoir properties (e.g. pressure, saturation) at the end of this simulation are defined as representative for the 'ultimate state' of the reservoir, after the field is shut in and depletion-induced seismicity is expected to have

ceased. We use multiple production scenarios that explore the complete range of reservoir depletion between near-immediate shut-in and maximum depletion to avoid introducing any bias (Fig. 5) [43].

2) The 'ultimate state' values of the dynamic reservoir simulation model properties are appended to the time series for each lat/lon cell. The seismicity rate for these additional points is set to zero.

From an ML perspective, this turns an extrapolation problem into an interpolation problem. Comparison between models which are given access to 'ultimate state' values with the same models that do not have such access, show that on historical forecast periods access to 'ultimate state' values does not statistically significantly affect forecast performance [43].

3 Prediction strategies, baselines, and hyper-parameter tuning

This section describes the experimental setup, the learning strategies employed and parameter tuning.

3.1 Modelling task

The task for the algorithms is to produce earthquake count predictions for each cell, 3 months ahead, based on contemporaneous (simulator) prediction features – that is, the k-th model is used to predict earthquake counts $\widehat{y}_k(s,t)$, for time intervals $t_a, \ldots, t_T$, where $a = 9$, thus t_a corresponds to the period Apr-Jun 1997. The duration of each time period for which cumulative counts are predicted is 3 months. To produce $\widehat{y}_k(s,t_j)$ for a particular month t_j (model training), every contender algorithm has access to features $x(s_i, t_{j'})$, for any i and $j' \leq j$, i.e., to predictor features that are possibly contemporaneous. Additionally the algorithms have access to past observations of earthquake counts, that is $y(s_i, t_{j'})$, for any i and $j' \leq j-1$. For prediction, the algorithms have access to the same data. In addition, predicted overall counts $\widehat{z}_k(t)$ by time bin are calculated as $\widehat{z}_k(t_j) := \sum_{i=1}^{B} \widehat{y}_k(s_i, t_j)$. Predictions are forced to be non-negative real numbers trough clipping to zero if needed. However, no rounding to the closest integer is carried out.

3.2 Experimental set-up: Model training, tuning, validation, and application

Our experiment is run in two steps: (a) model training combined with iterative walk-forward model performance evaluation and (b) prospective application.

In model training and performance evaluation, every contender algorithm produces three-months-forecasts for the period April 1st, 1997 to December 31st, 2016. Every model has access to all data up to the prediction period. Individual algorithms may choose to further sub-divide the data internally for hyper-parameter tuning. The model performance benchmarking for the algorithms is carried out using a walk-forward approach as detailed in Section 4. A manageable subset of "winning" algorithms and associated parameter settings are selected based on the calculated performance metrics. These models are then invoked (b) to produce earthquake count forecasts several years ahead beyond January 2017, using reservoir simulation based features $x(s, t)$, for $t >$ January 2017, by applying the algorithms iteratively as if the future time point were one time-step in the future.

3.3 Types of learning strategies

Two classes of learning strategies are tested as contender strategies for the time series forecasting problem at hand:

(i). Supervised regressors, trained on all feature-label pairs (s_i, t_j), $y(s_i, t_j)$ within the respective training period as part of the walk-forward training testing setup. For prediction, targets are queried for all feature vectors within the test period, that is, $x(s_i, t_j)$ post-April-1997 for model benchmarking, and simulator forecast features for prospective use. We list these in Table 2 with the associated hyper-parameters that were tuned as part of this study. Note that the simulator outputs, i.e. the reservoir pressure and saturation changes, prior to January 2017 have been history matched according to common state-of-art practice. Even though no earthquake data is used in the history matching process, potential temporal information leakage from future to past within our benchmarking set-up may have occurred since the study team was not involved in the history matching process.

(ii). Time series forecasters based on geomechanical drivers which serve as state-of-art baselines. These are described in Section 3.5 and come with only one hyper parameter q, being the number of previous time steps which are taken into account.

We note that within this manuscript we differentiate two sets of parameters: hyper-parameters are tuning parameters specific to a particular ML model (Table 2), whilst meta-parameters are related to the general experimental setup (Table 3). The hyper-parameters are auto-tuned via the IRace algorithm exposed via the MLR package, whereas the meta-parameter are chosen as discussed in Sections 3.6–3.7.

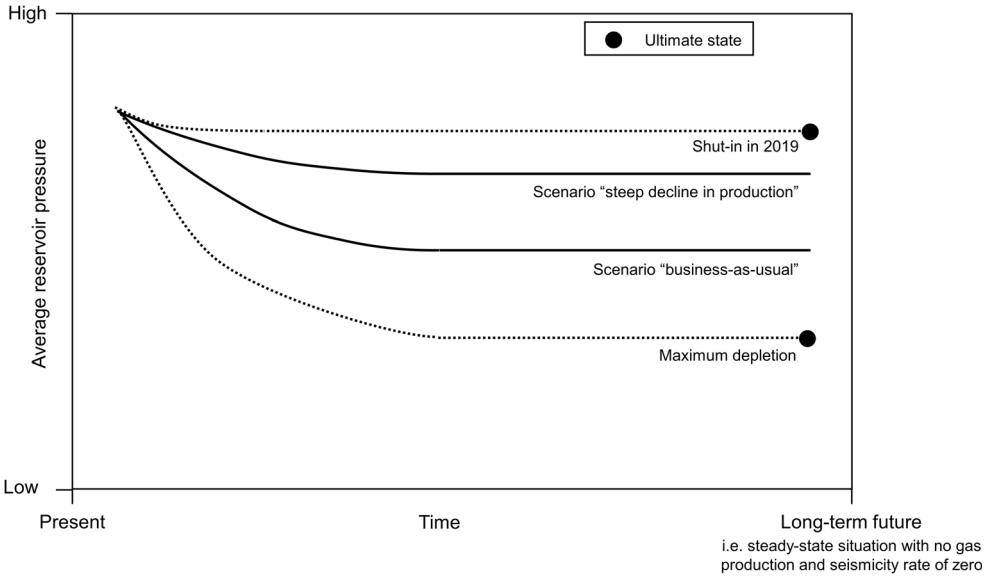

Fig. 5 Illustration of the range of physically possible depletion states that are used to generate the ultimate states for weighted mean pressure provided as training data to the ML pipeline

3.4 Supervised strategies for contemporaneous seismicity forecasting

The supervised learning strategies used, with hyper-parameters, are listed in Table 2. Note that by including lagged features in the input of the algorithms, the time series nature of the data and prediction problems remains honoured. Thus, a larger set of machine learning techniques is accessible compared to only considering classical techniques specific to time series forecasting. Additionally, the MLR pipelines can be used without major modifications which would easily allow to re-run the benchmarking and prediction pipeline with some of the additional regression models that are wrapped into MLR.

3.5 Baseline strategies

The baseline strategies are variations of a temporal moving average strategy.

The *exogeneous moving average with proportionality to a variable difference* assumes that moving averages of earthquake rate, and relative change in another variable's first differences, are proportional to each other. This baseline strategy is physically motivated by the common physical model assumption that earthquake rate is driven by another variable's gradient. Choices for the exogeneous variable, in our experiments, are average depletion thickness (Fig. 3e) and strain thickness (Fig. 3f), within the lon/lat/time cell, which have been found to better capture the spatial distribution of

Table 2 Summary of supervised learning strategies. 1st column is vernacular name; 2nd column is MLR (v2) learner ID; 3rd column are hyper-parameter settings; this only lists hyper-parameters that chosen differently from their MLR default; 4th column are tuning grids for hyper-parameters; 5th column is the used tuning strategy. Tuning is subject to random (non-temporal) sub-sampling, which is not a major issue as the evaluation set-up is temporal

Name of strategy, reference	MLR learner ID (if applicable)	Fixed hyper-parameters	Tuned hyper-parameters, tuning grid	Tuning strategy
Regularized GLM [56]	regr.glmnet	nlambda = 200	family $\in$ {gaussian, possion} $\alpha \in \{0, 0.1, \ldots, 1\}$	IRace package with max. 1000 iterations [57]
Regularized GLM [56] with top 5 features ("GLM top")	regr.glmnet	nlambda = 200 and applied to the top 5 features identified by permutation based variable importance	family $\in$ {gaussian, possion} $\alpha \in \{0, 0.1, \ldots, 1\}$	
K-Nearest Neighbours [58]	regr.kknn	(unchanged)	$k \in \{1, 2, \ldots, 10\}$	
Random forest [59]	regr.ranger	(unchanged)	mtry $\in \{1, 2, \ldots, 10\}$	
Kernel SVM [60, 61]	regr.ksvm	(unchanged)	type $\in$ {eps − svr, nu − svr, eps − bsvr} epsilon, sigma, nu, C $\in \{10^{-5}, 10^{-4}, \ldots, 10^{4}, 10^{5}\}$	

Table 3 Overview of model meta-parameters which are considered in the ML pipeline

Meta-parameter	Description
ML model (excl. baselines) including model parameters	Type of machine learning model and respective hyper parameters of the model.
Gridsize	Resolution of the grid cells for spatial gridding to generate x, y, feature values.
Time delay production	Delay (number of time steps) in production data versus target quantity.
Minimal magnitude	Lower bound for earthquake magnitudes to be used.
Time delay pressure	Delay (number of time steps) in pressure data.
Number of spatial blocks	Number of blocks that the grid cells in the field are divided in using k-means clustering
Time delay compaction	Delay (number of time steps) in compaction data.
Kernel smoothing bandwidth	Bandwidth in meters of the kernel smoothing applied to spatial predictor features.
Max. nr. Lags	Maximum number of lags to be added to the time-series data.
Feature correlation threshold	Threshold above which predictor features are defined as highly correlated. These features are then grouped, and one representative feature is used.
Interval length	Length of the period over which features are temporally aggregated.
Feature significance threshold	Minimum threshold for predictor features to be considered significant.
Interval start	Start of time interval for model training
Interval end	End of time interval for model training

seismicity compared to other reservoir properties [16]. Mathematically, predictions are obtained as

$$\widehat{y}_k(s,t_\tau) := \frac{v(s,t_\tau)-v(s,t_{\tau-1})}{v(s,t_{\tau-1})-v(s,t_{\tau-q-1})} \sum_{i=\tau-q}^{\tau-1} y(s,t_i) \quad (1)$$

where $v(s,t)$ is the exogeneous predictor feature available for the same cell at time up to t, i.e., the forecast assumes contemporaneous availability of v.

In the experiments, $v(s,t)$ is derived from the prediction features in Section 2.5. The two instances of these baselines are obtained from the pre-processed data by calculating depletion thickness as pressure×reservoir thickness, respectively strain thickness as vertical strain×reservoir thickness, at location s and time t, for $v(s,t)$. For tuning the lookback period q *is* considered a hyper-parameter.

3.6 Analysis and tuning of model meta- and hyper-parameters

The main model meta-parameters are related to data integration choices, and feature selection thresholds. When considering the values probed for each of the pipeline meta-parameters, we face the trade-off between range and resolution which is addressed via an iterative refinement of the parameter grid, i.e. starting with regularly sampling a broad parameter range with coarse resolution and iteratively re-sampling well performing and robust meta-parameter ranges with a higher resolution until the observed improvements in performance become insignificant. To enable insights into how the parameters impact model performance, the following data is recorded for each experiment, including:

- Meta-parameter choices;
- Model performance for each error metric with the associated standard errors;
- Random Forest (RF) based variable importance (estimated increase in MSE together with SE) for each feature [60];
- Listing of significant and potentially significant features as determined by the Boruta variable importance test [62].

The meta-parameters which we vary as part of this study are highlighted in Table 3. We use a factorial setup to study the impact of these parameters. As a major benefit, this setup permits sensitivity analysis of prediction results with respect to these modelling choices, allowing optimizing model performance in a robust way by exposing parameter combinations that yield similar and competitive performance under small perturbations of the meta-parameters. This is achieved by applying the so-called "meta-analysis pipeline". In brief, it is leveraging techniques from interpretable AI and is described in Section 3.7. By running a factorial experimental design, probing a range of plausible combinations of parameters, interaction effects between parameters with sufficient effect size can be detected.

In contrast to meta-parameter tuning, hyper-parameter tuning is fully automatic and embedded in the pipeline via the generic MLR model tuning interface. See Table 3 for a

list of parameters that have been optimized. Whenever model hyper parameter tuning is enabled within the pipeline, the same re-sampling strategy is used which is also employed for the outer validation loop, however with a coarser step size of $2a$, see Section 3.1. This means that we are using walk-forward cross validation to tune the hyper-parameters as part of the model training procedure. Again, the coarser step size represents a trade-off between computational complexity and our ability to find better hyper parameters. We recognize that his may potentially lead to sub optimal model performance.

3.7 Meta-analysis pipeline

Model based meta-analysis allows mostly automated assessment of modelling results using techniques from interpretable AI. We use it to obtain a combination of well-performing and, with respect to small perturbations, stable parameter settings.

The first step is establishing a functional relationship between the meta-parameter choices and model performance. We have chosen to use the Random Forest (RF) algorithm [59]. The resulting model is called the meta model. RF models are not particularly sensitive to data pre-processing choices like normalization and accept both categorial and numerical data. However, other regression models could also haven been considered.

First, we assess the quality of the meta model using the reported out-of-sample estimate of the explained variance R^2. Only if the model manages to explain a set fraction of the variance in the data, we proceed with further analysis. This includes testing for significant variables and meta model variable importance analysis using permutation based variable importance analysis.

This type of importance analysis is insensitive to monotonic transformations of the input data and still works well in the presence of several noise variables which are un-related to the prediction target. Variables for which confidence bands overlap are indistinguishably important within the meta model.

One subtlety regarding the interpretation of running iterative experiments with different subsets of predictor features is that if removal of a feature makes the model worse, it was important. The other direction however is not true, i.e., if removal does not lead to a degradation of model performance, it still could have been important, but the encoded information could be captured through other associated features. To mitigate this problem, the Boruta algorithm [62] is used to establish which features have significant impact on model performance.

The algorithm implements a sophisticated heuristic to determine which predictor features are relevant for the predictive performance of the model in the context of the full set of features that are used to create the model. In brief the algorithm iteratively tests the importance of specific features via two-sided t-tests in which the performance of the model with the original features is compared to the performance in which the feature has been randomly permuted. Several iterations are necessary to account for potential interaction effects between features. Full details about the underlying algorithm can be found in [62]. It should be noted that no statement about the importance of a variable in isolation can be made since the RF models do capture interaction effects between the given variables.

Eventually, stable parameter ranges for the most important features are established using standard ICE-plots which are explained in Appendix 2. This is done by visual inspection of the ICE-plots for the individual variables, by identifying troughs in the graph and choosing the deepest trough for which we observe relative stability of the model performance in proximity to the minimal value in the trough.

4 Model performance evaluation and comparison

We describe the model evaluation and comparison methodology to assess performance of the forecast methods. Performance is assessed with respect to the following two tasks:

(i). To accurately forecast the overall number of events $z(t)$ in the time bin t in the entire Groningen field, 3 months ahead, given all prior months of data [42];
(ii). To accurately forecast the number of events $y(s, t)$ in the time bin t, for all lon/lat bins s within the Groningen field, 3 months ahead, given all prior months of data [43].

As outlined in Section 3.1, each algorithm is queried to produce forecasts. The k-th algorithm produces forecasts $\widehat{z}_k(t)$ and $\widehat{y}_k(s,t)$ at time bins $t = t_a, \ldots, t_T$, of these quantities.

4.1 Confidence intervals for performance quantifiers of event predictions

The performance quantifiers we use are described in Table 4. We obtain confidence intervals on these quantifiers by common procedures applied to the mean of an uncorrelated sample. We observe that all quantifiers (up to squaring, for RSMLE) can be written as $\varepsilon_k := \frac{1}{T-a+1}\sum_{j=a}^{T} L_k(t_j)$ for a suitable choice of series $L_k(t)$ of performances per test time points. We can then obtain a standard error of the sample mean as $SE(\varepsilon_k) := \sqrt{\frac{1}{T-a+1}\sum_{j=a}^{T}\left(L_k(t_j)-\varepsilon_k\right)^2}$. A 95% confidence level can then be obtained from the standard error by normal approximation. This process makes the assumption that the values of $L_k(t)$ are uncorrelated due to the use of a 3-month aggregation period removing any autocorrelation. This could be rigorously confirmed using, for example, a Durbin Watson test.

Table 4 Performance quantifiers for total number of events used in this study. Each row describes a performance quantifier for the quality of predictions

Err. Metric	Formula	Properties
MAE (Mean Absolute Error)	$\frac{1}{T-a+1}\sum_{j=a}^{T}\lvert\hat{z}_k(t_j)-z(t_j)\rvert$	• very common choice in regression • measures how well the conditional median is predicted
RMSLE (Root Mean Square Logarithmic Error)	$\sqrt{\frac{1}{T-a+1}\sum_{j=a}^{T}\left(\log(\hat{z}_k(t_j)+1)-\log(z(t_j)+1)\right)^2}$	• common choice for count prediction • measures how well the conditional harmonic/logarithmic mean is predicted
MPL(Mean Poisson Loss)	$\frac{1}{T-a+1}\sum_{j=a}^{T}\left(\hat{z}_k(t_j)-z(t_j)\log(\hat{z}_k(t_j))+\log(z(t_j)!)\right)$	• common choice for count prediction • up to scaling, same as negative predictive log-likelihood under bin-wise Poisson assumption

4.2 Formal comparison of performances by hypothesis tests

For formal comparison of performances quantifiers as presented in Section 4.1, i.e., to certify whether methods have significantly different performance, we employ paired hypothesis tests in the frequentist paradigm [63].

Effect sizes for performance differences between methods k, k' are obtained as differences between mean performance estimates, $\varepsilon_k - \varepsilon_{k'}$, with ε_k as defined in Section 4.1. Significances are obtained by applying a paired sample test on location to samples of test performances $L_k(t_i)$, considered to be paired/blocked via the method index k. In particular, the Wilcoxon signed rank test was used, since it is non-parametric and hence makes no explicit distributional assumptions.

Since the final number of models which remain after applying the meta and hyper parameter tuning described in Section 4 is very small the uncorrected significances of the Wilcoxon signed rank test of each method is calculated against the best-performing baseline. Note, that in theory this could lead to overoptimistic performance estimates. The tests are applied to predictions which have been obtained within the walk-forward validation procedure.

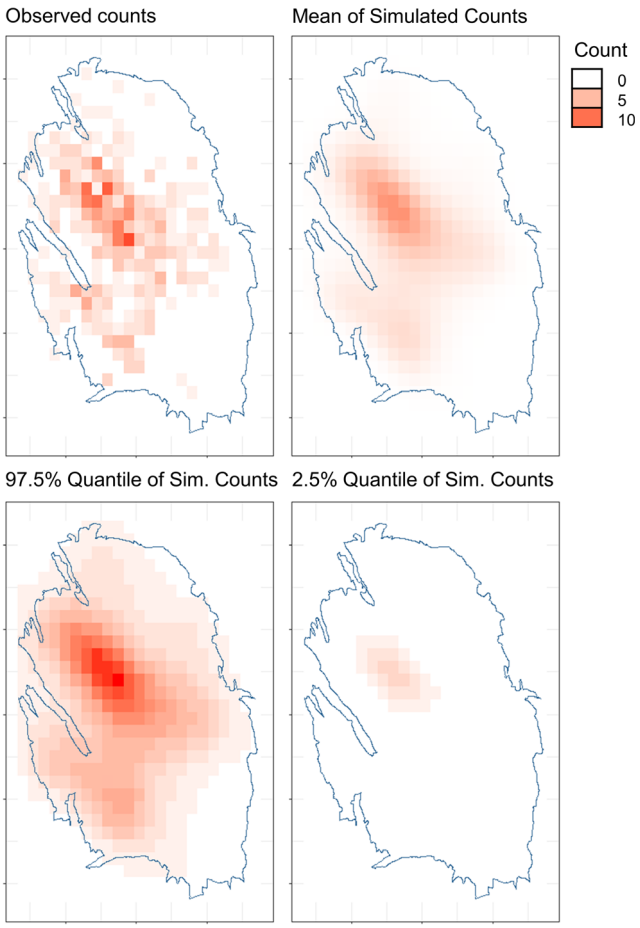

Fig. 6 Spatial plots of earthquake counts aggregated through time. Moving clockwise from the top left, the plots show the recorded earthquake counts for $M \geq 1.5$, the mean count of 1000 simulations from, the 97.5% quantile of the simulated counts and the 2.5% quantile of the simulated counts. The gradient from white to red depicts the range of 0–10

4.3 Assessing and comparing performance of bin-wise prediction – R-test variant

For testing the predictive performance of predictions across temporal and spatial dimensions, we propose a method which takes inspiration from the R-test as outlined in [64]. The basis of this test is the log-likelihood ratio

Table 5 Summary of relevant experiment meta-parameters (see Table 4 for complete list of available meta-parameters) used for the experiments discussed in the results

Parameter	Value
Seismicity data used for period	01-01-1995 – 31-12-2016
Aggregation period length	3 months
Minimum magnitude	1.5
Grid size	1500 m
Bandwidth of spatial smoothing	3500 m
Time shifts	None

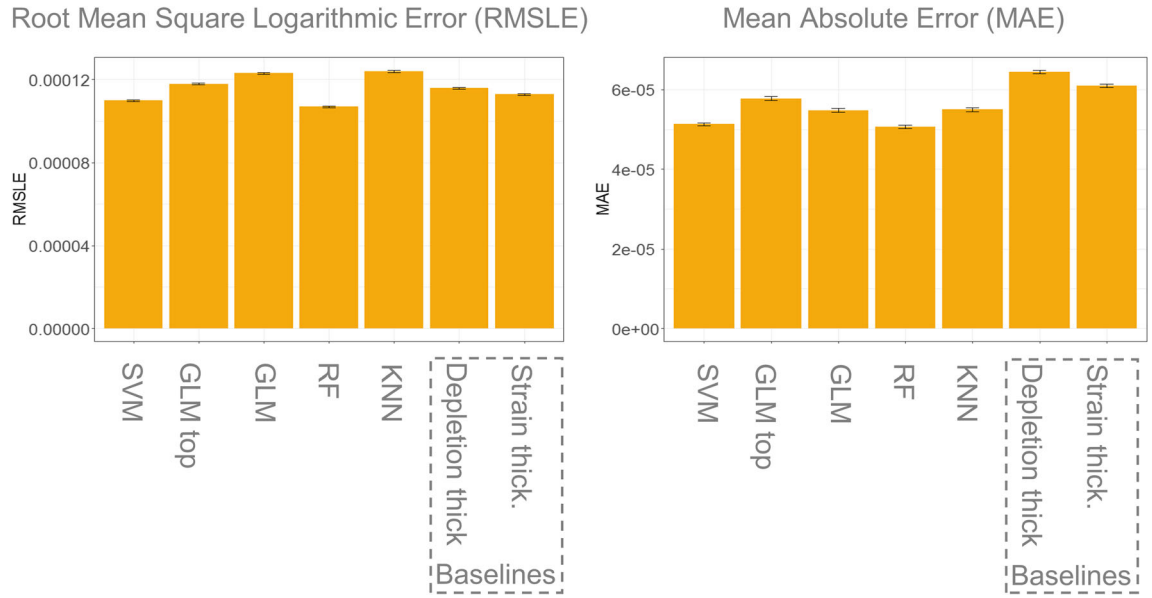

Fig. 7 The RMSLE (left) and MAE (right) error metrics on a cell by cell basis per model. The errors are derived by comparing predictions for each cell with the actual earthquake rates in 3 months. The black bars denote standard error calculated with the Jackknife resampling technique

$$r(\mathcal{M}_k, \mathcal{M}_0|\boldsymbol{y}) := \ell(\mathcal{M}_k|\boldsymbol{y}) - \ell(\mathcal{M}_0|\boldsymbol{y}), \tag{2}$$

with $\ell(\mathcal{M}_k|\boldsymbol{y})$ being the predictive log-likelihood of the k-th model which is denoted as $\mathcal{M}_k$. Under assumption of a bin-wise Poisson model, the explicit form of the log-likelihood is

$$\ell(\mathcal{M}_k|\boldsymbol{y}) = \sum_{i=1}^{B}\sum_{j=a}^{T} y(s_i,t_j)\log\left(\hat{y}_k(s_i,t_j)\right) - \hat{y}_k(s_i,t_j) - \log(y(s_i,t_j)!), \tag{3}$$

with the null case $k=0$ chosen to correspond to a baseline model of spatiotemporally smoothed null predictions $\hat{y}_0(s_i, t_j)$, which are detailed in Section 4.4. A positive log-likelihood ratio indicates that the k-th model is better than this baseline.

For the R-test significance for whether the log-likelihood ratio is positive is usually obtained from a mean-variance estimate, based on empirical sample simulations [64]. Though, this method has been criticized [65–67], amongst others for dependency on the simulation process rather than on the null model choice alone. We aim to address these issues by using a separate simulation model, rather than simulating from the proposed model and baseline. We then repeatedly evaluate the log-likelihood ratio on the data simulated from this model, denoted as $\tilde{\boldsymbol{y}}$, to give an estimated distribution for $r(\mathcal{M}_k, \mathcal{M}_0|\boldsymbol{y})$. The significance of the ratio being greater than zero can then be calculated directly as the proportion of simulations which are above 0.

For our particular choice of likelihood and simulation model, described below, we can side-step the simulation as the bin-wise Poisson assumption directly allows us to calculate expectation and variance of the log-likelihood ratio analytically. By some elementary calculation these are given by,

$$\begin{aligned}
\mathrm{E}\left[r\left(\mathcal{M}_l, \mathcal{M}_0|\tilde{\boldsymbol{y}}\right)\right] &= \mathrm{E}\left[\ell\left(\mathcal{M}_l|\tilde{\boldsymbol{y}}\right) - \ell\left(\mathcal{M}_0|\tilde{\boldsymbol{y}}\right)\right], \\
&= \mathrm{E}\left[\sum_{i=1}^{B}\sum_{j=a}^{T} \tilde{y}(s_i,t_j)\log\left(\hat{y}_l(s_i,t_j) - \hat{y}_0(s_i,t_j)\right) - \hat{y}_k\left(s_i,t_j\right) + \hat{y}_0\left(s_i,t_j\right)\right] \\
&= \sum_{i,j}\left[\theta(s_i,t_j)\left\{\log\left(\hat{y}_l(s_i,t_j) - \hat{y}_0\left(s_i,t_j\right)\right)\right\} - \hat{y}_l\left(s_i,t_j\right) + \hat{y}_0\left(s_i,t_j\right)\right].
\end{aligned} \tag{4}$$

Table 6 Model test results on a cell by cell basis for the selected models from Fig. 7 and the baselines for the period 1995–2016, including Mean Average Error (MAE), Root Mean Squared Logarithmic Error (RMSLE) and Mean Poisson loss error metrics together with the respective standard deviations for the period 1995–2016. The standard errors are calculated based on the Jackknife resampling method. "R VAL" shows the log likelihood ratio value calculated using the observed counts where a negative value indicates that the model does not perform better than the baseline, and the columns "E[R]", "VAR[R]" and "*P* VALUE" are the expected value, variance and *p* value calculated using the formulas in Section 4.3

	MAE	RMSLE	Mean Poisson loss	R VAL	E[R]	VAR[R]	*P* VALUE
RF	5.1E-05 (±4.0E-07)	1.1E-04 (±4.0E-07)	2.48E-02 (±1.4E-03)	110.8	97.5	282.0	3.2E-09
SVM	5.1E-05 (±4.1E-07)	1.1E-04 (±4.1E-07)	2.63E-02 (±1.5E-03)	32.1	38.4	254.0	8.1E-03
Depletion Thickness	6.4E-05 (±4.1E-07)	1.2E-04 (±4.1E-07)	2.76E-02 (±1.5E-03)	−44.1	−41.8	13.5	1
Strain Thickness	6.1E-05 (±4.1E-07)	1.1E-04 (±4.1E-07)	2.68E-02 (±1.4E-03)	−145.4	−133.4	350.8	1

$$\begin{aligned}&\mathrm{var}\left[r\left(\mathcal{M}_l,\mathcal{M}_0|\tilde{\boldsymbol{y}}\right)\right]\\&=\mathrm{var}\left[\sum_{i=1}^{B}\sum_{j=a}^{T}\tilde{y}(s_i,t_j)log\left(\hat{y}_l(s_i,t_j)-\hat{y}_0(s_i,t_j)\right)-\hat{y}_k\left(s_i,t_j\right)+\hat{y}_0\left(s_i,t_j\right)\right],\\&=\sum_{i,j}\left[\theta(s_i,t_j)\left\{log\left(\hat{y}_l(s_i,t_j)-\hat{y}_0\left(s_i,t_j\right)\right)\right\}^2\right],\end{aligned}\quad(5)$$

where $\theta(s,t)$ is the rate parameter used for the simulation model such that $\tilde{y}(s_i,t_j)$ ~Pois($\theta(s_i,t_j)$) and $\hat{y}_l(s,t)$ and $\hat{y}_0(s,t)$ are the rates predicted by models $\mathcal{M}_l$ and $\mathcal{M}_0$. For large effective sample size, $r(\mathcal{M}_l,\mathcal{M}_0|\tilde{\boldsymbol{y}})$ is approximately normal distributed due to the central limit theorem, and we can obtain a p value as the approximate proportion of simulated log ratios which are below zero as,

$$\mathrm{p}=1-\Phi\left(\frac{\mathrm{E}\left[r\left(\mathcal{M}_l,\mathcal{M}_0|\tilde{\boldsymbol{y}}\right)\right]}{\sqrt{\mathrm{var}\left[r\left(\mathcal{M}_l,\mathcal{M}_0|\tilde{\boldsymbol{y}}\right)\right]}}\right),\quad(6)$$

where $\Phi(.)$ is the standard normal cumulative distribution function.

Summarizing, we obtain a test for quality of the k-th method's predictions, i.e., for

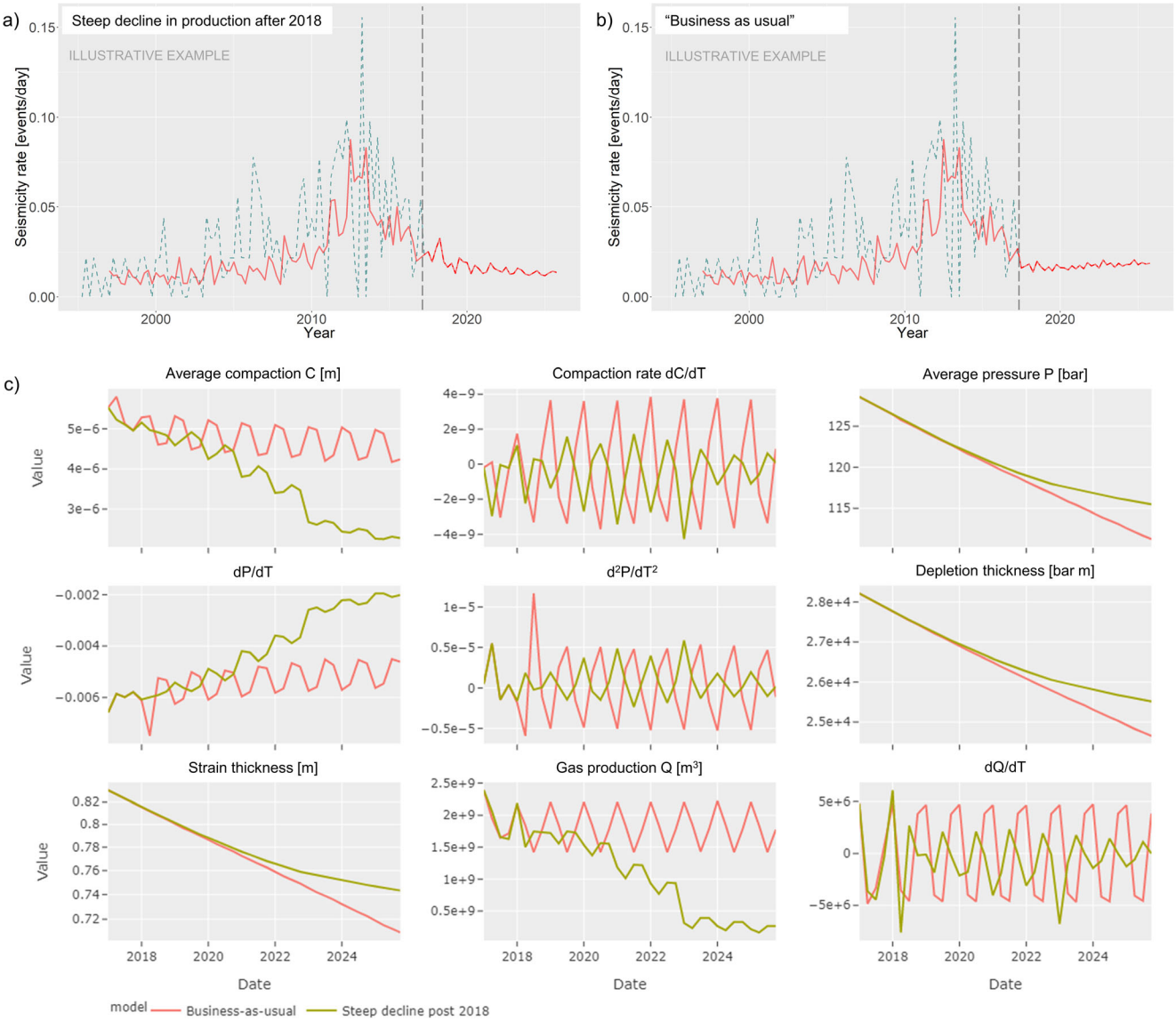

Fig. 8 Illustrative examples of the aggregated spatiotemporal forecasts (red) from the SVM model, shown together with the historical data (dotted blue) for a) Scenario with a steep decline in production after 2018 and b) Business-as-usual scenario. The vertical dotted-dashed line is on December 31st, 2016, marking the end of the dataset used for training and testing the models. Forecasts left the vertical line are rolling-forward with re-training, forecasts right of the line are without re-training. Seismicity data from January 1995 onwards is used, whereas the first quarter for which a forecast is available is Q2 1995; c) Illustrative comparison of differences in forecasted feature values of the two different reservoir flow simulator scenarios, using a subset of 9 time-dependent features

- H_0: The model performance is identical, $E[r(\mathcal{M}_k, \mathcal{M}_0|\boldsymbol{y})] = 0$, vs
- H_1: Model $\mathcal{M}_i$ is performing better than the baseline, $E[r(\mathcal{M}_k, \mathcal{M}_0|\boldsymbol{y})] > 0$,

with p value computed as above, and effect size $r(\mathcal{M}_k, \mathcal{M}_0|\boldsymbol{y})$.

4.4 Simulation model used in the R-test variant in section 4.3

The simulation model used as a null baseline in the Section 4.3 hypothesis test is assumed bin-wise Poisson with a discrete Poisson intensity $\theta(s,t)$ in the bin with index (s,t). This rate is estimated by fitting a GAM of the form,

$$\log(\theta(s,t)) = f_1(t) + f_2(s), \tag{7}$$

where $f_1(t)$ and $f_2(s)$ are smoothly varying spline functions. This model 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. Full details of this fitting scheme and smoothness estimation can be found in [68]. Figure 6 shows how the spatial fit of the simulation model compares to the observed counts aggregated over time.

5 Illustrative pipeline results

This section shows some illustrative pipeline results from the ML pipeline as sketched in Fig. 2 and explained in Sections 2–4. The general model performance is discussed in Section 5.1 and the forecasts are illustrated in Section 5.2.

5.1 Forecast model performance

Using predictor and seismicity rate target data from January 1, 1995 to December 31, 2016, the ML pipeline is deployed with the aim of forecasting seismicity for the period from January 1, 2017 to December 31, 2024. The experimental setup is summarized in Table 5. For the forecasting period, reservoir model data is used from the two production scenarios (Fig. 4). The performance of the trained ML models is quantified using the error metrics discussed in Section 4, i.e. the MAE, RMSLE, Mean Poisson Loss and likelihood ratio (Fig. 7; Table 6), using the models as discussed in Section 3.4 (Table 3) and the baselines as discussed in Section 3.5. The respective meta- and hyper-parameters have been obtained using the procedure that has been explained in Section 3.6. The performance metrics show that the SVM and RF models perform better than other models (Fig. 7), and that these models outperform the baselines (Table 6).

5.2 Qualitative inspection of long-term temporal forecasts

We proceed with qualitatively inspecting long-term forecasts – note that the performances reported in Section 5.1 are indicative only of short-term, rolling forecasts, into a future where there are no major regime shifts (e.g., under stationarity). Therefore, long-term forecasts are at best only indicative, without statistical guarantees attached. We also note that RF is a non-extrapolating model, i.e. this model cannot forecast

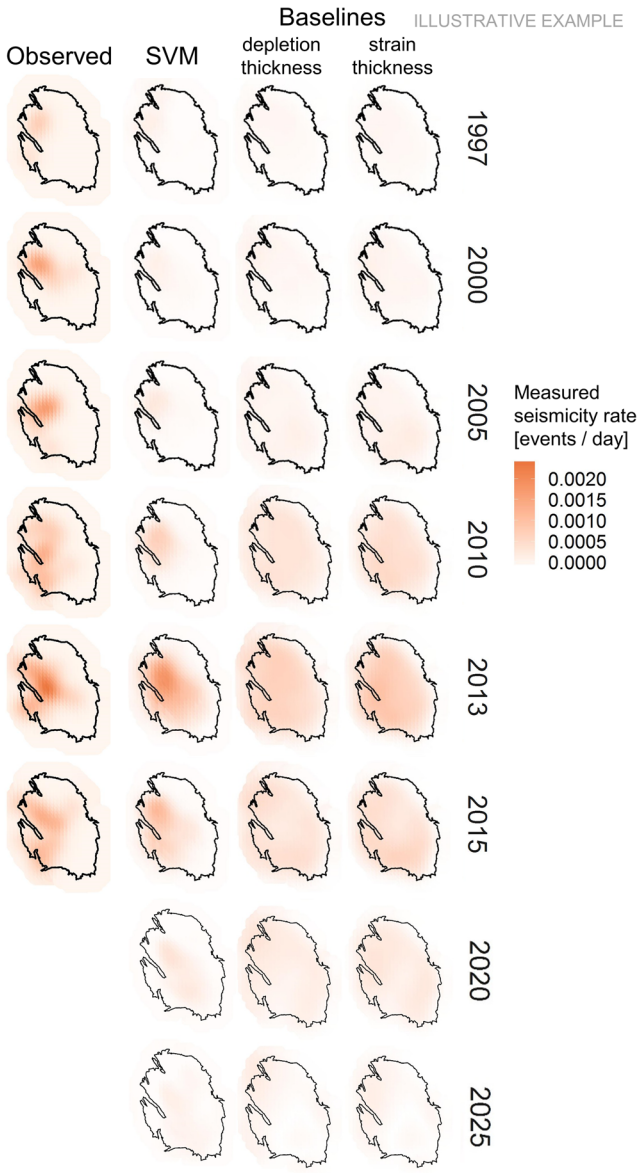

Fig. 9 Model forecasts of the number of events per day, averaged per year, for SVM (2nd column) and two baselines (3rd and 4th columns) compared to observed event rates (1st column) for scenario with steeply declining production rates after 2018

outside the range of calibration for the forecast target. Long-term forecasts from the RF model are qualitatively unphysical upon inspection, therefore we show only SVM long-term forecasts.

The SVM model results are analysed using a field-wide seismicity rate per 3-months aggregation period and shown together with historical data or baselines in Fig. 8. We emphasize that the forecasts are only indicative as the validation setting encompasses only short-term forecasts; even in the case of short-term forecasts, empirical confidence bands on rate predictions are typically in the range of ±70%. This situation may seem paradoxical – but it should be noted that it may well be possible that average performance differences (as presented in Section 5.1) are significant, while individual predictions remain highly uncertain. In summary, the confidence bands are very large, and not plotted in the indicative forecasts; trends, behaviours, or shapes are not subject to any statistical confidence and should hence not be relied on. As we can see in Fig. 8, up to December 2016, both SVM and observed seismicity show a continued increase in seismicity rate around 2008, and a decrease following 2012. The SVM does not capture extremes in the historical data, which is as expected, due to the typical variation in the underlying point process mechanism – as the forecasts are rates, and observations are samples.

Looking to the long-term forecasting period from 2017 onwards, the SVM model forecasts for two production scenarios appear visually different: seismicity rate appears to decline in the scenario with strongly decreasing production rates, whereas a stable or increasing rate may be seen in the 'business-as-usual' scenario. As discussed above, the two long-term forecasts come without any correctness guarantees, and are also not statistically different from each other.

5.3 Forecast model performance

We also inspect SVM model forecasts in their original, spatiotemporal form (Fig. 9). This is subject to the same caveats as before. Qualitatively, and visually, one may see a local difference in seismicity event rates, with the central northwest region of the field at higher rate than other regions, in line with historical observations (Fig. 9). The forecasts show a decrease in seismicity, with the highest relative density remaining in the central-northwest region. A comparison in spatial trends in seismicity between the two investigated production scenarios shows a decrease in seismicity rate over time for the scenario with steeply declining production, whereas this decrease is absent for the other scenario (Fig. 10).

6 Discussion

The ML forecasts for seismicity rates in the period between April 1st, 1995 and December 31st, 2016 for the two investigated production scenarios (Fig. 9) demonstrate that the developed ML pipeline generates results that are significantly better than the investigated baseline models. Spatial

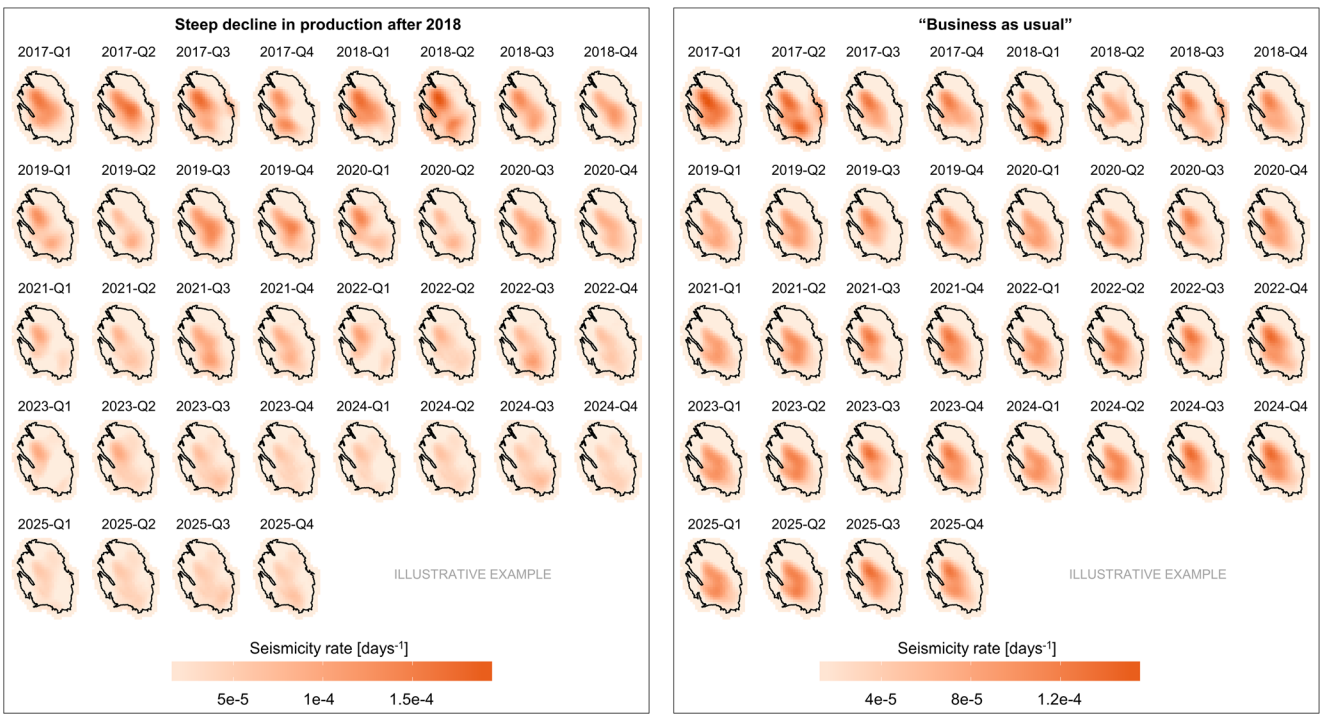

Fig. 10 Comparison of spatial seismic density, expressed as the daily seismicity rate averaged over 3-months intervals, at 3-months time steps between 2017 and 2025 for the two analysed production scenarios: left: Steep decline in production after 2018; right: Results for 'business as usual'

information is also captured, as the models forecast the highest seismicity in the Central-Northwest area and the lowest seismicity in the south and towards the edges of the field, specifically in the southeast corner of the field.

Qualitatively, the future forecasts capture the expected spatial high-density localization of seismicity, as well as the expected stronger decline in seismicity for the scenario where production is significantly limited compared to a 'business as usual' scenario (Fig. 10), although the decline was only observed after the introduction of 'ultimate state' points, representing the reservoir pressure long after production is ceased. The values of these points, as well as the forecasted dynamic reservoir model properties, are the outcome of simulation models, and hence carry uncertainties and modelling assumptions. For the ultimate state points, a limited range of scenarios has been run to assess these uncertainties, but for the forecasts up to 2025 these uncertainties are not accounted for. In the specific case of the Groningen field, the model uncertainties are relatively small as the model is history-matched to decades of production data and the physical behaviour of this gas field is relatively well known. However, for more complex subsurface reservoirs simulation model uncertainties would need to be incorporated into the ML pipeline. Another limitation of using simulation models for the forecasted dynamic reservoir properties is that these forecasts are based on historical data, which is in-sample rather than out-of-sample data.

Additionally, it should be noted that the tuning of model meta and hyper parameters was focused on short term prediction intervals of 3 months. Short- and long-term predictive performance are not necessarily equivalent. Even though it was experimentally shown that the relative ranking of models stays intact independently of a short- or long-term forecasting horizon, potentially better model parameters could be obtained if the optimization would be carried out on a longer forecasting horizon than 3 months [42, 43].

Although there is a relative match with observed trends, the ML models do not capture most of the extreme values that are observed in the measured seismicity dataset in both time and space. Alternatively, physics-based rules could be derived to constrain ML model behaviour, but the current ML approach provides no option for implementing such rules. To address these limitations, a model for production-induced seismicity could be envisioned that combines the physical mechanisms driving seismicity with ML to address the aspects of seismicity forecasting that cannot be modelled with either deterministic physics or ML alone, in line with what others have suggested for geoscience-related applications [27, 28, 69].

7 Conclusions

We have developed a Machine Learning (ML) pipeline for automated tuning of forecasting methods, model selection and benchmarking, experiment meta-analysis, for rolling short-term forecasts of seismicity rates in space in the Groningen gas field (Netherlands), based on production and reservoir simulator outputs. The framework enables quantifying which methodology is most performant, model-agnostic interpretation of forecasts, and testing of whether a specific data source or variable increases predictive performance or not.

Our results show that seismicity forecasts generated using auto-tuned, sliding-window Support Vector Machine (SVM) and Random Forest (RF) models outperform the physics-informed baselines in this report. Other investigated models were not significantly better than the baselines.

One of the highest performing forecasters from the automated pipeline (sliding window tuned SVM) was then used to produce spatiotemporal long-term forecasts of seismicity in the Groningen gas field, using long-term reservoir simulator forecasts for two distinctly different production scenarios. These forecasts are only indicative as they come without any substantive statistical guarantees (see discussion in Section 5); the SVM model forecasts that seismicity rates decrease in the conservative production scenario and seismicity rates remain constant or slightly increase in the original 'business as usual' production. These forecasts are to be handled with caution as the forecasts between scenarios are not different from each other, subject to any meaningful statistical confidence.

Acknowledgements We are grateful to colleagues from Shell (Peter van den Bogert, Xander Campman, Pandu Devarakota, Hadi Jamali-Rad, Gerard Joosten, Kees Hindriks, Roger Yuan, Rick Wentinck, Alan Wood), NAM (Hermann Baehr, Leendert Geurtsen, Per Valvatne, Assaf Mar-Or, Remco Romijn, Richard Vietje, Clemens Visser, Onno van der Wal) and IBM (Munish Goyal, Stephen Lord, Mo Zhang) for providing their expertise input and for fruitful discussions and review comments. We thank NAM for funding this work and allowing for its publication. We thank the three anonymous reviewers for their useful comments that greatly improved this paper.

Funding Funding was provided by the Nederlandse Aardolie Maatschappij (NAM).

Data availability All data used in this study is publicly available: The seismic data is available through the KNMI Seismic and Acoustic Data Portal and EPOS (European Plate Observatory System). The reservoir models are also available through EPOS. Velocity models can be downloaded from NAM report repository website.

Compliance with ethical standards

Conflicts of interest/competing interest The authors declare that they have no conflict of interest.

Code availability The code is available upon individual request, with guidance on its usage and integration with the data.

Appendix

Table 7 Predictor feature overview

Predictor 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 reservoir 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 for lateral density variations in the Zechstein, caused by the anhydrite floaters.		SeisVint.val.mean
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).	Reservoir 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.	Reservoir geological model (interpolated well data)	gc_vs_wc.val.mean

Table 7 (continued)

Predictor feature name	Description	Data source	Model label
Gas saturation in the aquifer	Gas saturation in the Carboniferous (aquifer), as a potential proxy for lateral variations in the fault friction behaviour.	Reservoir geological model (limited well data, interpolated using kriging without constraint to other features).	sg_carb.val.mean
Porosity	Mean reservoir porosity (weighted vertical average)	Reservoir 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.	Reservoir 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	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 reservoir flow model pressure, reservoir thickness and compressibility.	Calculated from reservoir flow model 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

Appendix 2: Individual Conditional Expectations (ICE) plots

Variable importance plots provide information on which predictor features drive model behaviour and the relative impact of a feature, but they provide no information on how changes in the predictor features impact the target feature. These questions can be partially addressed using Individual Conditional Expectation (ICE) plots that illustrate the average impact of a variable on model response [70]. ICE plots show the marginal response of the model with respect to changes in one predictor feature usually indicated by a thick black line. Additionally, actual data points (black dots) and the model response conditioned to all other predictor features except the one shown on the x-axis of the plot assuming the values of the actual data point are shown in form of thin black lines. How far the model response deviates from a linear response can be estimated by how much a response curve deviates from a straight line. If the individual curves are not approximately parallel and show significantly different behaviour this hints at interaction

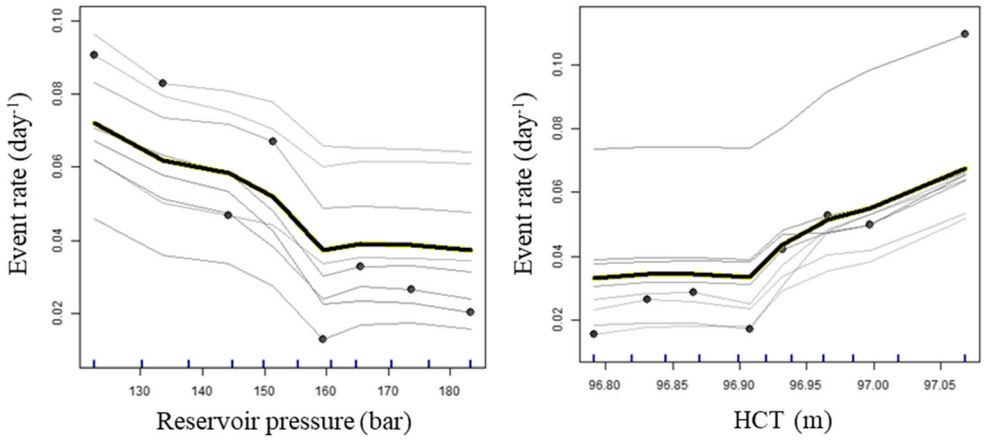

Fig. 11 Illustrative examples of ICE plots for the predictor features reservoir pressure (left) and hydrocarbon column thickness HCT (right). The bold line is equivalent to a partial dependence plot. The black dots correspond to actual data points. Through each of these points passes one thin line which is the model response conditioned to the data of the actual data point excluding the predictor feature on the horizontal axis which is varied in the model

effects, which can subsequently be analysed for instance by creating 2D partial dependence plots. An illustrative example of an ICE plot can be seen in Fig. 11. An overview of the meta-parameters which are considered as part of the pipeline are given in Table 4.

By partitioning the experiments into specific subsets and applying the meta-analysis pipeline steps mentioned above we can address a range of questions including:

- How consistent are the effects of the predictor features and meta-parameters across the different models?
- Can optimal and stable parameter ranges for the significant meta-parameters be established?
- What ranges of model meta-parameters lead to better model performance?
- How sensitive is model performance to the choice of model meta-parameters?

References

1. van Elk, J., Doornhof, D., Bommer, J.J., Bourne, S.J., Oates, S.J., Pinho, R., Crowley, H.: Hazard and risk assessments for induced seismicity in Groningen. Neth. J. Geosci. **96**, s259–s269 (2017). https://doi.org/10.1017/njg.2017.37
2. Bourne, S.J., Oates, S.J., Van Elk, J., Doornhof, D.: A seismological model for earthquakes induced by fluid extraction from a subsurface reservoir. J. Geophys. Res. Solid Earth. **119**, 8991–9015 (2014). https://doi.org/10.1002/2014JB011663
3. Statistics Netherlands (CBS): Natural gas revenues almost 417 billion euros, https://www.cbs.nl/en-gb/news/2019/22/natural-gas-revenues-almost-417-billion-euros. Accessed 31 Jan 2020
4. Foulger, G.R., Wilson, M.P., Gluyas, J.G., Julian, B.R., Davies, R.J.: Global review of human-induced earthquakes. Earth-Sci. Rev. **178**, 438–514 (2018). https://doi.org/10.1016/j.earscirev.2017.07.008
5. Candela, T., Wassing, B., ter Heege, J., Buijze, L.: How earthquakes are induced. Science. **360**(80), 598–600 (2018). https://doi.org/10.1126/science.aat2776
6. van Thienen-Visser, K., Breunese, J.N.: Induced seismicity of the Groningen gas field: history and recent developments. Lead. Edge. **34**, 664–671 (2015). https://doi.org/10.1190/tle34060664.1
7. van Thienen-Visser, K., Pruiksma, J.P., Breunese, J.N.: Compaction and subsidence of the Groningen gas field in the Netherlands. Proc. Int. Assoc. Hydrol. Sci. **372**, 367–373 (2015). https://doi.org/10.5194/piahs-372-367-2015
8. Rutqvist, J., Rinaldi, A.P., Cappa, F., Moridis, G.J.: Modeling of fault activation and seismicity by injection directly into a fault zone associated with hydraulic fracturing of shale-gas reservoirs. J. Pet. Sci. Eng. **127**, 377–386 (2015). https://doi.org/10.1016/j.petrol.2015.01.019
9. Fan, Z., Eichhubl, P., Gale, J.F.W.: Geomechanical analysis of fluid injection and seismic fault slip for the M w 4.8 Timpson, Texas, earthquake sequence. J. Geophys. Res. Solid Earth. **121**, 2798–2812 (2016). https://doi.org/10.1002/2016JB012821
10. Stabile, T.A., Giocoli, A., Perrone, A., Piscitelli, S., Lapenna, V.: Fluid injection induced seismicity reveals a NE dipping fault in the southeastern sector of the high Agri Valley (southern Italy). Geophys. Res. Lett. **41**, 5847–5854 (2014). https://doi.org/10.1002/2014GL060948
11. Izadi, G., Elsworth, D.: Reservoir stimulation and induced seismicity: roles of fluid pressure and thermal transients on reactivated fractured networks. Geothermics. **51**, 368–379 (2014). https://doi.org/10.1016/j.geothermics.2014.01.014
12. Walsh, F.R., Zoback, M.D.: Oklahoma's recent earthquakes and saltwater disposal. Sci. Adv. **1**, e1500195 (2015). https://doi.org/10.1126/sciadv.1500195

13. Van Wees, J.-D., Fokker, P.A., Van Thienen-Visser, K., Wassing, B.B.T., Osinga, S., Orlic, B., Ghouri, S.A., Buijze, L., Pluymaekers, M.: Geomechanical models for induced seismicity in the Netherlands: inferences from simplified analytical, finite element and rupture model approaches. Neth. J. Geosci. **96**, s183–s202 (2017). https://doi.org/10.1017/njg.2017.38
14. Spiers, C.J., Hangx, S.J.T., Niemeijer, A.R.: New approaches in experimental research on rock and fault behaviour in the Groningen gas field. Neth. J. Geosci. **96**, s55–s69 (2017). https://doi.org/10.1017/njg.2017.32
15. Hunfeld, L.B., Niemeijer, A.R., Spiers, C.J.: Frictional properties of simulated fault gouges from the Seismogenic Groningen gas field under in situ P - T -chemical conditions. J. Geophys. Res. Solid Earth. **122**, 8969–8989 (2017). https://doi.org/10.1002/2017JB014876
16. Bourne, S.J., Oates, S.J.: Extreme threshold failures within a heterogeneous elastic thin sheet and the spatial-temporal development of induced seismicity within the Groningen gas field. J. Geophys. Res. Solid Earth. **122**, 299–320 (2017). https://doi.org/10.1002/2017JB014356
17. Orlic, B., Wassing, B.B.T.: A study of stress change and fault slip in producing gas reservoirs overlain by elastic and viscoelastic caprocks. Rock Mech. Rock. Eng. **46**, 421–435 (2013). https://doi.org/10.1007/s00603-012-0347-6
18. Postma, T., Jansen, J.D.: The small effect of Poroelastic pressure transients on triggering of production-induced earthquakes in the Groningen natural gas field. J. Geophys. Res. Solid Earth. **123**, 401–417 (2018). https://doi.org/10.1002/2017JB014809
19. van der Linden, A., Makurat, A., Marcelis, F., Hol, S., Bierman, S.: Rock physical controls on production-induced compaction in the Groningen field. Sci. Rep. **8**, 1–13 (2018). https://doi.org/10.1038/s41598-018-25455-z
20. Mignan, A., Broccardo, M., Wiemer, S., Giardini, D.: Induced seismicity closed-form traffic light system for actuarial decision-making during deep fluid injections. Sci. Rep. **7**, 1–10 (2017). https://doi.org/10.1038/s41598-017-13585-9
21. Broccardo, M., Mignan, A., Wiemer, S., Stojadinovic, B., Giardini, D.: Hierarchical Bayesian modeling of fluid-induced seismicity. Geophys. Res. Lett. **44**(11), 357–11,367 (2017). https://doi.org/10.1002/2017GL075251
22. van Elk, J., Doornhof, D.: Review and Update of: Study and Data Acquisition Plan Induced Seismicity in Groningen - Update Post-Winningsplan 2016, Assen, Netherlands (2019). https://nam-feitenencijfers.data-app.nl/download/rapport/529d284a-a8e9-4aa8-a52e-3aa17761f40d?open=true. Accessed 31 Jan 2020
23. Nederlandse Aardolie Maatschappij: Technical Addendum to the Winningsplan Groningen 2016. (2016). https://www.nam.nl/algemeen/mediatheek-en-downloads/winningsplan-2016.html. Accessed 31 Jan 2020
24. Nederlandse Aardolie Maatschappij: Winningsplan Groningen Gasveld 2016. , Assen, Netherlands (2016). https://www.nam.nl/algemeen/mediatheek-en-downloads/winningsplan-2016.html. Accessed 31 Jan 2020
25. Melnikov, A.A., Nautrup, H.P., Krenn, M., Dunjko, V., Tiersch, M., Zeilinger, A., Briegel, H.J.: Active learning machine learns to create new quantum experiments. Proc. Natl. Acad. Sci. **115**, 1221–1226 (2018). https://doi.org/10.1073/PNAS.1714936115
26. Carrasquilla, J., Melko, R.G.: Machine learning phases of matter. Nat. Phys. **13**, 431–434 (2017). https://doi.org/10.1038/nphys4035
27. Bergen, K.J., Johnson, P.A., de Hoop, M.V., Beroza, G.C.: Machine learning for data-driven discovery in solid Earth geoscience. Science. **363**(8), eaau0323 (2019). https://doi.org/10.1126/science.aau0323
28. Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H.A., Kumar, V.: Machine learning for the geosciences: challenges and opportunities. IEEE Trans. Knowl. Data Eng. **31**, 1544–1554 (2019). https://doi.org/10.1109/TKDE.2018.2861006
29. Pathak, J., Hunt, B., Girvan, M., Lu, Z., Ott, E.: Model-free prediction of large spatiotemporally chaotic systems from data: a reservoir computing approach. Phys. Rev. Lett. **120**, (2018). https://doi.org/10.1103/PhysRevLett.120.024102
30. DeVries, P.M.R., Viégas, F., Wattenberg, M., Meade, B.J.: Deep learning of aftershock patterns following large earthquakes. Nature. **560**, 632–634 (2018). https://doi.org/10.1038/s41586-018-0438-y
31. Perol, T., Gharbi, M., Denolle, M.: Convolutional neural network for earthquake detection and location. Sci. Adv. **4**, e1700578 (2018). https://doi.org/10.1126/sciadv.1700578
32. Rouet-Leduc, B., Hulbert, C., Lubbers, N., Barros, K., Humphreys, C.J., Johnson, P.A.: Machine learning predicts laboratory earthquakes. Geophys. Res. Lett. **44**, 9276–9282 (2017). https://doi.org/10.1002/2017GL074677
33. Panakkat, A., Adeli, H.: Recurrent neural network for approximate earthquake time and location prediction using multiple seismicity indicators. Comput. Civ. Infrastruct. Eng. **24**, 280–292 (2009). https://doi.org/10.1111/j.1467-8667.2009.00595.x
34. Asencio-Cortés, G., Morales-Esteban, A., Shang, X., Martínez-Álvarez, F.: Earthquake prediction in California using regression algorithms and cloud-based big data infrastructure. Comput. Geosci. **115**, 198–210 (2018). https://doi.org/10.1016/j.cageo.2017.10.011
35. Asencio-Cortés, G., Martínez-Álvarez, F., Morales-Esteban, A., Reyes, J.: A sensitivity study of seismicity indicators in supervised learning to improve earthquake prediction. Knowl.-Based Syst. **101**, 15–30 (2016). https://doi.org/10.1016/j.knosys.2016.02.014
36. Last, M., Rabinowitz, N., Leonard, G.: Predicting the maximum earthquake magnitude from seismic data in Israel and its neighboring countries. PLoS One. **11**, 1–16 (2016). https://doi.org/10.1371/journal.pone.0146101
37. Mignan, A., Broccardo, M.: One neuron versus deep learning in aftershock prediction. Nature. **574**, E1–E3 (2019). https://doi.org/10.1038/s41586-019-1582-8
38. Mignan, A., Broccardo, M.: Neural Network Applications in Earthquake Prediction (1994–2019): Meta-analytic and statistical insights on their limitations. Seismol. Res. Lett. **1–25**, 2330–2342 (2020). https://doi.org/10.1785/0220200021
39. Meade, B.J.: Reply to: one neuron versus deep learning in aftershock prediction. Nature. **574**, E4–E4 (2019). https://doi.org/10.1038/s41586-019-1583-7
40. Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D.: Do we need hundreds of classifiers to solve real world classification problems? J. Mach. Learn. Res. **11**, 015020 (2014). https://doi.org/10.1117/1.JRS.11.015020
41. Makridakis, S., Spiliotis, E., Assimakopoulos, V.: Statistical and machine learning forecasting methods: concerns and ways forward. PLoS One. **13**, e0194889 (2018). https://doi.org/10.1371/journal.pone.0194889
42. Limbeck, J., Lanz, F., Barbaro, E., Harris, C., Bisdom, K., Park, T., Oosterbosch, W., Jamali-Rad, H., Nevenzeel, K.: Evaluation of a Machine Learning methodology to forecast induced seismicity event rates within the Groningen Field, Assen, Netherlands (2018). https://nam-feitenencijfers.data-app.nl/download/rapport/d5be89f6-fcea-4237-bc07-6cda25e151d9?open=true. Accessed 31 Jan 2020
43. Lanz, F., Bisdom, K., Barbaro, E., Limbeck, J., Park, T., Harris, C., Nevenzeel, K.: Evaluation of a Machine Learning methodology for spatiotemporal induced seismicity forecasts within the Groningen field, Assen, Netherlands (2019). https://nam-onderzoeksrapporten.data-app.nl/reports/download/groningen/en/e5535713-46e2-4523-a479-4124f674c55f. Accessed 31 Jan 2020
44. R Core Team: R: A Language and Environment for Statistical Computing, https://www.r-project.org/, (2018)

45. Wright, M.N., Ziegler, A.: Ranger : A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J. Stat. Softw. **77**, 1–17 (2017). https://doi.org/10.18637/jss.v077.i01
46. CGAL Project: CGAL User and Reference Manual, (2018). https://www.cgal.org/
47. Gärtner, B., Schönherr, S.: 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, NY, USA, 110–118 (2000). https://doi.org/10.1145/336154.336191
48. Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G., Jones, Z.M.: mlr: Machine Learning in R. J. Mach. Learn. Res. **17**, 1–5 (2016)
49. van Oeveren, H., Valvatne, P., Geurtsen, L., van Elk, J.: History match of the Groningen field dynamic reservoir model to subsidence data and conventional subsurface data. Neth. J. Geosci. **96**, s47–s54 (2017). https://doi.org/10.1017/njg.2017.26
50. Bierman, S., Kraaijeveld, F., Bourne, S.: Regularised Direct Inversion to Compaction in the Groningen Reservoir Using Measurements from Optical Leveling Campaigns. Tech. Report. Shell Glob. Solut. Int. (2015). https://nam-feitenencijfers.data-app.nl/download/rapport/cc5ea278-c093-457b-b930-1869a3c26c21?open=true. Accessed 31 Jan 2020
51. Burkitov, U., van Oeveren, H., Valvatne, P.: Groningen Field Review 2015 Subsurface Dynamic Modelling Report. (2016)
52. Ministry of Economic Affairs and Climate Policy: Kamerbrief over gaswinning Groningen (2018) https://www.government.nl/documents/parliamentary-documents/2018/03/29/kamerbrief-over-gaswinning-groningen. Accessed 1 July 2019
53. Rydelek, P.A., Sacks, I.S.: Testing the completeness of earthquake catalogues and the hypothesis of self-similarity. Nature. **337**, 251–253 (1989). https://doi.org/10.1038/337251a0
54. Dost, B., Goutbeek, F., van Eck, T., Kraaijpoel, D.: Monitoring Induced Seismicity in the North of the Netherlands: Status Report 2010, De Bilt, Netherlands (2012). KNMI report: http://bibliotheek.knmi.nl/knmipubWR/WR2012-03.pdf. Accessed 31 Jan 2018
55. Bourne, S.J., Oates, S.J.: Development of statistical geomechanical models for forecasting seismicity induced by gas production from the Groningen field. Geol. Mijnbouw/Neth. J. Geosci. **96**, s175–s182 (2017). https://doi.org/10.1017/njg.2017.35
56. Nelder, J.A., Wedderburn, R.W.M.: Generalized linear models. J. R. Stat. Soc. Ser. A. **135**, 370 (1972). https://doi.org/10.2307/2344614
57. López-Ibáñez, M., Dubois-Lacoste, J., Pérez Cáceres, L., Birattari, M., Stützle, T.: The irace package: iterated racing for automatic algorithm configuration. Oper. Res. Perspect. **3**, 43–58 (2016). https://doi.org/10.1016/j.orp.2016.09.002
58. Hu, L.Y., Huang, M.W., Ke, S.W., Tsai, C.F.: The distance function effect on k-nearest neighbor classification for medical datasets. Springerplus. **5**, 1304 (2016). https://doi.org/10.1186/s40064-016-2941-7
59. Breiman, L.: Random Forrests. Mach. Learn. (2001)
60. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer New York, New York (2009)
61. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**, 273–297 (1995). https://doi.org/10.1007/BF00994018
62. Kursa, M.B., Rudnicki, W.R.: Feature Selection with the Boruta Package. J. Stat. Softw. **36**, i11 (2010). https://doi.org/10.18637/jss.v036.i11
63. Fagerland, M.W., Sandvik, L.: The Wilcoxon-Mann-Whitney test under scrutiny. Stat. Med. **28**, 1487–1497 (2009). https://doi.org/10.1002/sim.3561
64. Schorlemmer, D., Gerstenberger, M.C., Wiemer, S., Jackson, D.D., Rhoades, D.A.: Earthquake likelihood model testing. Seismol. Res. Lett. **78**, 17–29 (2007). https://doi.org/10.1785/gssrl.78.1.17
65. Gerstenberger, M., Rhoades, D., Stirlin, M., Brownrigg, R., Christophersen, A.: Continued Development of the New Zealand Earthquake Forecast Testing Centre. (2009). https://www.eqc.govt.nz/sites/public_files/3753-Development-NZ-EQ-Forecast-Centre.pdf. Accessed 31 July 2020
66. Rhoades, D.A., Schorlemmer, D., Gerstenberger, M.C., Christophersen, A., Zechar, J.D., Imoto, M.: Efficient testing of earthquake forecasting models. Acta Geophys. **59**, 728–747 (2011). https://doi.org/10.2478/s11600-011-0013-5
67. Bray, A., Schoenberg, F.P.: Assessment of point process models for earthquake forecasting. Stat. Sci. **28**, 510–520 (2013). https://doi.org/10.1214/13-STS440
68. Wood, S.N., Pya, N., Säfken, B.: Smoothing parameter and model selection for general smooth models. J. Am. Stat. Assoc. **111**, 1548–1563 (2016). https://doi.org/10.1080/01621459.2016.1180986
69. Karpatne, A., Atluri, G., Faghmous, J.H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., Kumar, V.: Theory-guided data science: a new paradigm for scientific discovery from data. IEEE Trans. Knowl. Data Eng. **29**, 2318–2331 (2017). https://doi.org/10.1109/TKDE.2017.2720168
70. Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E.: Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. J. Comput. Graph. Stat. **24**, 44–65 (2015). https://doi.org/10.1080/10618600.2014.907095