WP3: Proposal for an alternative seismogenic source representation

Review of seismogenic source models for the Groningen gas reservoir

April 2020

GFZ

Helmholtz-Zentrum
POTSDAM

NORSAR

Contents

1 Executive summary **3**

2 Introduction **4**

3 Potential of models to deal with production stop **6**

4 Testing of seismicity models **10**

4.1 CSEP . . . 11

4.1.1 Test metrics . . . 11

4.2 CSEP2.0 . . . 13

4.3 CSEP-tests for induced seismicity . . . 13

5 Proposal **15**

5.1 Project objectives . . . 15

5.2 Scope of work . . . 15

5.3 Potential for publications . . . 19

5.4 Collaboration between NORSAR and GFZ . . . 19

5.5 Price . . . 20

5.5.1 Personnel costs . . . 20

5.5.2 Costs per task . . . 20

5.6 Schedule and key deliverables . . . 21

5.6.1 Deliverables . . . 21

5.7 Quality assurance . . . 21

5.8 Commercial qualifications . . . 22

Appendix A Literature **24**

1 Executive summary

Modelling seismicity after a production stop is a big challenge for seismicity models. The reason is that the dominant forces change, and in consequence, the correlations observed may also change. Based on our present knowledge, we can only speculate about what model constitutes the best alternative to forecast seismicity in the Groningen field during the following years when production is planned to stop, as well as during the subsequent decades. In part, this is a consequence of our limited knowledge of the physical processes taking place in and properties of the Groningen field. But more importantly, the performance of the present models, differing in processes and assumptions, is difficult to assess, because no systematic comparative tests have been performed so far.

Seismicity models provide forecasts of average rates and probabilities of earthquake activity rather than deterministic predictions of the occurrence of single events. Hence, the absence or occurrence of single events can never verify or falsify such models, and more sophisticated test metrics have to be applied to judge their quality. Moreover, because alternative models are often suggested by different research teams, comparative metrics are required to statistically determine whether one model performs better than another over an evaluation period. Properly done, such tests can show the consistency of a model with the observations. Although all published models have provided some set of such statistical tests, some were tested for the time-dependency of the overall seismicity, while others were tested for spatio-temporal patterns. Further, due to the complexity of most models and the fact that some of them were specifically developed and adapted to the Groningen field, it remains unclear whether or not hidden parameters were set based on the knowledge of the full data set. All these factors render an objective comparison using only published information impossible.

The only way to discriminate between potentially useful forecasts and to assess the hypotheses on which they are based is to evaluate their associated models using rigorous and transparent forecasting experiments. An example of such systematic assessment for models that forecast natural seismicity can be found in the efforts of the Collaboratory for the Study of Earthquake Predictability (CSEP). However, not much work has been done so far on CSEP-type model selection and comparison in the context of induced seismicity. Compared to the full cyber-infrastructure developed in the frame of CSEP, a simpler, less demanding approach consists of CSEP-tests running on a web-interface. This type of online toolbox would enable modellers to run their codes independently but employing identical input data. Then, based on uniform input and test data, this would also allow for meaningful model comparisons and rankings. In WP3 (this document), we propose a project consisting of three work packages where such online toolbox can be developed for the testing of models that forecast induced seismicity in Groningen. An additional benefit of such toolbox is that it can be extended to be used elsewhere in other similar types of projects (e.g., geothermal fields) and to incorporate additional variables (e.g., event locations, moment tensor solutions).

2 Introduction

The Groningen gas reservoir is located in the north-east of the Netherlands at a depth of approximately 3000 m. The presence of gas in Groningen was discovered in 1959 with the start of production taking place in 1963. Since the beginning, interest was placed in the monitoring of subsidence in the field as this had already been linked to oil and gas production in the past with undesirable effects (e.g., Mayuga and Allen, 1969; Yerkes and Castle, 1969).

The first seismic event related to gas extraction in Groningen was detected in 1991 (Van Eck et al., 2006). From that moment on, the sustained observation of seismic activity prompted the need for the development of seismic hazard analyses. As part of these analyses, seismogenic models that take into account the conditions of a producing reservoir such as Groningen had to be developed.

A seismogenic model is aimed at forecasting the distribution, rate and magnitudes of seismicity. Leading forecast models for natural seismicity rely on statistical stationarity, such that forecasts are based on historical frequencies (e.g., European Committee for Standardization, 2005). However, induced seismicity requires a fundamentally different approach, since the operational activities causing or triggering seismicity are not stationary in time and may even be adapted to reduce the seismic hazard (Bourne et al., 2018).

The origins of the current NAM seismogenic model for Groningen can be traced back to the work of Bourne et al. (2014). This initial model (V0) highlighted a connection between compaction and seismicity in Groningen, which was represented via an empirical factor called strain partitioning. A subsequent version of the model (V1; Bourne and Oates, 2014) was reformulated in terms of a Poisson Point Process (PPP), where the relationship between the rate of seismic activity and compaction defined the PPP through an intensity function. Subsequent versions of the seismogenic model (V2 and V5) continued with the PPP formulation, although with intensity functions that incorporated more and more sophisticated physical constraints.

A constant observation made throughout the development of the NAM seismogenic model was the lack of sufficient data to constraint its parameters. This limitation became more accentuated as the number of model parameters increased with the model sophistication. Additional challenges present for all seismogenic models that apply to Groningen are related to the accuracy limits of the different types of input data (e.g., location accuracy of event locations, accuracy of subsidence measurements, extrapolation of rock properties away from control points, etc.) and other related assumptions (e.g., compaction models, relationship between compaction and subsidence). A review of the evolution of the NAM seismogenic model and the different elements that affect its potential to produce reliable forecasts was the subject of WP1 of this project (Dahm et al., 2020a).

The NAM seismogenic model is of hybrid type, based on a physical model, but combined with parameter uncertainties resulting in forecasts that are reported in terms of probabilities. Other alternative

models that can be relevant for Groningen exist in the literature. These can be classified into: (1) deterministic physical models, (2) machine learning approaches, (3) purely statistical and empirical models, and (4) so-called hybrid models integrating physical and statistical approaches. WP2 of this project (Dahm et al., 2020b) describes a collection of more than 20 seismogenic models that fit into one of these categories. The description covers their strengths and weaknesses, but also highlights the difficulty of elaborating a fair and objective comparison between them and with the NAM model if only the results published by the respective authors are at one's disposal.

Therefore, this WP3 report ultimately proposes the continuation of the current project through the development of a publicly available toolbox for testing and comparing different seismogenic models. An important benefit of such toolbox is that it would provide standardised input data sets and evaluation metrics that would allow an objective comparison between sundry models to forecast seismicity as the Groningen field enters its final phase of production-stop.

Thus, the first part of this report discusses the potential of different types of seismogenic models to produce reliable forecasts over a period of production stop, but more importantly, it elaborates on the challenge of comparing alternative seismogenic models having access only to published results obtained under different conditions. In the second part, it is observed that the same challenge has been recognised in the forecasting of natural seismicity. In this case, an ongoing initiative to compare between alternative approaches to forecast natural seismicity is described, as it can serve as a template to develop similar tools for induced seismicity. Thus, the final part of the report presents the details of our proposal. These include three work packages related to the development and testing of the above-mentioned toolbox, related personnel costs and other relevant associated elements.

3 Potential of models to deal with production stop

Modelling the seismicity after production stop is a big challenge for seismicity models, this is because the dominant forces change and in consequence the observed correlations may also change. A shift in the prominent mechanisms can take place and new processes can be triggered. In the case of Groningen, the significant production reduction in 2014 was followed by a clear decrease in seismicity, but still larger earthquakes of magnitudes $M \geq 3.0$ occurred. This raises the question how a production stop will influence the future seismic energy release. One possibility is that the earthquake activity immediately stops, because the system is stabilised. In another, more likely case, the seismic activity continues due to e.g. stress corrosion reducing the strength of faults, ongoing aftershock triggering, or ductile processes leading to a continuing stress buildup. In particular, the production in Groningen significantly changes the stress regime in the region covered by the gas field, which results in interaction with adjacent formations until a new steady-state condition in the stress regime is reached. Here, the soft salt layers can act either as buffer or distributor of stress. In any case, such ductile processes are slow and might influence the stress evolution over a long period.

It is worth noting that several models expect stress changes or volume changes as input parameters, which in turn are output from independent models such as the NAM dynamic reservoir model. Therefore, the ability of these secondary models to account for the frictional response of the overburden comprising viscoelastic thick salt layers and the more rigid underburden containing preexisting faults are crucial for the success of modelling the seismicity after a production stop.

Ability of physical models to deal with the production stop

The results of the physical models help to identify the dominant mechanisms and outline the necessary parameter space that should be considered in the seismogenic models. In the following, abbreviations given in **bold** refer to the description of seismogenic models in Dahm et al. (2020b). The generic poroelastic model (**basicRes**) of Segall (1989) and the extended model (**basicRes+fault**) of Jansen et al. (2019) can be used to identify in which distance range poroelastic pressure diffusion can occur. Whereas the 3D FEM model (**3D-Res**) of Sanz et al. (2015) may provide the magnitude of cumulative slip that is expected, the results of the **basicRes+salt** model of Kettermann et al. (2017) are necessary to obtain information on rheological changes in case of more pronounced viscoelastic material behaviour. Such behaviour may lead to changes in the frequency-magnitude distribution of the generated seismicity and to reactivation of other faults due to salt intrusion. The influence of the fluid flow modelled by Zbinden et al. (2017) (**basicRes+flow**) will be another important factor in the stress modelling after production stop. Buijze et al. (2019) found in their **basicRes+dynRup** model that the rupture of faults with small offsets tends to propagate into the underburden; this means that the seismicity models should be capable to include seismicity migrating into the underburden. The influence of non-elastic friction responses of the fault system (e.g. rate-and-state) is already im-

plemented in some models that describe the seismicity of the whole gas field. Therefore, the results of the **basicRes+quasidynRup-RS** (DeDontney and Lele, 2018) and **basicRes+quasistaticRup** (Van Wees et al., 2017) models describing only parts of the complex interaction of the fault system are already implemented.

Ability of statistical and hybrid models independent of NAM's dynamic reservoir model dealing with the production stop

The statistical and hybrid seismicity models can be divided into two groups: one based on NAM's dynamic reservoir model and the other group relying on other input data. The models that are independent from NAM's dynamic reservoir model are (1) the **K2/K3** model (Helmstetter and Werner, 2012; Helmstetter and Werner, 2014), (2) the pure **ETAS** model (Ogata, 1998), both based solely on the observed seismicity, (3) the **NV-models** of Hettema et al. (2017) and Vlek (2019) and (4) the seismogenic index model (**SI**; Shapiro et al., 2010; Shapiro, 2018). The models (3) and (4) are based on the extracted gas volume. The other ten models discussed in Dahm et al. (2020b) are based on the results of NAM's dynamic reservoir model. It could be an advantage of a model to be independent of the NAM dynamic reservoir model if the model assumption on the input parameters do not change after production stop, but e.g., the **K2/K3** model requires stable seismicity patterns for a reliable prediction. This will obviously not be true for the time period until and after a production stop. Here an adapted dynamic reservoir model to a changing rheology would provide stable input data. The **ETAS** model on the other hand assumes the stress interaction of the seismic events as driving force, which has only been of minor importance in the data set until now (Bourne et al., 2018; Muntendam-Bos, 2020). Therefore, this model should only be used in combination with another seismicity model. The models based on extracted gas volume pose the opportunity to obtain an estimate of the total stress amount accumulated by production independently of the assumptions on which the NAM dynamic reservoir model is based. Both models (**NV-model** and **SI**), though, assume an instantaneous seismic response to volume changes given by the extracted gas volume resulting in a sudden cessation of seismicity after production stop.

Ability of statistical and hybrid models based on NAM's dynamic reservoir model dealing with the production stop

Comparing the ten models based on the results of the NAM dynamic reservoir model, the **NC-model** of DeDontney (2017) resulted in a wide range of activity rates for the tested forecast and could not discriminate the best relation of compaction or slip to seismicity rates. This will be especially the case if the reaction on the production stop will be analysed. The machine learning model (**ML**) of Lanz et al. (2019) was adapted to model a shut-in scenario by setting an ultimate state of zero activity if all time-derivatives of the features are zero. While the consideration of such an ultimate state is reasonable, it is likely not sufficient to guarantee reliable forecasts. In general, a limitation in the predictive power

of the machine learning model is set by conditions outside of the training parameter-set range as e.g. the time after a production stop. The model of **Wentinck** (Wentinck, 2015) showed a significantly higher event rate than observed after pressure reduction stopped in 2005 for the application to the Eleveld field. This will be the case for the Groningen field as well after a production stop, because the induced stresses are only released by the stress drop of the earthquakes. Inelastic mechanisms altering the induced stress, e.g. within the salt layers, are not considered.

The three modifications of the NAM model, namely the 'linear' **NAM** model V5, the viscoelastic model **visco-NAM** and the rate-and-state model **NAM-RS** (Bourne and Oates, 2018) have in common the underlying thin-sheet extreme threshold failure model. This model results in increasing (doubling) seismicity rates for every increment of strain in the long term, which constitutes an overestimation when the extreme value theory does not hold anymore. In this case, a transition to a linear dependence on Coulomb stress increments is expected (Bourne and Oates, 2018) that was not observed yet. Hence, the model may provide an upper limit for the seismicity forecast. The viscoelastic and rate-and-state modification can identify long term effects, but the extreme threshold failure model and the fitting period from 1995 to 2018 are limiting their influence. The model of Dempsey and Suckale (2017) (**CMsub**) assumes an instantaneous coupling of stress changes to frictional failure on the faults inside the reservoir layer similar to the NAM model. On the other hand, the rupture propagation and consequently the magnitude calculation is based on an alternative approach to the commonly used Gutenberg-Richter model, which provides another perspective on the seismic response after shut-in. The reduction of seismicity in 2014 was reproduced by the model, but the interaction with the adjacent formations is not implemented. Such interaction can be considered in the NAM dynamic reservoir model, in case that the seismicity remains restricted to the reservoir layer. The model of Candela et al. (2019) (**Candela-RS**) may provide important information on the response of the stress field and accordingly, the seismicity on the production stop within the two analysed subregions (one in the central region where most seismicity occurs and the other in the SW region where the seismic response is weaker). These results can be used to improve the stress models that cover the whole gas field. The model of Richter et al. (2020) (**Richter-RS**) is another alternative model providing the rate-and-state frictional response of the fault system that is independent from the extreme threshold failure model. The model enables to easily implement new parameters and test their impact. These are all properties valuable for the application after production stop.

The most promising models to provide reasonable seismicity estimates for the time after a production stop are the **NAM** model, the **CMsub** model, the **Richter-RS** model and the **ML**. Comparing the results of all these models may provide a first impression of the seismicity after a production stop. In any case, the adaptation of the models to seismicity occurring outside of the gas fields (laterally and vertically) is likely necessary to provide improved forecasts. Non-elastic processes may partially be implemented through the NAM dynamic reservoir model, but the implementation of other effects like delayed rate-and-state frictional failure of faults or ongoing aftershock activity might be similarly

important for reliable seismicity forecasts after production shut-in.

Based on our present knowledge, we can only speculate about what model constitutes the best model to forecast seismicity in the Groningen field during the next years and decades. This is partly due to our limited knowledge of the physical processes taking place in and properties of the Groningen field. However, more importantly, the performance of the present models, which differ in processes included and assumptions, is difficult to compare because no systematic comparative tests have been performed so far.

4 Testing of seismicity models

Seismicity models provide forecasts of average rates and probabilities of earthquake activity rather than deterministic predictions of the occurrence of single events. Instead of yes/no answers, they can only provide probabilities of observing an earthquake within some predefined space-time-magnitude region. Thus, the absence or occurrence of single events can never verify or falsify such models, and more sophisticated test metrics have to be applied to judge their quality. Here, likelihood-based evaluation metrics are necessary, which evaluate whether observed seismicity is statistically consistent with the forecasts. Additionally, because usually alternative models are suggested by different research teams, comparative metrics are required to statistically determine whether one model performs better than another over an evaluation period.

Usually, model developers carefully evaluate and test their seismicity models before publication. This holds also for the Groningen case, where each of the models was compared and tested with existing data by the corresponding authors, partly employing sophisticated test procedures. For example, Bourne et al. (2018) performed pseudo-prospective tests of the NAM-model ensemble forecasts for the period 2012–2017 using the joint posterior distribution of parameter values estimated from the period 1995–2012. Properly done, such tests can show the consistency of the model with the observations. However, some models were tested for the time-dependency of the overall seismicity, while others were tested for spatio-temporal patterns. Also, due to the complexity of most models and the fact that a part of the models was specifically developed and adapted for the Groningen field, it remains unclear whether or not hidden parameters were set based on the knowledge of the full data set. Such hidden model optimisations often occur without explicit perception of the authors. Moreover, individual tests cannot answer the question of whether or not one model is superior to other alternative models. This is because such comparative tests are hampered by the fact that (i) often not all information concerning the tests is provided; (ii) input data are diverse and not provided; and (iii) test data and quantities differ in various publications. Thus, usually no quantitative comparisons of the forecast ability can be performed based on publications alone.

Due to these difficulties, the only way to discriminate between potentially useful forecasts and to assess the hypotheses they implement is to evaluate them using rigorous and transparent forecasting experiments. In the context of natural seismicity, the Collaboratory for the Study of Earthquake Predictability (CSEP) has established philosophies and cyber-infrastructure to conduct earthquake forecasting experiments in an unbiased and transparent fashion (Jordan, 2006; Schorlemmer et al., 2018). This can be employed as a template for testing seismicity in the Groningen field or other anthropogenic earthquake activity.

4.1 CSEP

CSEP experiments use predefined rules to test forecasts against observations using a number of different evaluation metrics (Jordan, 2006; Jordan et al., 2011; Michael and Werner, 2018; Schorlemmer and Gerstenberger, 2007; Schorlemmer et al., 2018). The CSEP test metrics evaluate whether the observed seismicity is statistically consistent with the forecasts. Additionally, they include comparative metrics to statistically determine whether one model performs better than another over an evaluation period.

Within the CSEP framework fully prospective regional tests are performed, where seismicity models are tested in independent testing centres against new earthquake data that are unknown during model development and adjustment. Currently, CSEP conducts prospective experiments in five testing centres around the world and more than 400 models are under evaluation (Michael and Werner, 2018). GFZ is one of those testing centres, where tests for Japan as well as new tests for Italy and Europe are currently conducted. While such fully prospective tests provide the most objective and rigorous results, they require a long-term effort and cyber-infrastructure which is sustained for years and decades. Furthermore, significant test results are only available with a few years' delay, earliest at the end of the forecast period. Therefore, within the CSEP framework comparative pseudo-prospective experiments are conducted additionally. In this case, the analysed models are tested for a sequence of test periods, where models are only provided with data available before a certain time period and evaluated against the observed events within those periods. Such tests have been conducted with homogeneous input and test data as well as test metrics e.g. within the EU-projects SAFER and REAKT for the Landers aftershock sequence in California and the Canterbury sequence in New Zealand, respectively (Woessner et al., 2011; Cattania et al., 2018). The GFZ-group participating in this project and proposal was involved in these experiments as well as in the ongoing EU-project RISE, where a similar test is planned for central Italy with new test procedures described in subsection 4.2.

4.1.1 Test metrics

All technical details of the CSEP-tests can be found e.g. in Zechar et al. (2010) and Rhoades et al. (2011). Here, we summarise them only briefly:

1. **N-test**
   The N-test is intended to measure how well the total forecasted number of earthquakes in the whole region matches the observed number of events. Because the forecast rate in each bin is considered independently and Poisson-distributed, the overall forecast rate is also Poissonian. Then, one calculates the probabilities δ_1 and δ_2 of observing (1) at least and (2) at most N_{obs} events given a forecast number N_{fore}, respectively. These probabilities can be written as $\delta_1 = 1 - F(N_{obs} - 1|N_{fore})$ and $\delta_2 = F(N_{obs}|N_{fore})$, where $F(x|\mu)$ is the right-continuous Poisson cumulative distribution function with mean μ evaluated at x. To maintain a Type I error rate of

$\alpha = 5\%$, both δ_1 and δ_2 have to exceed a critical value of 0.025.

2. **L-test**
   The L-test (likelihood test) considers the entire space-rate-magnitude forecast. It addresses the question: what is the likelihood of the observation assuming that the forecast is correct? Because the forecast in each space-time-magnitude bin follows a Poisson distribution, the log-likelihood is, by definition, $L(N_{obs}|N_{fore}) = -N_{fore} + N_{obs}\log(N_{fore}) - \log(N_{obs}!)$. Because the rates in each bin are assumed to be independent, the joint log-likelihood (for all bins) of the observation given the forecast is simply the sum over all bins. The observed joint log-likelihood has a negative value, and values that are closer to zero indicate a more likely observation; in other words, such a value indicates that the forecast shows better agreement with the observation. To account for forecast uncertainty, catalogues that are consistent with the forecast are simulated and the corresponding L-values of these simulations calculated. Then, the L-value for the observed events is compared with the distribution of L-values for the simulated catalogues. If it falls in the lower tail of the distribution, it indicates that the observation is not consistent with the forecast. The accurracy is quantified by the quantile score, which is the fraction of simulated joint log-likelihoods less than or equal to the observed joint log-likelihood.

3. **M-test**
   The objective of the M-test (magnitude test) is to consider only the magnitude distributions of the forecast and the observation. To isolate these distributions, the forecasts are summed over all bins and normalised such that the sum matches the observation. Then, the joint log-likelihood for the observed total number of events in the different magnitude bins is calculated given the forecast numbers in these bins. Similarly to the L-test, a quantile score is calculated for random simulations in agreement with the forecast model.

4. **S-test**
   The S-test (space test) analyses the spatial forecast. It is the spatial equivalent of the M-test, where now only the spatial distributions of the forecast and the observation are considered. Similar to the M-test, the spatial information is considered by summing and normalising, which removes the effect of the rate and magnitude components of the original forecast. Again, the corresponding log-likelihood is compared to the distribution of simulated joint log-likelihoods and a corresponding quantile score is calculated.

5. **T-test**
   The T-test is a comparative test between models which examines the information gain $I_{A,B}$ of a model A with respect to a baseline model B. In particular, the information gain is given by the average difference of the log-likelihood values, $I_{A,B} = (L_A - L_B)/N_{obs}$. Loosely speaking, $\exp(I_{A,B})$ is the per-event average of the ratio between the likelihoods of both models. By means of the t-distribution, it is tested whether or not $I_{A,B}$ differs significantly from zero. If it

does, the model with the lower likelihood could be rejected in favour of the other.

4.2 CSEP2.0

The tests developed during the first phase of CSEP (2007-2017; Schorlemmer et al., 2007; Werner et al., 2011) rooted in the concepts that 1) the earthquake activity is independent in different space-time-magnitude bins, 2) earthquakes follow the Poisson distribution in each bin, and 3) modelers provide the *true* parameter of the distribution in each bin. Thus, CSEP-tests required only forecasts in form of expected number of earthquakes in discrete space-time-magnitude bins. This simplification allows multiple types of models, including those without explicit likelihood functions, to participate in the experiments. However, evaluations can incorrectly reject forecasts when the likelihood function of a forecast does not match a Poisson distribution. For example, the epidemic-type aftershock sequence model (ETAS) is overdispersed with respect to a Poisson process causing forecasts to be incorrectly rejected (Lombardi and Marzocchi, 2010; Nandan et al., 2019). To account for this issue, the new CSEP2.0 framework is going to use the full distribution of forecasted seismicity in likelihood tests developed in previous CSEP experiments (Schorlemmer et al., 2007; Nandan et al., 2019). The new validation methods (consistency tests) are based on families of synthetic catalogues (stochastic event sets), which account for the aleatory variability of the generating process as well its epistemic uncertainty. Exhaustive sets of simulated catalogues retain the full spatio-temporal dependencies amongst modelled earthquakes, i.e., they can reflect the full complexity of the model simulations. The new CSEP2.0 tests, which are developed within the EU-project RISE under participation of GFZ, will operate directly on these sets of simulated catalogues. Because the evaluations consider the full distribution of forecasted seismicity instead of just the expected number in each bin, the tests do not need to assume independence of events and can probe the dependence structure adequately.

4.3 CSEP-tests for induced seismicity

Not much work has been done so far on CSEP-type model selection and comparison in the context of induced seismicity, besides a first attempt by Kiraly-Proag et al. (2016) for the case of two models applied to geothermal stimulations in Basel (2006) and Soultz-sous-Forêts (2004). In contrast to natural tectonic-induced seismicity, anthropogenic-induced seismicity clearly shows the influence of the non-stationary forcing through the technical operations. While CSEP usually only considers earthquake catalogs in the test procedures, induced seismicity also requires data sets for the load input in the reservoir. As soon as these are taken into account, a transfer of the statistical test procedures from CSEP is straightforward. Optimally, fully prospective tests would be performed in an independent testing centre. However, it would require that a reasonable number of model developers are able and willing to participate and provide software codes that run automatically, given the required input. This cannot be guaranteed and is maybe also unlikely because of commercialisation and licensing. A simpler and less demanding approach consists of CSEP-tests running on a web-interface,

which enables modellers to run their codes independently employing identical input data. Based on uniform input and test data, this allows for a meaningful model comparisons and rankings. Such a toolbox is proposed in Sec. 5.

5 Proposal

5.1 Project objectives

The objectives will be met via:

1. Providing a toolbox together with a web front-end, database and test-data management system to enable users to independently test their source models for Groningen and compare (and publish) them against other models;
2. Identifying the input data that needs to be made available to the users in standardised format;
3. Definition of criteria for statistical testing individual models and benchmarking between them.

5.2 Scope of work

The problem of massive data and complex analysis tools is to calibrate and validate the methods with standardised protocols and comparable input data. Especially for the Groningen field, a multitude of seismicity source models has been developed by various authors, comprising different types of model categories, parameter sets, assumptions and input data. From a literature review alone, it is impossible to completely assess the abilities and weaknesses of each of those models and model classes. Further, in order to perform an objective comparison of model results, it has to be ensured that models are run on exactly the same set of input data and use an uniform test data set. Since reservoirs and engineering sites develop induced seismicity over many years, it is also necessary to repeatedly adjust the input data and compare the models. This requires a standardised database approach in close connection with test procedures.

For a long time, a similar problem existed for natural seismicity. To tackle this problem, the Collaboratory for the Study of Earthquake Predictability (CSEP) was established (Jordan, 2006). Within the CSEP-framework, fully prospective tests of earthquake forecasts are carried out, using standardised statistical tests and authoritative data sets to assess the predictive skills of forecast models and to objectively compare them. Not much work has been done so far on model selection and comparison in the context of induced seismicity. However, the establishment of a fully functional testing centre is beyond the scope of this proposal; instead, we propose to develop a toolbox and repository for Groningen seismicity models to:

- provide comparable testing and standardised input data,
- validate results against "ground truth data sets" (e.g. data sets for bygone time periods) and
- benchmark results by comparing against results provided by other users using standardised tools.

The toolbox will have a web front-end, such that users can access authoritative data sets marked with

a unique ID, import their results and benchmark them with others already present in the database. Input data sets are analysed by the users themselves, who work remotely and independently employing their own routines. During upload of the results, a given format needs to be ensured. A comparison of all uploaded results is performed automatically. As part of a user agreement (to be developed), the comparisons could optionally be made available to independent, authorized users or openly published. Such a toolbox provides a procedure for testing and verification independently from the model developers.

For now, we concentrate on the statistical testing of seismicity models, but in principle, the toolbox could be extended to comprise other features of induced and triggered seismicity, as e.g. event locations in 3D media, magnitudes, moment tensors etc., and potentially include tests employing synthetic data sets. Similarly, although the toolbox will be developed with the aim to compare seismicity models for the Groningen field, it will be easily extendable to other gas fields or other engineering work as for example geothermal applications.

Requirements are the accessibility and preparation of a standardised, dynamic earthquake catalogue as well as space-time dependent input forcing data required by the different model types (e.g. pressure reduction, strain, precalculated Coulomb stress according to a given model), the integration of 1D and 3D velocity models and - in case of a subsequent extension towards other methodologies, a potential integration of waveform data via services in close cooperation with KNMI. A participation of users in the platform should be promoted by SodM as the outcome of the benchmarking may potentially be used in justifying weighting schemes for ensemble models in hazard studies.

Some technical aspects of the toolbox development will benefit from the development of a IASPEI testing platform for automatic moment tensor inversion (e.g. CoSOI commission, `http://iaspei.org/commissions/commission-on-seismological-observation-and-interpretation`) where GFZ is involved.

To reach our goal, we propose the following work packages:

WP1: Toolbox and repository for Groningen seismicity models (GSTB)

WP2: Retrospective experiments

WP3: Prospective tests (optional)

WP1: Toolbox and repository for Groningen seismicity models (GSTB)

The first work package will deal with the development of the toolbox and cover the following tasks:

Task 1.1: Toolbox design:

The requirements for functionality and accessibility will be assessed, the toolbox management devised and the GUI sketched *(Specification functionality, definition of tools, integration with pyrocko toolbox, specification of interfaces & standardised formats, internal structure)*.

Task 1.2: Toolbox development:

The designed concept will be implemented and instructions will be prepared. In particular, input and output protocols will be put into effect and the test procedures developed in the framework of CSEP (see Sec. 4) will be implemented *(buildup server infrastructure, front-end, admin and user managements, unique ID for data sets, implementing statistical test modules, basic reports, graphical outputs)*.

Task 1.3: Identification and preparation of input data sets:

An important task will be to identify potential test users and to define - together with SodM - required input data sets. These data will then be collected and prepared in order to provide them as downloadable from the platform *(coordination with test users, data collection, quality control, wrapper for data accessibility, etc.)*

Task 1.4: Characterisation of test data:

Target events have to be defined. For that purpose, the completeness magnitude will be carefully evaluated to include, on the one hand, as many events as possible to increase the power of the tests, but ensure, on the other hand, the completeness of the test data set *(authoriative catalog preparation, characterisation)*

Task 1.5: Toolbox testing:

The functionality of the toolbox will be tested. In particular, we plan to run automatised tests for seismicity models developed at GFZ *(debugging, adjusting functionality, improving / adapting result tables and figures, implementation of different seismicity models)*.

Task 1.6: Publish toolbox:

Finally, the toolbox and instructions will be made available to users; its usage will be advertised via SodM and presentations on conferences will be given to attract potential users *(technical software release, user manual and case examples, presentations, dissemination, publications)* .

WP2: Retrospective experiments

Ideally, seismicity models are tested against unknown data. However, such a fully prospective test involves a time delay between forecast collection and testing. To obtain significant test results, this time delay might be in the order of years, depending on the level of seismic activity. Therefore, we plan to additionally run a retrospective test to compare the performance of the different models in response to the production reduction since 2014, which may provide an indication of the forecasting ability after complete shut-in. For this purpose, we want to replay the past in a pseudo-prospective test mode, where model inputs are limited to the data available at a certain time. In particular, this work package will cover the following tasks:

Task 2.1: Test design:

Firstly, the test with its detailed plan and processing steps will be formulated: (i) time frame, (ii)

target events, (iii) forecast periods (e.g. one year), (iv) spatial gridding of the test region (e.g. 5 km x 5 km), and (iv) test quantities (e.g. seismicity rates, magnitudes, spatial distribution).

Task 2.2: Data preparation:

According to the test design, the input data will be prepared; i.e. the reservoir and production data as well as the earthquake catalogue including only earthquakes that occurred before the start of the forecast.

Task 2.3: Test realisation:

Via the web-interface, all input data will be provided to users for each forecast period separately. Users will be asked to clarify that their parameter settings and model forecasts use solely this information.

Task 2.4: Benchmark models:

We will implement and run simple statistical models for benchmarking more sophisticated models of the users. In particular, we want to apply (i) a homogeneous Poisson model; (ii) an inhomogeneous Poisson model, in which the seismicity rate is proportional to the spatiotemporal stress changes, and (iii) an epidemic type aftershock sequence (ETAS) model with a similar inhomogeneous Poisson model simulating the background activity, in which aftershocks are additionally considered according to the modified Omori law. Furthermore, we will include the models analysed in Richter et al. (2020), in particular a version of the rate-state model.

Task 2.5: Test evaluation:

Based on the CSEP-tests, the model performance will be evaluated and compared to the benchmark models. Comparative tests between all submitted models will be performed and models will be ranked according to their performance. The results will be shared with the users.

WP3: Prospective tests

Parallel to or after WP2, we intend to start fully prospective tests of alternative seismicity models. As long as production is ongoing, models can make use of production plans and stress scenarios as input for their forecasts. Every year, we plan to call for model forecasts starting at a given deadline. The submitted models will be evaluated only after the end of the forecast period (e.g. one year) based on the observed seismicity in this time interval. In this case, target events and model development are completely separated and no hidden optimisation can potentially bias the test results. Only in the case that production differs significantly from production plans, we plan to reevaluate the models based on the true production. The detailed working steps are the following tasks:

Task 3.1: Input data collection:

We will provide all necessary input data including the planned production scenario for the next years. This input data needs to be obtained with the assistance of SodM.

Task 3.2: Start experiment:

Call for model submissions for the prospective test with given deadline.

Task 3.3: Evaluation:

Perform implemented CSEP-tests (see Sec. 4) after the end of the forecast period, once the KNMI earthquake catalogue is available for the forecast period.

Task 3.4: Ranking:

Comparative test results will be provided which can be used for weighting models e.g. in seismic hazard assessment.

The prospective test is offered independently from the proposal outlined in WP1 and WP2; it can be ordered anytime after that WP1 has been finished and the toolbox is functioning. Therefore, no specific time frame has been assigned to WP3 and its cost are not contained in the project phase 2 budget. The length of WP3 depends on the forecast period that is supposed to be tested. The prospective testing is offered by GFZ. Prices available on request after WP1 has been finished.

5.3 Potential for publications

We see the potential for at least one technical and one scientific ISI publication. Depending on the users of the platform, doi-referenced data publications for test results could be developed in the medium term.

5.4 Collaboration between NORSAR and GFZ

NORSAR and GFZ will enter into the contract as partnership, in which each partner is individually and jointly and severally liable for the performance of the obligations arising from the project. This partnership between NORSAR and GFZ is covered by a letter of intent already in existence dated 28 May 2015. NORSAR will act as consortium leader of the partnership and the authorised representative towards SodM.

Under this letter of intent, NORSAR and GFZ collaborated successfully in the past, e.g. in the framework of the ACISM project (see section 2.4 for details), the SECURE project (see section 2.4 for details) and the present project.

Furthermore, Dr. Daniela Kühn, who will act as project manager, holds a 20% position at GFZ in addition to her position at NORSAR, simplifying and guaranteeing good communication between both groups.

5.5 Price

5.5.1 Personnel costs

The price associated with the different personnel categories to be involved in this project in 2020 are listed below. Categories are according to seniority and experience of researchers. The hourly rates are assumed to increase by 3% every year for each organisation. These rates have been applied to the prices quoted for each of the work packages, after assigning tasks to relevant personnel and applying an estimated number of working hours.

Category	NORSAR pr. hr. (NOK)	GFZ pr. hr. (EUR)
A	2 150	155
B	1 770	103
C	1 320	82

5.5.2 Costs per task

The following tables provide the price for:

- work packages (as outlined in section 5.2);
- the administrative costs associated with the project;
- and a final price for the project as a whole.

A contingency of 5% is added to NORSAR's costs. This contingency comprises among others the (currently more than ever) unforeseeable variability in the exchange rate between EUR and NOK.

Work packages				**Price (EUR)**
	2020	**2021**	**2022**	**Total**
WP1	39 252	361 885	85 554	486 691
WP2		39 233	140 629	179 862
All work packages subtotal:	39 252	401 118	226 183	**666 553**

Administration and communication				**Price (EUR)**
	2020	**2021**	**2022**	**Total**
Project management	9 655	10 939	19 461	40 054
Quality control	2 345	4 832	6 221	13 398
Publications		17 360	17 881	35 241
Travel	153	10 652	9 422	20 226
Administration subtotal:	12 153	43 782	52 985	**108 919**

Total project costs				Price (EUR
	2020	**2021**	**2022**	**Total**
Work packages	39 252	401 118	226 183	666 553
Administration	12 153	43 782	52 985	108 919
Contingency	1 169	9 147	5 840	16 156
Project total:	52 575	454 047	285 007	**791 629**

Project share				
	NORSAR (EUR)	**GFZ (EUR)**	**% NORSAR**	**% GFZ**
Work packages	22 2656	443 897	33	67
Administration	100 473	8 446	92	8
Contingency	16 156	0	100	0
Project total:	339 285	452 343	43	57

5.6 Schedule and key deliverables

We propose to start the project in Q4-2020. Its duration is approximately two years (with an additional three months at the end of the project to ensure publication of toolbox and results) until end of the year 2022. Reporting towards SodM will be via video meetings, Powerpoint presentations and a final report. GFZ and NORSAR will conduct monthly progress meetings and invite SodM personnel regularly to participate.

5.6.1 Deliverables

- draft of toolbox structure and overview on collected data (end of Q2-2021),
- implementation of toolbox (end of Q3-2021),
- results of benchmark models (end of Q1-2022),
- result of retrospective tests (end of Q3-2022),
- publication of toolbox (end of Q4-2022).

5.7 Quality assurance

At NORSAR, quality assurance applies to two main areas:

- Operational quality or processes that ensure a robust, consistent and professional business. This is controlled through the use of NORSAR's management system. The management system is based on principles from the ISO 9001 standards.
- Quality of primary functions. This has been controlled for each project through project planning and quality assurance.

Risk management is the basis for prioritisations and corrective measures both at a management level and for each individual activity. Risk is assessed with the use of risk matrices for both business operations and for HSE. Risk is regularly followed up at management meetings, and is always defined before beginning work on a new primary activity.

NORSAR is a member of Achilles Joint Qualification System ensuring sustainable management of Quality, Safety, Health and Environment.

5.8 Commercial qualifications

Invoices shall be paid within 30 days after issue date.

Appendices

A Literature

References

Bourne, S. and S. Oates (2014). *An activity rate model of induced seismicity within the Groningen field*. Tech. rep. Nederlandse Aardolie Maatschappij BV.

Bourne, S. and S. Oates (2018). *The influence of stress rates on induced seismicity rates within the Groningen gas field*. Tech. rep. NAM.

Bourne, S.J., S.J. Oates, and J. Van Elk (2018). "The exponential rise of induced seismicity with increasing stress levels in the Groningen gas field and its implications for controlling seismic risk". In: *Geophysical Journal International*.

Bourne, S.J., S.J. Oates, J. Van Elk, and D. Doornhof (2014). "A seismological model for earthquakes induced by fluid extraction from a subsurface reservoir". In: *Journal of Geophysical Research: Solid Earth* 119.12, pp. 8991–9015.

Buijze, L., P. A. J. van den Bogert, B. B. T. Wassing, and B. Orlic (2019). "Nucleation and arrest of dynamic rupture induced by reservoir depletion". In: *J. Geophys. Research*. DOI: 10.1029/2018JB016941.

Candela, T., S. Osinga, J.P. Ampuero, B. Wassing, M. Pluymaekers, P. A. Fokker, J.D. van Wees, H. A. de Waal, and A. G. MuntendamBos (2019). "Depletioninduced seismicity at the Groningen gas field: Coulomb rateandstate models including differential compaction effect". In: *J. Geophys. Res.* DOI: 10.1029/2018JB016670.

Cattania, C. et al. (2018). "The forecasting skill of physics-based seismicity models during the 2010-2012 Canterbury, New Zealand, earthquake sequence". In: *Seismological Research Letters* 89, pp. 1238–1250.

Dahm, T., S. Hainzl, D. Kühn, V. Oye, G. Richter, and I. Vera-Rodriguez (2020a). *Review of seismogenic source models for the Groningen gas reservoir (WP1: review of the existing NAM seismogenic source model)*. Tech. rep. NORSAR and GFZ.

Dahm, T., S. Hainzl, D. Kühn, V. Oye, G. Richter, and I. Vera-Rodriguez (2020b). *Review of seismogenic source models for the Groningen gas reservoir (WP2: review of existing alternative approaches and possible knowledge gaps)*. Tech. rep. NORSAR and GFZ.

DeDontney, N. (2017). *Groningen seismological models for activity rate forecasting*. Tech. rep. NAM.

DeDontney, N. and S. Lele (2018). *Impact of production fluctuations on Groningen seismicity – Geomechanical modelling using Rate and State friction*. Tech. rep. NAM.

Dempsey, D. and J. Suckale (2017). "Physics-based forecasting of induced seismicity at Groningen gas field, the Netherlands". In: *Geophysical Research Letters* 44.15, pp. 7773–7782.

European Committee for Standardization (2005). *Eurocode 8: Design of structures for earthquake resistance-part 1: general rules, seismic actions and rules for buildings*. Tech. rep. Brussels.

Helmstetter, A. and M. J. Werner (2012). "Adaptive spatio-temporal smoothing of seismicity for long-term earthquake forecasts in California". In: *Bull. Seismol. Soc. Am.* Pp. 2518–2529.

Helmstetter, A. and M. J. Werner (2014). “Adaptive smoothing of seismicity in time, space, and magnitude for time-dependent earthquake forecasts for California”. In: *Bull. Seismol. Soc. Am.* Pp. 809–822.

Hettema, M.H.H., B. Jaarsma, B.M. Schroot, and G.C.N. Van Yperen (2017). “An empirical relationship for the seismic activity rate of the Groningen gas field”. In: *Netherlands Journal of Geosciences* 96.5, s149–s161.

Jansen, J. D., P. Singhal, and F. Vossepoel (2019). “Insights from closed-form expressions for injection- and production-induced stresses in displaced faults”. In: *J. Geophys. Res.* DOI: 10.1029/2019JB017932.

Jordan, T. H. (2006). “Earthquake predictability, brick by brick”. In: *Seismol. Res. Lett.* Pp. 3–6.

Jordan, T. H., Y. T. Chen, P. Gasparini, R. Madariaga, I. Main, W. Marzocchi, and G. Papadopoulos (2011). “Operational earthquake forecasting. State of knowledge and guidelines for utilization”. In: *Annals of Geophysics* 54.4. DOI: 10.4401/ag-5350.

Kettermann, M., S. Abe, A. F. Raith, J. de Jager, and J. L. Urai (2017). “The effect of salt in dilatant faults on rates and magnitudes of induced seismicity first results building on the geological setting of the Groningen Rotliegend reservoirs”. In: *Netherlands Journal of Geosciences*, s87–s104.

Kiraly-Proag, E., J. D. Zechar, V. Gischig, S. Wiemer, D. Karvounis, and J. Doetsch (2016). “Validating induced seismicity forecast models – Induced Seismicity Test Bench”. In: *J. Geophys. Res.* Pp. 6009–6029.

Lanz, F., K. Bisdom, E. Barbaro, J. Limbeck, T. Park, C. Harris, and K. Nevenzeel (2019). *Evaluation of a Machine Learning methodology for spatio-temporal induced seismicity forecasts within the Groningen field - Machine Learning*. Tech. rep. NAM.

Lombardi, A. M. and W. Marzocchi (2010). “The assumption of Poisson seismic-rate variability in CSEP/RELM experiments”. In: *Bulletin of the Seismological Society of America* 100, pp. 2293–2300. DOI: 10.1785/0120100012.

Mayuga, M. and D. Allen (1969). “Subsidence in the Wilmington oil field, Long Beach, California”. In: *1st International Symposium on Land Subsidence*.

Michael, A. J. and M. J. Werner (2018). “Preface to the focus section on the Collaboratory for the Study of Earthquake Predictability (CSEP): New results and future directions”. In: *Seismological Research Letters* 89, pp. 1226–1228.

Muntendam-Bos, A. G. (2020). “Clustering characteristics of gas-extraction induced seismicity in the Groningen gas field”. In: *Geophys. J. Int.* in press.

Nandan, S., G. Ouillon, D. Sornette, and S. Wiemer (2019). “Forecasting the full distribution of earthquake numbers is fair, robust, and better”. In: *Seismological Research Letters* 90, pp. 1650–1659.

Ogata, Y. (1998). “Space-time point-process models for earthquake occurrences”. In: *Annals of the Institute of Statistical Mathematics* 50.2, pp. 379–402.

Rhoades, D. A., D. Schorlemmer, M. C. Gerstenberger, A. Christophersen, J. D. Zechar, and M. Imoto (2011). “Efficient testing of earthquake forecasting models”. In: *Acta Geophys.* 59, pp. 728–747. DOI: 10.2478/s11600011-0013-5.

Richter, G., Hainzl S., T. Dahm, and G. Zöller (2020). "Stress-based, statistical modeling of the induced seismicity at the Groningen Gas Field, The Netherlands". In: *Environmental Earth Sciences*.

Sanz, P. F., S. P. Lele, K. H. Searles, S.-Y. Hsu, J. L. Garzon, J. A. Burdette, W. E. Kline, B. A. Dale, and P. D. Hector (2015). "Geomechanical analysis to evaluate production-induced fault reactivation at Groningen gas field". In: *SEG Annual Technical Conference and Exhibition, Houston, Texas*.

Schorlemmer, D. and M. C. Gerstenberger (2007). "RELM Testing Center". In: *Seismological Research Letters* 78, pp. 30–36.

Schorlemmer, D., M. Gerstenberger, S. Wiemer, D. D. Jackson, and D. A. Rhoades (2007). "Earthquake likelihood model testing". In: *Seismological Research Letters* 78, pp. 17–29.

Schorlemmer, D. et al. (2018). "The Collaboratory for the Study of Earthquake Predictability: achievements and priorities". In: *Seismological Research Letters* 89, pp. 1305–1313.

Segall, P. (1989). "Earthquakes triggered by fluid extraction". In: *Geology* 17.10, pp. 942–946.

Shapiro, S. A. (2018). "Seismogenic index of underground fluid injections and productions". In: *J. Geophys. Res.* Pp. 7983–7997.

Shapiro, S. A., C. Dinske, C. Langenbruch, and F. Wenzel (2010). "Seismogenic index and magnitude probability of earthquakes induced during reservoir fluid stimulations". In: *Lead. Edge*, pp. 304–309.

Van Eck, T., F. Goutbeek, H. Haak, and B. Dost (2006). "Seismic hazard due to small-magnitude, shallow-source, induced earthquakes in The Netherlands". In: *Engineering Geology* 87.1-2, pp. 105–121.

Van Wees, J.-D., S. Osinga, K. Van Thienen-Visser, and P.A. Fokker (2017). "Reservoir creep and induced seismicity: inferences from geomechanical modeling of gas depletion in the Groningen field". In: *Geophysical Journal International* 212.3, pp. 1487–1497.

Vlek, C. (2019). "Rise and reduction of induced earthquakes in the Groningen gas field, 19912018: statistical trends, social impacts, and policy change". In: *Environmental Earth Sciences*. DOI: 10.1007/s12665-019-8051-4.

Wentinck, R. (2015). *Induced seismicity in the Groningen field: A statistical assessment of tremors along faults in a compacting reservoir*. Tech. rep. NAM.

Werner, M. J., A. Helmstetter, D. D. Jackson, and Y. Y. Kagan (2011). "High-resolution long-term and short-term earthquake forecasts for California". In: *Bulletin of the Seismological Society of America* 101, pp. 1630–1648.

Woessner, J., S. Hainzl, W. Marzocchi, M. J. Werner, A. M. Lombardi, F. Catalli, B. Enescu, M. Cocco, M. C. Gerstenberger, and S. Wiemer (2011). "A retrospective comparative forecast test on the 1992 Landers sequence". In: *Journal of Geophysical Research-Solid Earth* 116, B05305. DOI: 10.1029/2010JB007846.

Yerkes, R. and R. Castle (1969). "Surface deformation associated with oil and gas operations in the United States". In: *1st International Symposium on Land Subsidence*.

Zbinden, D., A. P. Rinaldi, L. Urpi, and S. Wiemer (2017). "On the physics-based processes behind production-induced seismicity in natural gas fields". In: *J. Geophys. Res.* Pp. 3792–3812.

Zechar, J. D., M. C. Gerstenberger, and D. A. Rhoades (2010). “Likelihood-based tests for evaluating spaceratemagnitude earthquake forecasts”. In: *Bulletin of the Seismological Society of America* 100, pp. 1184–1195. DOI: 10.1785/0120090192.

Report Number: 20-002	Confidential: X Unlimited:	External: X Internal:	NORSAR Project No.: 10178
Title:	Review of seismogenic source models for the Groningen gas reservoir (WP3: Proposal for an alternative seismogenic source representation)		
Client:	Staatstoezicht op de Mijnen, Netherlands		
Project manager:	D. Kühn		
Authors/prepared by (alphabetical order):	T. Dahm (GFZ), S. Hainzl (GFZ), D. Kühn (NORSAR), V. Oye (NORSAR), G. Richter (GFZ), I. Vera Rodriguez (NORSAR)		
Submitted to:	Staatstoezicht op de Mijnen, Netherlands		
Contract reference:			
Archive reference:			
Approved by: Department head:	Name: V. Oye	Signature: Volker Oye	Date: 30.04.2020

NORSAR

info@norsar.no
www.norsar.no

GFZ
Helmholtz-Zentrum
POTSDAM