Addendum to report "Ensemble Based Subsidence application to the Ameland gas field – long term subsidence study part two (LTS II)"

NAM

Doc nr. EP201703202079
Version 0. Status: FINAL
DD 17-03-2017

Table of Contents

1 Summary

This addendum describes the modifications that have been incorporated in the work flow since the publication of the NAM (2017) report. The addendum demonstrates that NAM is able to apply the work flow to a real field case (Ameland) when incorporating the changes proposed and agreed by the LTS-2 consortium. This conclusion confirms that the objective of the LTS-2 study is now satisfied.

An important objective that was missing in the main report was the comparison to previously published results in the measurement and control documents for a known area like the sand-sharing area Pinkegat. This comparison can be found in this addendum as well.

The modifications to the workflow resulted in:

- A clear preference for the reservoir scenarios with limited aquifer depletion.
- A weighted average that forms the stochastic expectation case. This result matches satisfactorily with the model and forecast published in NAM (2016a) and gives confidence in the subsidence predictions above the Wadden Sea.
- The "remweg" scenarios tested with this method show a clear deceleration of the subsidence rate over the Pinkegat sand sharing area.

The upscaling of the method to the full Wadden-area in future M&R cycles will be very challenging and requires further discussions with the M&R cycle committee (Auditcommissie).

2   Reason for this addendum.

A precursor to these results was the identification of the following three issues in the LTS 2 main report (NAM, 2017)

1. Is the variance/co-variance matrix in itself a stochastic parameter? The contribution of the idealisation noise to this matrix is based on a model with five parameters. Since their estimation of these parameters turned out not to have a unique and stable solution, at least one parameter had to be fixed to a given value. Was this choice adequate and what is its impact on the test statistic?
2. Despite the observation of a reasonable match of the best model members with the data (σ around 1cm) the lowest value for the χ^2/N value is around 16. Why do we obtain such counter intuitive values and what are the possibilities to reduce these?
3. Despite the fact that the method did dismisses most depleting aquifer scenarios in the Wadden Sea, the contribution or weight of these scenarios is still considerable. This results in a stochastic expectation case that does not match to the GPS data in the Wadden Sea and therefore largely over predicts subsidence volumes and rates.

During the discussion of these issues in the consortium meeting, the viscosity values for the salt were challenged and additional investigation on this topic is described in this addendum as well.

2.1   Understanding the role of parameter choices in the population of the variance-covariance matrix.

The variance-covariance matrix of the "data" comprises two contributions. They reflect the uncertainty of geodetic measurements on the one hand and benchmark movements due to shallow processes on the other hand. The first is modelled by the measurement noise, and the latter is described by the idealisation noise. Both components depend on multiple parameter choices. Whereas the parameter choices of the measurement noise are relatively robust and can be based on geodetic experience, the choice of an appropriate idealisation noise model is subject to substantially larger uncertainties.

Regarding idealisation noise, it turned out that the simultaneous estimation of all model parameters at a time does not converge to a sufficiently stable and unique solution. As a remedy it was chosen to constrain one parameter to an a priori value. In order to narrow down the possible causes for the high χ^2/N test statistic, this choice shall be reviewed in section 3.1. In addition, the impact of the range of possible choices shall be assessed.

2.2 Why do we observe a high value for the test statistic?

The test statistic "χ^2/N" is a sum of squared residuals that are weighted by their variance-covariance matrix and divided by the number of observations. It is assumed that the residuals are normally distributed with zero mean and appropriately weighted by their covariance matrix. Seeing that the number of observations is large and the number of geomechanical model parameters is small, the test statistic follows a $F_{N,\infty}$-distribution with expectation value 1. Obtaining values much larger than 1 indicates a model mismatch that can be interpreted in different ways and localised in a four-quadrant model (see Figure 1).

	Functional model	Stochastic model
Geodetic observations	I	II
Geomechanical models	III	IV

Figure 1: Quadrant model for the localisation of model errors as an explanation of statistical misfit. The functional model describes the functional relationship between observations and model parameters. The stochastic model describes the expectable misfit of the functional model due to residual uncertainties. Geodetic observations quantify relative displacements of surface benchmarks. Geomechanical models describe the expected surface deformation caused by hydrocarbon extraction.

   I. **Functional model of geodetic observations:** The functional relationship between measurements and benchmark movement is generally very well known. It does not cover exceptional outliers though that can occur if the actually observed benchmark differs from the supposedly measured benchmark. Undetected outliers can be responsible for a too high test statistic (see section 2.2.2).
   II. **Stochastic model of geodetic observations:** The uncertainty of geodetic observations is well quantifiable by parametric models for measurement noise. Even though these models always have some residual imperfections, they are considered good enough to be excluded a potential cause of the observed mismatch.
   III. **Functional model of geomechanical modelling:** Imperfections of surface subsidence predictions, which have been the primary motivation to carry out this study, can be considered a potential cause for the high test statistic. (see section 2.2.1)
   IV. **Stochastic model of geomechanical modelling:** In the pursued approach, surface subsidence predictions are assumed deterministic. Residual imperfections are accounted for by appropriately defining and sampling the solution space. However, if the perfect model is not covered by any member from any ensemble, it is worth consideration to incorporate residual imperfections in a stochastic manner (see section 3.5.1)

Idealisation noise: Surface deformation due to shallow sources form a missing link between relative benchmark displacements observed by geodetic techniques and surface deformation predicted from geophysical models. There is no unique rationale to locate shallow deformation effects in a specific quadrant. In the current study, they are subsumed by idealisation noise, which contributes together

with measurement noise (quadrant II) to the covariance matrix of the "data". Thus, idealisation noise models can potentially impact the test statistic and be partly responsible for its high value (see section 3.1).

2.2.1   High test statistic as a result of a not perfect geomechanical model

The most logical explanation we found to explain the high values for the test statistic is the relative high precision of the levelling measurements on the Ameland Island. An estimate of the standard deviation of the levelling points leads to a value of around 3 mm, which is much smaller than the values for sigma of the model mismatch with the data of around 1cm. Especially the data points on the eastern part of the island, highlighted by the red rectangles, are over predicted by all models up to values of around 5 cm. We therefore conclude that the high value of the test statistic is dominantly controlled by the precise levelling double differences on the Ameland Island and using them in a confrontation to each discrete member. The high value could also be explained by a sub optimal geomechanical model with a remark that also more sophisticated finite element models produce similar mismatches.

Figure 2; Subsidence vs time at certain benchmark locations (bottom) for the model members having the highest probability, the reservoir scenario with the least aquifer depletion and an influence function based on a viscous salt (modified from figure 73, NAM, 2017).

The question now arises whether we will ever find higher order geomechanical models that could come closer to the data. The search to such a model certainly leads to more complexity and therefore introduces a larger parameter space making stochastic work flows even more complex. More important is the question what the impact of the high value of the test statistic is on the quality of the subsidence forecasts for the sand sharing areas.

2.2.2 High test statistic because of outliers in the Double Differences

In LTS-1 it was proposed to deviate from a more thorough geodetic approach for outlier detection. A different approach was chosen due to difficulties in quantifying geomechanical modelling uncertainties. According to that approach, only the most obvious outliers should be identified in individual time series of double differences with respect to a small number of different reference benchmarks. An a-priori geomechanical model needs to be subtracted from these time series beforehand. This approach has been pursued in LTS-2.

Statistical testing was based on measurement uncertainties and idealisation noise only. To account for residual modelling uncertainties, the acceptance threshold has been relaxed in a way that outliers are only flagged, if there is no doubt that the abnormal deformation pattern is not related to driving mechanisms in the deep underground. Thus, a priori modelling assumptions are limited to the necessary minimum. On the other hand, this strategy has a poor sensitivity and does not exploit spatial relationships between benchmarks. Thus, the high test statistic could (partly) be caused by undetected outliers.

2.3 Poor distinctiveness between reservoir scenarios

One of the aims of the LTS II study was to demonstrate that the confrontation of the models to the data would bring in a clear distinction between the chosen reservoir scenarios. NAM (2017) showed that this objective is partially met. Some differentiation is observed but the scenarios with high depletion in the Wadden Sea do still have a high posterior probability value even for those cases that show a clear mismatch with the measured data in the Wadden Sea. We concluded in NAM (2017) that this observation arises from the dominant role of the geodetic data on the Island over the GPS points in the Wadden Sea.

An error was however discovered in the formula that was used to calculate the probability of the member, which was embedded in the code that was provided by TNO.

The probability according to TNO (2016) is defined as:

$$P\left(dd_r^{prior}\middle|dd\right) = \frac{P\left(dd_r^{prior}\right) \cdot P\left(dd\middle|dd_r^{prior}\right)}{\sum_{i=1}^{Ne} P\left(dd_i^{prior}\right) \cdot P\left(dd\middle|dd_i^{prior}\right)}$$
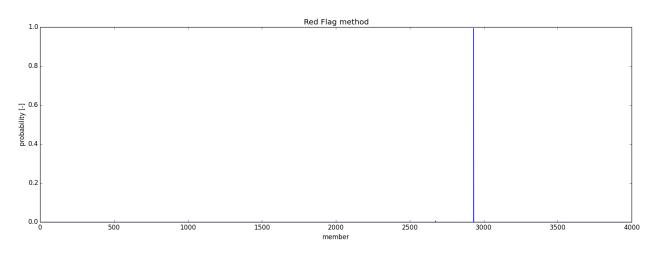
$$P\left(dd\middle|dd_r^{prior}\right) = exp[-\frac{\chi^2}{2N}]$$

Where N is the number of double differences. This N should be removed from the formula when applying the theory of Red Flag correctly (Nepveu et al., 2010). The effect of removing the N on the probability calculation has a significant effect on the probability values for the members in the ensemble. Applying the new formula:

$$P\left(dd\middle|dd_r^{prior}\right) = exp[-\frac{\chi^2}{2}]$$

results in a collapse or divergence of the ensemble into a few member having a high probability 1 and all other members having a probability of 0.

This means that in case of χ^2 values being close to each other, the weighing by this probability definition causes an extreme discrimination between them which leads to the conclusion that the calculation diverges. An example of a posterior probability density plot of this result is shown in Figure 3.

Figure 3; Pdf of posterior distribution using the correct red flag equations.

The next chapter will describe the work that has been executed to address the gaps that were concluded in NAM (2017).

3 Description of the topics to address identified issues

3.1 Parametric model for idealisation noise and outlier removal

The idealisation noise model applied in the Long Term Study can be decomposed into two components (NAM, 2017):

- *Temporal component:* This subsumes all shallow effects that are correlated in time and uncorrelated in space. It can be considered as an autonomous movement of an individual benchmark that has nothing in common with the behaviour of neighbouring benchmarks. An example would be settlement of the individual building a benchmark is attached to.
- *Spatio-temporal component:* This subsumes all shallow effects that are correlated in both space and time. It can be considered as a coherence in movement with neighbouring benchmarks, whereas the level of coherence decreases with distance. An example would be the compaction of a shallow peat layer.

Evaluated for the variance of a double difference observation, the five-parametric model reads:

$$\sigma_{DD}^2 = \underbrace{2\sigma_t^2 \Delta t^{p_t}}_{\text{temp. comp.}} + \underbrace{2\sigma_s^2 \left(1 - e^{-\left(\frac{\Delta d}{L}\right)}\right) \Delta t^{p_s}}_{\text{spatio-temp. component}}$$

Two datasets have been used to estimate idealisation noise parameters within the scope of the Long Term Study (see Figure 4):

1. Calibration of both components to an onshore levelling dataset from the LTS1 study area in Northern Friesland. In that area, both components can be estimated due to absence of gas production (Samiei-Esfahany and Bähr, 2015 and 2017). Both components turned out to be significant.
2. Calibration of the temporal component to the relative intra-cluster movements of the offshore benchmarks (LTS2; van Leijen et al., 2017). Levelling between these benchmarks, that are placed in clusters of three with mutual distances ≤ 15 m, provide the opportunity to estimate the temporal component in isolation. This is possible, because all spatially correlated signal components cancel out for short distances.
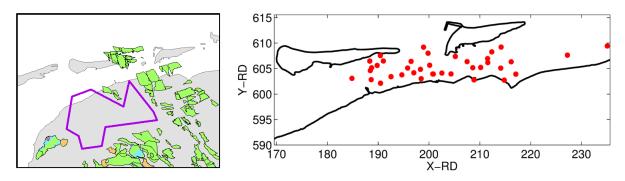

*Figure 4: Study areas for the determination of idealisation noise parameters. **Left:** LTS1 onshore dataset (Samiei-Esfahany and Bähr, 2015 and 2017). **Right:** LTS2 offshore dataset (van Leijen et al. 2017).*

To review the robustness of the model parameter estimation, we started with the offshore dataset, because a two-parametric model for the temporal component is less complex than a five-parametric model for both components. From Figure 5 can be seen that even though the estimation is stable, it

is not well-constrained. Whereas point A marks the global minimum, the parameter set marked by B gives a slightly less optimal but still reasonable fit. In the figure on the right-hand side it can be seen that models A and B yield similar standard deviations for time spans up to 9 years, which is the maximum time span covered by the underlying dataset. Model A, however, which has been used in (NAM, 2017), performs poor for longer time spans. It suggests quite high standard deviations, increasing almost linearly with time. This is not an expected behaviour for benchmark settlement, which is suspected to be the primary driving mechanism for spatially uncorrelated autonomous movements. Settlement would be expected to decelerate over time.
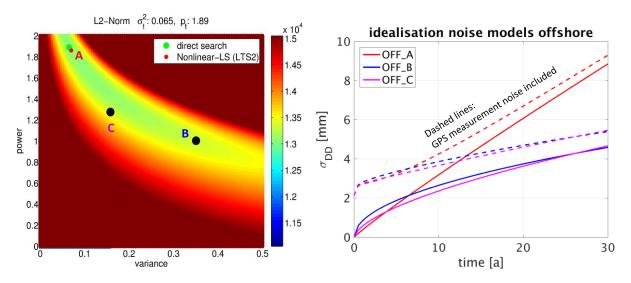

Figure 5: Estimation of the temporal idealisation noise component from intra-cluster levellings between the Wadden Sea benchmarks. **Left:** *Solution space: A greenish colour (corresponding to a low L2 norm) indicates a good agreement between data and model. Point A marks the global minimum. Point B is arbitrarily chosen to represent an extreme case with a still reasonable fit. Point C represents the parameters estimated by Houtenbos and Kenselaar (2001).* **Right:** *Contribution of the temporal idealisation noise component to the standard deviations of double difference measurements for the models indicated on the left hand side. See Table 2 for the model parameters.*

For comparison, also a third model is considered in Figure 5: Model C (Houtenbos and Kenselaar, 2001)[1] is close to model B, still reasonably close to the minimum area in the solution space (see Figure 5) but based on a considerably larger dataset (see Table 1). Model C is slightly less optimal than model A for describing the idealisation noise in the LTS2 offshore dataset, because it does not minimise the objective function and is not tailored to the benchmark type in the Wadden Sea. However, it is more appropriate for longer time spans that also occur in the offshore part of LTS2, originating from hydrostatic levelling. Due to the size of the underlying dataset, which includes observations from the whole of the Netherlands, it is also considered a more precise estimate. Therefore, model C is considered most appropriate for the use for offshore benchmarks in LTS2.

---

[1] 3-component model for "other" (i. e., non-underground) benchmarks with outlier removal up to 2.5 σ from Houtenbos and Kenselaar (2001), Table 5.5. Mind the different definition of p_t in that publication.

	Houtenbos and Kenselaar (2001)	LTS1 onshore	LTS2 offshore
Data	unadjusted sections	adjusted networks	benchmark clusters
Area	whole Netherlands	Northern Friesland	part of Wadden Sea
Temporal sampling	0.5...40 a (Δt = 1 a)	5...25 a (Δt ≈ 5 a)	1...9 a (Δt ≈ 1 a)
Spatial sampling	0.05...1.6 km (Δs = 0.1 km)	0.25...12 km (Δs = 0.5 km)	0 km
temporal component	most stable estimation (large dataset)	poorly constrained for t < 5 a	poorly constrained for t > 9 a
spatio-temporal component	not significant (poorly constrained?)	well estimable	not estimable

Regarding the calibration of the full five-parametric model (including both temporal and spatio-temporal component) to the LTS1 onshore dataset, it was already concluded that this estimation is very unstable. It fails for most of the simulations (Samiei-Esfahany and Bähr, 2017). Therefore, it was deemed necessary to constrain at least one parameter to a given value. It was chosen to constrain p_t to the value estimated from the LTS2 offshore dataset (model OFF_A), which seemed to be the least arbitrary choice at the time. However, since there are doubts about the appropriateness of model OFF_A (see above), some alternatives are explored in the following (see Table 2 and Figure 6).

Table 2: Parameters of different idealisation noise models. **OFF_A:** *The temporal component of this model was used for offshore benchmarks in (NAM, 2017).* **OFF_C:** *These are the parameters from the 3-component model for "other" (i. e., non-underground) with outlier removal up to 2.5 σ from Houtenbos and Kenselaar (2001, Table 5.5; mind the different definition of p_t). They have been used for all computations within the scope of this addendum.* **ON_C4_A:** *The temporal component of this model was used for onshore benchmarks in (NAM, 2017).*

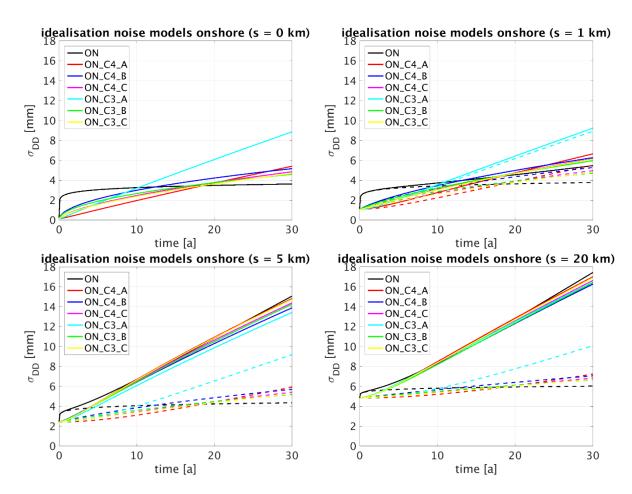|	σ_s^2 [$\mathbf{mm^2/km/a^{p_s}}$]	L [m]	p_s [−]	σ_t^2 [mm^2/a^{p_t}]	p_t [−]	model description
OFF_A	−	−	−	0.07	1.86	LTS2 offshore dataset, global minimum
OFF_B	−	−	−	0.35	1.00	LTS2 offshore dataset, extreme case
OFF_C	−	−	−	0.16	1.24	Houtenbos and Kenselaar (2001)
ON	0.277	4086	1.817	3.361	0.194	LTS1 onshore dataset, unconstrained
ON_C4_A	0.684	4069	1.515	0.026	1.860	LTS1 onshore dataset, p_t constrained to OFF_A
ON_C4_B	0.406	4276	1.640	0.442	1.000	LTS1 onshore dataset, p_t constrained to OFF_B
ON_C4_C	0.515	4117	1.590	0.172	1.240	LTS1 onshore dataset, p_t constrained to OFF_C
ON_C3_A	0.749	5471	1.390	0.070	1.860	LTS1 onshore dataset, σ_t^2 and p_t constrained to OFF_A
ON_C3_B	0.411	4062	1.650	0.350	1.000	LTS1 onshore dataset, σ_t^2 and p_t constrained to OFF_B
ON_C3_C	0.511	4023	1.606	0.160	1.240	LTS1 onshore dataset, σ_t^2 and p_t constrained to OFF_C

Figure 6: Different idealisation noise models estimated from the LTS1 onshore dataset, evaluated for standard deviations of double differences covering spatial distances of 0, 1, 5 and 20 km. The continuous lines show both idealisation noise components together with levelling measurement noise. The dashed lines show only the temporal component together with measurement noise. Note that the plot for s = 0 km shows the temporal component almost in isolation, because the spatio-temporal component is zero and constant the measurement noise of 0.14 mm is negligible. See Table 2 for an explanation of the different models and the model parameters.

Sticking to the unconstrained solution seems inappropriate from a technical point of view. The dataset from the LTS1 study area poorly constrains the model for time spans shorter than five years. Hence, the unconstrained model ("ON") suggests an unrealistically large temporal component for short time spans (e. g., σ = 2.0 mm for a double difference spanning one month).

Since it has been concluded that model OFF_C is preferable with respect to model OFF_A, it would be consistent to constrain p_t to the estimate from OFF_C instead of OFF_A. This results in model ON_C4_C. Seeing that the estimated σ_t^2 in model ON_C4_C is almost identical with σ_t^2 in model OFF_C and taking into account the high uncertainty of this estimation ($D\{\sigma_t^2\} = 0.055 \, \text{mm}^2/\text{a}^{p_t}$ for model ON_C4_C), it seems justified to constrain both σ_t^2 and p_t to yield model ON_C3_C as the most appropriate model to describe the idealisation noise in the LTS1 study area.

In conclusion, the model ON_C3_C is considered the most appropriate model to describe idealisation noise. Its temporal component coincides with the model OFF_C (Houtenbos and Kenselaar, 2001).

There is still some uncertainty whether it is appropriate to apply the spatio-temporal idealisation noise component to the LTS2-dataset. As in the main report (NAM, 2017), this component has been assumed zero in all modelling calculations presented in this addendum.

The implication for the test statistic is expected to be small. For onshore benchmarks, the temporal component of model OFF_C/ON_C3_C is used instead of ON_C4_A. These models almost coincide (see Figure 6). Only for a small number of double differences with observations of offshore benchmarks and time spans longer than ten years, a substantial drop in standard deviation can be observed. This drop stems from using model OFF_C instead of OFF_A (see Figure 5).

3.1.1 Outlier detection

In order to assess the impact of potentially undetected outliers, a couple of pragmatic detection approaches have been pursued to get more insight in potentially undetected blunders:

- decreasing the threshold of the approach proposed in LTS1 and applied in (NAM, 2017)
- geodetic w-test in the simplified model $E\left\{\begin{pmatrix} y_{\text{data}} \\ y_{\text{model}} \end{pmatrix}\right\} = \begin{pmatrix} I \\ I \end{pmatrix} x$
- approximating the contribution of the i-th residual to the χ^2/n test statistic:

  $\Delta_i \dfrac{\chi^2}{n} \approx \dfrac{\partial \frac{\chi^2}{n}}{\partial y_{res,i}} \cdot y_{res,i}$, where $y_{res,i}$ = $y_{data,i}$ - $y_{model,i}$ for i = 1...n
- spatio-temporal analysis
- consultation of background information on benchmarks

Exploiting the insight from all these approaches, two issues could be identified that reduce the χ^2/N statistic by a factor of around 50% (see Figure 7). Three more issues with a limited impact could reduce the χ^2/N statistic even further to 6.4. For the investigations described in this addendum, all five of these issues have been appropriately handled by further exclusions of observations.
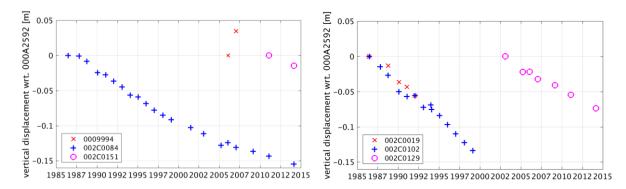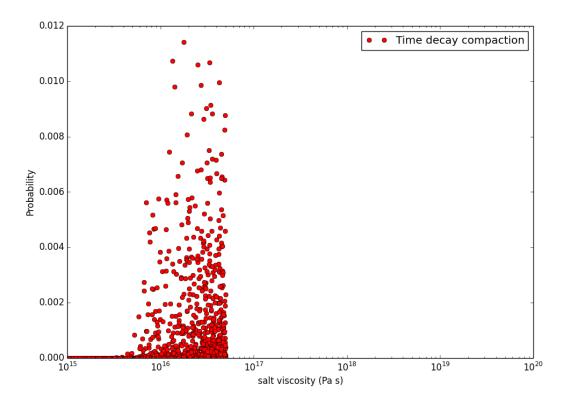

Figure 7: Two outliers with huge impact on the χ^2/n test statistic. **Left:** Benchmark 0009994 shows an uplift of 3.5 cm within one year that is not reflected by the 390 m distant benchmark 002C0084. This obvious outlier has been missed due to a too high threshold in the outlier identification algorithm. **Right:** The relative displacement of benchmarks 002C0019 and 002C0102 differs by 13.5 mm, even though these to benchmarks have a nominal distance of only 10 m. The impact of this outlier is high, because the high precision of the difference is taken into account by the full covariance matrix. This type of outliers cannot be detected by the LTS-1 approach, which neglects spatial correlations.

Inspection revealed that there are more potential outliers in the dataset (identification errors, discontinuities or abnormal autonomous movements of benchmarks). However, these are not as obvious that physical processes in the deep underground can be excluded as their cause with sufficient confidence. A more differentiated outlier assessment is not possible with the objective to create a practically assumption-free dataset prior to geomechanical modelling.

3.2 Constraining the range for salt viscosity based on observation from the Barradeel salt mine

One of the possible explanations for the high values of the test statistic was a discrepancy in range and expectation value for the salt viscosity between Ameland and Barradeel. The range of viscosity used for Ameland spanned from $1x10^{16}$ Pa s to $1x10^{19}$ Pa s, whereas the Barradeel study (pers. com. TNO) found a range of $5x10^{15}$ Pa s to $2.5x10^{16}$ Pa s to match the subsidence caused by the solution mine with the measured data. To test if using such a confined range improves the ESIP results in terms of the test statistic, we did a run using the original parameter ranges for the compaction model (time decay only) and the Young's modulus and Poisson's ratios for the different layers, but confining the viscosity in the range $2x10^{15}$ Pa s to $5x10^{16}$ Pa s. The value of χ^2/N did, however, not improve.

As Figure 70 in the main report (NAM, 2017) suggested already, the members with the highest probability in the AEsubs runs with salt had a viscosity in the range ~$1x10^{17}$ Pa s to $1x10^{18}$ Pa s, which have not been probed in the narrow range here. As a comparison to Figure 70 of NAM (2017), the probabilities for the members tested here, with the Barradeel viscosity range, are plotted in Figure 8, which probability distribution appears to be truncated at the higher end of the viscosity spectrum.

Figure 8; Member probability for a time decay model with AEsubs and salt, using the Barradeel viscosity range.

To look in a bit more detail at the time constants of the models, we plotted the probability of the members versus the Maxwell time (viscosity / Young's modulus) for the original set in and for the Barradeel viscosity range in Figure 9. The characteristic time scales of both the salt and the compaction model span similar ranges explaining why the time decay model showed a slightly wider probability distribution in terms of viscosity compared to linear and bilinear compaction models; the

temporal effect can be exchanged between the salt viscosity and time-decay parameter, allowing for a wider range in both these parameters. In the results in Figure 10 however, the characteristic time scales over which the salt relaxes are much faster, where the high-end of the Maxwell time is constrained to ~10 weeks (0.2 years) with the Barradeel parameters. The salt relaxes much faster in these cases.
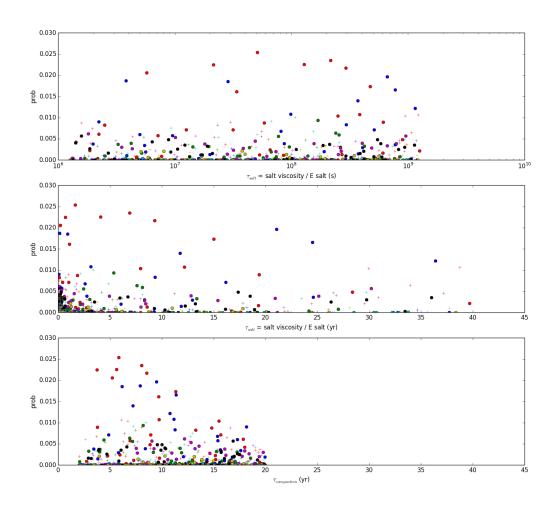

Figure 9: Results for ESIP run using a log-uniform distribution for the salt viscosity in the range 1x10^{16} Pa s to 1x10^{19} Pa s (as originally used). The obtained member probabilities are plotted against the salt Maxwell time for each member (in seconds in the top figure, in years in the middle figure), and the time-coefficient of the time-decay compaction model (bottom figure).
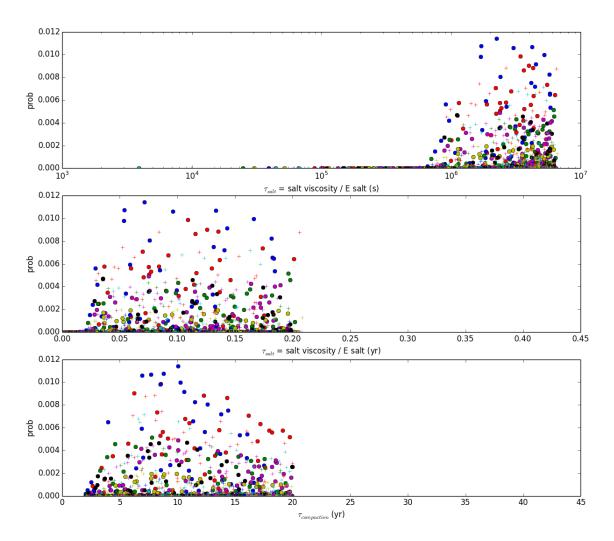
Figure 10: Same as Figure 9but now for an ESIP run using a viscosity range based on the Barradeel parameteres for the viscosity, ranging from 2x10^{15} Pa s to 5x10^{16} Pa s. Note the much faster Maxwell times for the salt in this range (0.2 yr ≈ 10 weeks).

Although constraining the viscosity did not improve the results, to not disregard the Barradeel range, while also including the ranges used in the past in the Geomec models (6.7x10^{17}Pa s), and not deviating too much from the ranges used in other studies [Marketos], we propose to now continue with a viscosity ranging from 1x10^{15} Pa s to 1x10^{18} Pa s, thereby covering both the Geomec and Barradeel ranges.

3.3 Possible water production for the reservoir scenarios used in the study to estimate a different prior probability distribution for the reservoir scenarios

In the original study water production by the model was not included as a parameter to assess model quality. As scenarios with aquifer depletion are likely to result in more water production, a final check was done to see whether some weighing needed to be applied to the various 'history matched' scenarios based on the simulated water production. As it turns out those models with a relatively high aquifer depletion actually forecast only moderate historical water production and hence, even from a water production perspective all subsurface scenarios seem equally likely.

It must be noted that, in order to get aquifer depletion, one needs:

1) and aquifer mobility,
2) and permeable (non sealing) faults,
3) and no residual gas.

Hence, based on this statement one could a priori observe that the likelihood of significant aquifer depletion is smaller than the likelihood of limited to no aquifer depletion. However, there is no clear subsurface data that underpins this general line of reasoning. However, if subsidence data suggest that scenarios with a low aquifer depletion match better, then such a general analysis could be used to support reducing the probability of the high depletion scenarios.

3.4   Offering improved pressure depletion coverage as input to ESIP

When comparing the probability of different pressure members based on the data above the Wadden Sea area, one can observe that in the current set of pressure members used, only one or two members have a little to no depletion away from the gas field, and it is also these two models that show the best match to the data. For a probabilistic workflow like the Red Flag approach to work properly and to return a weighted average that lies close to the data, it is important that none of the tested parameters has its most probable value close to its probed range. In such a case, the weighted average will be biased. Ideally, members are tested with each parameter on either side of the most-likely value. For the pressure depletion, this might not be the case since the best pressure scenarios seem to both lie on the edge of what is currently probed in terms of aquifer depletion.

Additionally, in the full set of ~200 pressure members available, members exist that have a similar low level of depletion in the aquifer but different depletion profiles across the gas part. A broader range of lateral pressure depletion profiles can therefore be probed in the workflow, adding more variability to the development of the subsidence bowl.

We therefore revisited the pressure member selection, to ensure a better and less biased coverage of the pressure depletion, both across the field as well as over time. To compare the models, we looked at the pressure depletion over time on nine locations across the Ameland area, as defined in Figure 11. The depletion profiles of the original selection of 13 scenarios, together with the additional scenarios in the set used here, are plotted at these nine locations in Figure 12. The pressure depletion of all available 193 scenarios is plotted in light blue, while the original 13 scenarios are plotted in blue and the additional scenarios in red. Compared to the original set, the new set gives a better spread of depletion levels particularly at locations 3, 4 and 5.

To demonstrate that the distribution of the full set of 193 scenarios is still honoured, we plotted histograms of the pressures of the selection of 20 scenarios and compared these with the full set of 193 scenarios at these nine locations, for the year 2005 in Figure 13, 2015 in Figure 14 and 2045 in Figure 15. The 2D depletion profiles for the 20 scenarios as they are in 2045 are plotted in Figure 16.
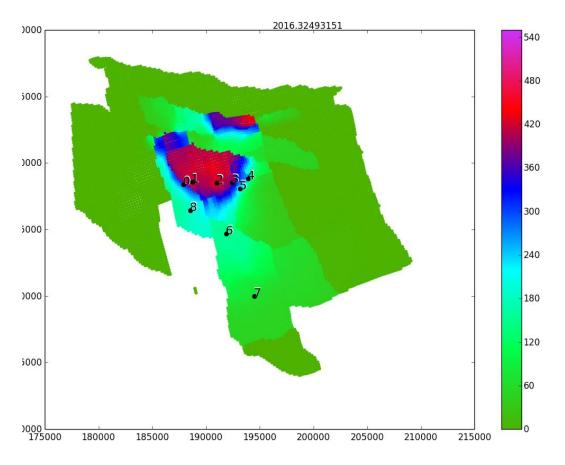
Figure 11: Locations across the field for which pressure histograms are shown in Figure 13 for 2005, in Figure 14 for 2015, and in Figure 15 for 2045, and depletion profiles in Figure 16.
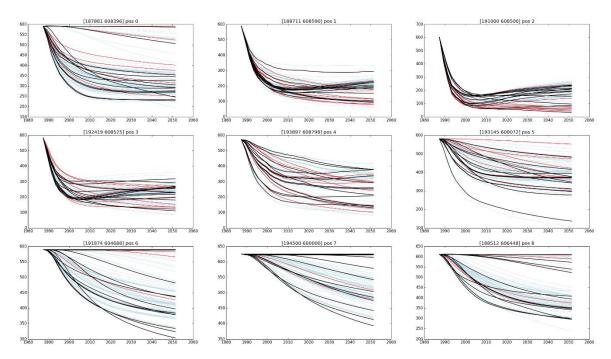

Figure 12: Pressure profiles of the selected members. All available members (with a reservoir engineering based RMS < 45) are shown in light blue. The original selection of 13 pressure depletion scenarios is shown in black, while the additional scenarios in the 20-member selection are shown in red. The added scenarios provide a better coverage of the possible depletion across the area, particularly around the gas field (locations 0 – 5 in the figure).
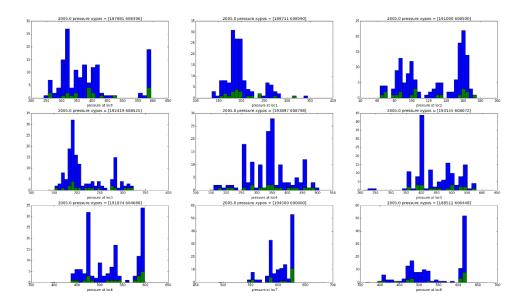
Figure 13; Pressure histograms at nine locations across the field (see Figure 11 for location) in the year 2005. The pressure distributions of the full set of 193 available pressure scenarios is shown in blue, those for the 20 selected pressure scenarios to be used in the workflow in green.
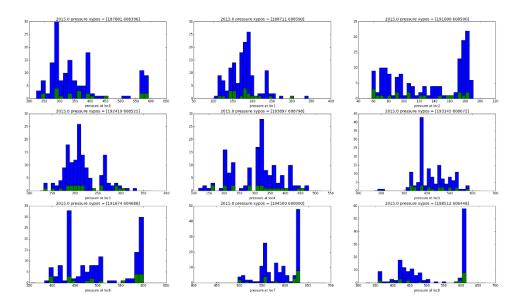

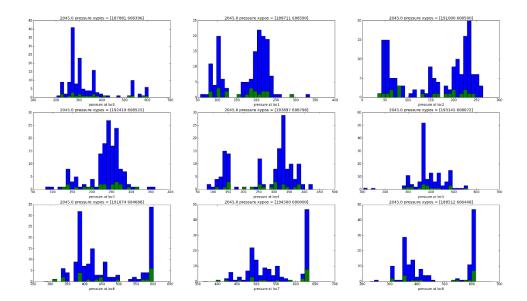*Figure 14; Same as Figure 13 but now for pressures in the year 2015.*

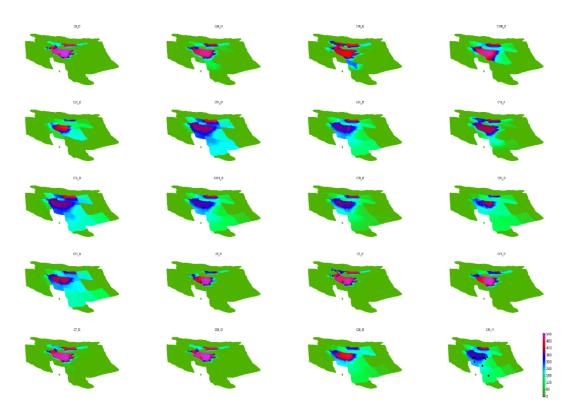*Figure 15; Same as Figure 13 but now for pressures in the year 2045.*

Figure 16; 2D depletion profiles (in 2045) for the twenty models selected.

3.5   Changes to the test statistic and probability function

This paragraph describes the reasoning why we propose to modify the test statistic by bringing in the uncertainty of the geomechanical models. New values for the parametric model that describe the idealisation noise and the removal of previously undetected outliers as described in paragraph 3.1.1 resulted in a significant reduction for the χ^2/N . Still the lowered values are not close to unity, which indicates that the uncertainty in either the data or model is underestimated. Based on the observed mismatch between the data and models on particular the eastern side of the Ameland island, we believe that this uncertainty arises from the subsurface models than from the geodetic data.

There are multiple reasons that can be identified this mismatch like:

- shallow compaction not appropriately described by idealisation noise
- 3D Geometry effects, like a varying salt thickness not reflected by the used models
- Possible viscous and plastic behaviour of overburden layers other than the salt layer.

3.5.1   Changes to the test statistic

These possible factors are not addressed by the current geomechanical models and therefore we propose to bring in a penalty variable in the test statistic that is related to the mismatch between the models and data. The penalty factor in our proposed modification is realised by the addition of a prior $\boldsymbol{C_{mm}}$. This will lead to the following change of the test function:

$$\chi^2 = \left(\boldsymbol{dd}_r^{prior} - \boldsymbol{dd}\right)^T \boldsymbol{C}_{dd}^{-1}\left(\boldsymbol{dd}_r^{prior} - \boldsymbol{dd}\right)$$

to

$$\chi^2 = \left(\boldsymbol{dd}_r^{prior} - \boldsymbol{dd}\right)^T (\boldsymbol{C}_{dd} + \boldsymbol{C}_{mm})^{-1}\left(\boldsymbol{dd}_r^{prior} - \boldsymbol{dd}\right)$$

The $\boldsymbol{C_{mm}}$ will be compiled in simplified form by bringing in an identity matrix and variance multiplication factor based on the standard deviation that is retrieved from the residuals probability density function as visualised in Figure 17 . The red lines in this figure indicate a value of the standard deviation of around 1 cm.
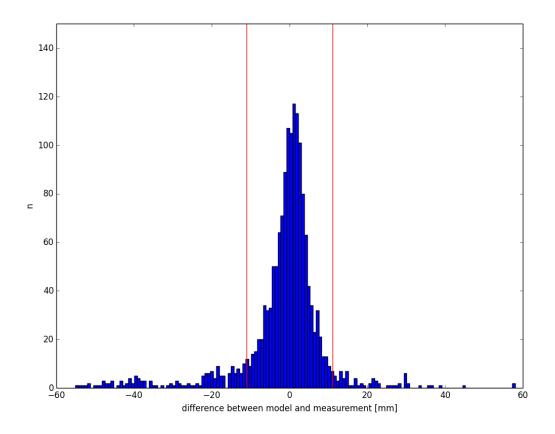
Figure 17; Probability density function of the residuals between model and data for the member having the lowest χ^2/N value

3.5.2 Changes to the probability function

We described in paragraph 0 the identification of an error in the equation for the Bayesian probability:

$$P(dd_r^{prior}|dd) = \frac{P(dd_r^{prior}) \cdot P(dd|dd_r^{prior})}{\sum_{i=1}^{Ne} P(dd_i^{prior}) \cdot P(dd|dd_i^{prior})}$$

$$P(dd|dd_r^{prior}) = exp[-\frac{\chi^2}{2N}]$$

With the correct formula for $P(dd|dd_r^{prior})$:

$$P(dd|dd_r^{prior}) = exp[-\frac{\chi^2}{2}]$$

The application of the correct formula results in a collapse of the ensemble into few member having a high probability and all other members having a probability of 0. This means that even in the case of χ^2 values being close to each other, the weighing by this probability definition causes an extreme discrimination between them, which is a counter intuitive result. This feature of particle filtering methods like Red Flag is observed as well in many other disciplines in case where a large set of independent variables are used (e.g. Snyder et al. 2008).

Besides the introduction of C_{mm}, a solution may be found in using the χ^2 distribution itself to estimate the probability for a specific member (i.e. the probability of finding that value for χ^2). We would then obtain

$$f_\nu(u) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}u^{\frac{\nu}{2}-1}e^{-u/2}$$

where u is the squared sum of normalised double difference residuals also referred to as χ^2, and ν is the number of residuals. For large ν, the χ^2 distribution can be approximated by a normal distribution with expectation ν and variance 2ν:

$$P(dd|dd_r^{prior}) \propto \exp\left\{-\frac{(\chi^2 - \nu)^2}{4\nu}\right\}$$

This distribution would give a reduction of the penalizing effect of single points deviating severely from predicted values. The reduction is presumably less than upon using an additional factor ν in the denominator of the exponent of the original equation, because of the extra power of 2 in the numerator in the new formula. The effect the introduction of C_{mm} plus this distribution function compared to the Red Flag method is shown in Figure 18.
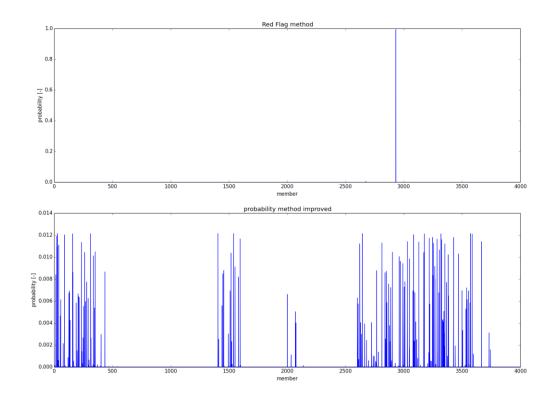

Figure 18; probability per member for the Red Flag method, showing above an ensemble collapse using the original equation and below showing more members with a higher probability following the proposed probability calculation method.

3.5.3 Analogy with Geodetic Adjustment and Testing theory

In this paragraph the analogy of the test statistic in paragraph 2.5.1 and the probability density function in paragraph 2.5.2 is shown with geodetic adjustment and testing theory as described in Teunissen (2000a) and Teunissen (2000b).

The application of geodetic adjustment and testing theory starts with the formulation of the functional model and the stochastic model. The functional model describes the relation between the observations and the unknown parameters that need to be estimated, whereas the stochastic model describes the precision of the observations.

We start from the vector of observations and its corresponding covariance matrix:

$$\underline{y} = \begin{bmatrix} \underline{dd}_{dd} \\ \underline{dd}_{mm} \end{bmatrix} \quad Q_y = \begin{bmatrix} C_{dd} & 0 \\ 0 & C_{mm} \end{bmatrix}$$

with the vector $\underline{y}$ comprising of the 'observed' double-differences from the geodetic data ($\underline{dd}_{dd}$) and the geomechanical modelling ($\underline{dd}_{mm}$). The precision of the 'observed' double-differences is described by the covariance matrices C_{dd} and C_{mm} for respectively the geodetic data and the geomechanical modelling. The covariance matrix of the geodetic data contains the measurement precision of geodetic technique (and additionally the idealisation precision), whereas the covariance matrix of geomechanical modelling represents the uncertainties in the surface displacements originating from uncertainties in the subsurface parameters. The stochastic models of the geodetic data and geomechanical modelling are here considered to be independent and hence uncorrelated.

The unknown parameters we define as the double-difference surface displacements imposed by hydrocarbon extraction. Under the hypothesis that both the geodetic and the geomechanical modelling double-differences represent the unknown double-difference surface displacements, we can formulate the functional model in terms of condition equations, stating that the expectation value of the misclosures (t) between the vectors $\underline{dd}_{dd}$ and $\underline{dd}_{mm}$ is 0:

$$B^T = \begin{bmatrix} -I & I \end{bmatrix} \quad \underline{t} = B^T \underline{y} \quad B^T E\{\underline{y}\} = 0$$

$$Q_t = B^T Q_y B = C_{dd} + C_{mm}$$

where B is the matrix containing the conditions on the observations, I is the identity matrix, and Q_t the covariance matrix of the misclosures ($\underline{t}$). Q_t can be obtained by the propagation law, assuming normally distributed observations. Since the geodetic and geomechanical modelling double-differences are considered uncorrelated, Q_t is the addition of the covariance matrices of the geodetic data and the geomechanical modelling.

In geodetic testing theory (hypothesis testing on the functional and stochastic model), the test statistic $\underline{T}_{q=v}$ is defined as:

$$T_{q=v} = \underline{t}^T Q_t^{-1} \underline{t} = (\underline{dd}_{mm} - \underline{dd}_{dd})^T (C_{dd} + C_{mm})^{-1} (\underline{dd}_{mm} - \underline{dd}_{dd})$$

$$\underline{T}_{q=v} \sim \chi^2(v,0)$$

where v equals the redundancy in the estimation (assuming the matrix B has full rank), equal to the number of conditions (misclosures), which is equivalent to the number of observations minus the number of unknown parameters. The test statistic $\underline{T}_{q=v}$ has a central chi-squared distribution with v degrees of freedom. Please note that $\chi^2(v,0)$ denotes the chi-squared **distribution**, which should not be confused with the χ^2 **test statistic variable** as defined in paragraph 2.5.1:

$$\chi^2 = \left(dd_r^{prior} - dd\right)^T (C_{dd} + C_{mm})^{-1}\left(dd_r^{prior} - dd\right)$$

Comparing the LTS-2 χ^2 test statistic to the $\underline{T}_{q=v}$ test statistic definition, one can see that they are similar, since dd_r^{prior} is a specific realization of geomechanical model double-differences $\underline{dd}_{mm}$, and dd is the vector of geodetic data double-differences $\underline{dd}_{dd}$.

For large degrees of freedom (v) the chi-squared probability density function can be approximated by a normal distribution with expectation value μ equal to v, and standard deviation σ equal to √(2v):

$$p(\chi^2) \approx \exp\left\{-\frac{(\chi^2 - v)^2}{4v}\right\}$$

Instead of the test statistic $\underline{T}_{q=v}$, the Overall Model Test is generally used in geodesy to assess both the functional and stochastic model considering the observations. The Overall Model Test statistic is defined as:

$$\hat{\underline{\sigma}}^2 = \frac{\underline{T}_{q=v}}{v} = \frac{\underline{t}^T Q_t^{-1} \underline{t}}{v}$$

which has a Fisher distribution,

$$\hat{\underline{\sigma}}^2 \sim F(v,\infty,0)$$

and an expectation value of 1 (both functional and stochastic model are correct).

Note that the Overall Model Test is equal to the LTS-2 test statistic χ²/N (when including the covariance matrix of the geomechanical model C_{mm}), where N is equal to the number of double-difference misclosures (residuals) between the geodetic data and the geomechanical model realization.

Summarizing, it can be concluded that the LTS-2 test statistic χ^2/N (when including the covariance matrix of the geomechanical model C_{mm}) is equal to the geodetic Overall Model Test as formulated in this paragraph.

4 Implication of findings on subsidence rates and emergency scenarios for the Pinkegat sandsharing area

A final proof of the applicability of the work flow to possible subsidence predictions above the Ameland field is a comparison with the published results in the measurement and control documents (M&R, e.g. NAM, 2016). The Pinkegat sand sharing area is one of the two areas that is monitored for subsidence in this procedure and the one that is dominated by the influence of the Ameland subsidence bowl. The standard M&R procedure to calculate the contribution of the subsidence due to gas production in a certain (sand-sharing) area is that first the subsidence volume per year is calculated. The contribution of this volume in a certain area is subsequently divided by the surface area to derive an average subsidence rate. Finally, a 6-year moving average of these results allow for a better assessment of possible ecological effects in the Wadden Sea. This is what is presented in most of the figures in this chapter.

This chapter we will describe the results when applying the modified work flow to the Ameland case.

The amendment to the ESIP calculation, explained in previous paragraph, is in summary:

- A more appropriate definition of idealisation parameter set and a more rigorous outlier detection procedure (paragraph 3.1).
- A more representative distribution of the reservoir scenarios (paragraph 3.4).
- A modification of the work flow explained in paragraph 2.5

The implementation of these modifications led to the recalculation of both the χ^2 and probability for all members. It turned out that some members got χ^2/N values below 1. These members would get an unreasonable lower weighting in the equation and therefore we used a condition that set all values lower than 1 to 1. The probability is directly related to the weight of each member to both assess a weighted average as the 95% confidence bounds, determined by a 2.5% cut off on the posterior pdf. All members per influence function have been combined in this case to span the complete ensemble (Figure 19). The total ensembles for the different influence functions show a 95% confidence range that is relatively narrow after the application of the weight to these members. The same figure points out as well that the expectation case, based on a weighted average of the posterior pdf of all members is close to the Geomec line (Geomec M&R) except for the Knothe influence function. Especially, the temporal behaviour of the "Knothe ensemble" is different when comparing to the other models. This is a result of the Knothe ensemble giving more weight to the reservoir scenarios with aquifer depletion (see also Figure 21) caused by the possibility to change the angle of the subsidence bowl in the Knothe function. By allowing this freedom, the Knothe model tends to produce narrow subsidence bowls with low values for the compaction coefficient (Cm). A low Cm value still gives a reasonable match in the Wadden Sea for those reservoir scenarios with higher aquifer depletion. Most of these scenarios start to deplete in a later stage than the depletion of the gas field resulting in a delay of the subsidence shown by the green line in Figure 19.
The Geomec line in this figure results from a model with a Time decay compaction model and a viscous salt. More important is that the model is based on a reservoir scenario with limited aquifer depletion. We conclude that this scenario is confirmed by the LTS-2 result with the knowledge that also scenarios with high aquifer depletion were evaluated in the same confrontation.
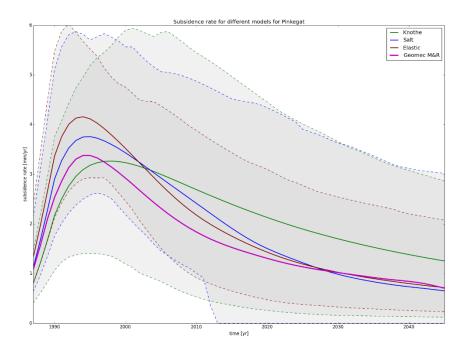
Figure 19; Weighted average and 95% confidence bound for 6-year averaged subsidence rates for all members in the Pinkegat sand sharing area.

Using this ensemble, we can also calculate the effects of the hypothetical production stops for 1996 and 2016 in the Pinkegat sand sharing area to demonstrate the effectiveness of the hand on the tap procedure (Figure 20). This figure shows clear deviations (represented by the dashed lines) of the expectation case from the "normal" production case for all ensembles, especially for the hypothetical stop in 1996. This case is more representative for the other Wadden fields like Nes and Moddergat in a scenario where we would halt the production in in the near future.
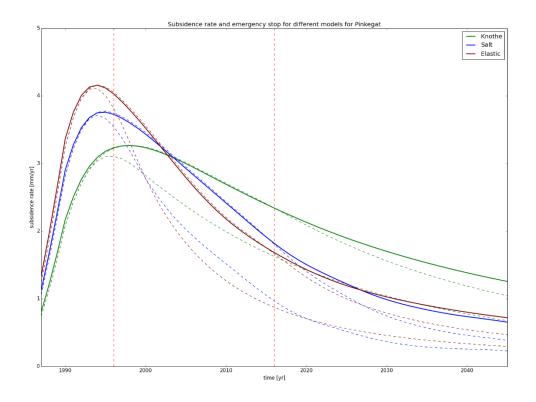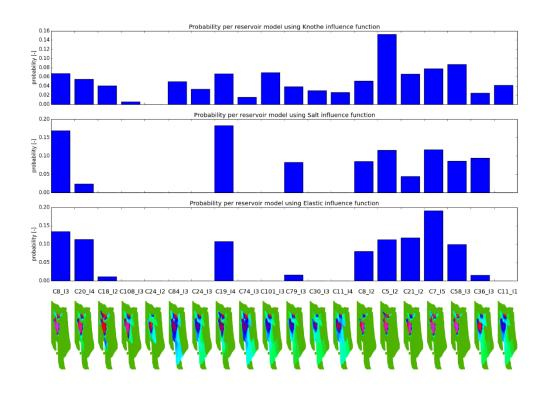
Figure 20; effect of a hypothetical production emergency stop of the Ameland field in 1996 and 2016.

The close match to the Geomec results gives us confidence that the modified work flow is able to make a distinction in the likelihood value of reservoir scenarios with or without aquifer depletion. The Geomec result is based on a scenario with limited depletion and the fact that the weighted averages of the ensembles are close to this result implies that the posterior likelihood for these scenarios is relatively high. To visualize the capability of the work flow making a distinction between the reservoir scenarios in more detail, we made a probability plot for each reservoir scenario (Figure 21) for the three influence functions. This plot shows in the case of an elastic and salt influence function, a clear distinction between the scenarios with high aquifer depletion and the scenarios with low to limited aquifer depletion. Therefore, we conclude that the work flow is capable of selecting only those scenarios with no or limited depletion in the aquifers south of the Ameland gas field.

Figure 21; posterior probability value for each reservoir scenario (bottom row) for based three ensembles characterized by the influence functions.

The three ensembles that were discussed in the previous section can be combined into one big ensemble to assess the uncertainty and expectation case, based on a 95% confidence bound and weighted average of the total ensemble. This final result is presented in the Figure 22 and Figure 23 with again the conclusion that the average line is close to the results that were published in NAM (2016a).
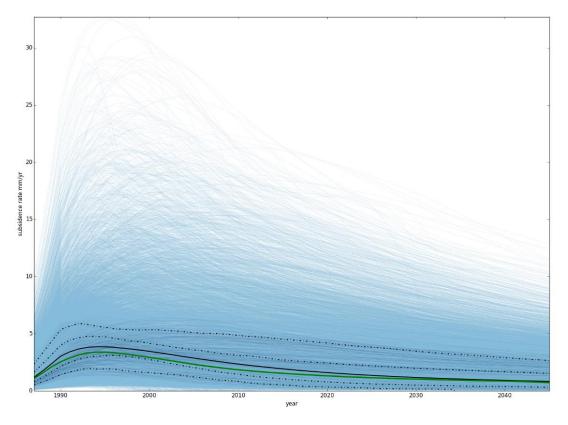
Figure 22; 6-year moving average of the Pinkegat subsidence rates for the total ensemble that is confronted to the data showing in black the posterior stochastic expectation case with ~1σ and ~2σ uncertainty bandwidth.
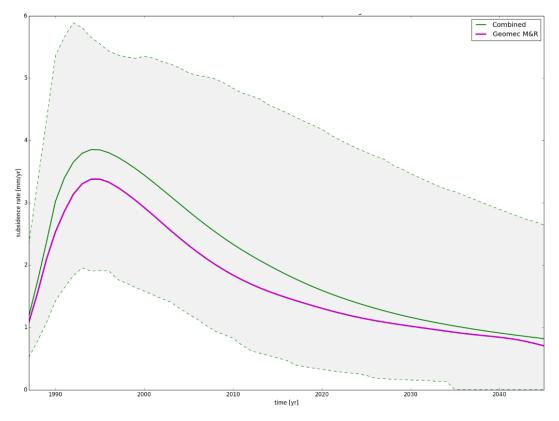

Figure 23; Same figure as Figure 22, now zooming in on a 95% confidence bound of the result.

4.1 Outlook to possible implementation of the work flow in the measurement and control cycle for the Wadden fields.

One of the objectives of this study was to investigate the potential application of a similar work flow in the measurement- and control cycle procedure for the Wadden fields. In this procedure we would then calculate a more objective expectation case then currently used "base case" (NAM, 2016a) to compare to the ecological limit as defined for the two sand-sharing areas Pinkegat and Zoutkamperlaag.

With the application of the work flow, we demonstrated how to derive an expectation case that is honouring both the observations on the island and the observations in the Wadden Sea. This expectation case closely matches the base case that is reported in the measurement and control documents of latest years, which can be seen as an independent quality check of the reservoir scenario and Geomec model used so far. However, the Red Flag theory had to be modified in order to arrive at these results. The applied modifications are based on first principles but the blending of the geodetic testing theory based on condition equations with the Bayesian theory is a novel approach. With the given methodology we are in a position to apply the methodology to other gas fields with the remark that a next level of complexity will arise from the introduction of multiple fields in this application. While the Ameland gas field is relatively isolated we can't make a similar statement for the other Wadden fields. In this case data points are effected by multiple fields that possibly require the development of one large Ameland/Wadden reservoir- and geomechanical model. Also it is likely that parameter ranges need to be varied per individual reservoir rather than using a uniform parameter set for all fields with a single Monte Carlo draw of parameter values. These are non-trivial and fundamental requirements that have to be implemented in the code, which requires a phase of proper testing on dummy models like what was done in the LTS-2 project. All these efforts seriously challenge the feasibility to implement such a work flow into the "Meet- en regel" cycle over the year 2017.

Besides these fundamental changes to the work flow, the LTS-2 consortium indicated that also the geomechanical models could be improved. At the start of the project, we discussed that the test statistic should have a value of around unity, an objective not being met using the original Red Flag equation. To achieve this, we added a penalty parameter that in essence introduces a model noise component in the test statistic. In other words, we need to add more noise when our models deviate further from the observations to achieve a test statistic value close to unity that leads to larger uncertainty bandwidths. The value for the variance factor in the noise parameter is currently around 100 mm^2 matching the standard deviation of the difference between the best model and the data. During the meetings with the LTS-2 consortium opportunities were identified to improve the geomechanical models to reduce the value of the residuals:

- Test the geometry effects of the 3D Geomec FE model in combination with lower values for the salt viscosity. This could lead to a shift of the bowl toward the north, also indicated by models shown in the LTS-1 study. A confirmation by data could possibly being achieved by investigating the horizontal GPS movements in more detail.
- Apply modifications to the current Time decay and RTCiM compaction models. The Ameland and Wadden field are characterized by an over pressure of around 200 bar. This overpressure may have an influence on the geomechanical behaviour of the reservoir rock not currently reflected by the aforementioned compaction models. Currently the derivation of the time dependent constants of both models are dominated by the compaction

mechanism (mainly elastic) of the field in the early years while this behaviour may be different from the compaction mechanism (elastic plus plastic) exceeding the pre-consolidation stress, which is the highest effective stress seen by the reservoir rock before over pressuring as a result of the large gas column in that area.

Improvements to the model should lead to a closer fit to the data and therefore a lower variance factor and hence a narrower uncertainty bandwidth.

5 Conclusions

This addendum describes the modifications that have been incorporated in the work flow since the publication of the NAM (2017) report. We have demonstrated with the proposed and agreed (by the LTS-2 consortium) changes that we are able to apply a stochastic workflow to a real field case with many independent observations judging on the probability of each member in the ensemble.

An important objective that was missing in the main report was the comparison to previously published results in the measurement and control documents for a known area like the sand-sharing area Pinkegat.

These modifications resulted in the following findings:

- The method applied results in a clear preference for the reservoir scenarios with limited aquifer depletion.
- A weighted average forms the expectation case and this result matches satisfactorily the model and forecast published in NAM (2016a), which gives confidence in the subsidence predictions above the Wadden Sea. The confrontational method is based on the original Red Flag method (Nepveu et al. 2010) but with modified equations for both the goodness-of-fit (χ^2/N) and probability equation.
- The "remweg" scenarios tested with this method show a clear deceleration of the subsidence rate over the Pinkegat sand sharing area.

6 References

Houtenbos, A.P.E.M., F. Kenselaar (2001): Peilmerk hoogte variaties – stochastische analyse van peilmerkbeweging in Nederland. TUDelft, 2001.

van Leijen F., S. Samiei-Esfahany, H. van der Marel, R. F. Hanssen (2017): Uniformization of Geodetic data for deformation analysis – Contribution to the research project: Second phase of the long-term subsidence study in the Wadden Sea Region (LTS2). Delft University of Technology.

NAM (2016a) Gaswinning vanaf de locaties Moddergat, Lauwersoog en Vierhuizen. Resultaten uitvoering Meet- en regelcyclus 2015.
http://www.commissiemer.nl/projectdocumenten/00000929.pdf?documenttitle=MR%20rapportage%20%202015.pdf

NAM (2017) Ensemble Based Subsidence application to the Ameland gas field – long term subsidence study part two (LTS II)

Nepveu M., Kroon, I. C., Fokker, P. A. (2010) Hoisting a Red Flag: An Early Warning System for Exceeding Subsidence Limits. Mathematical Geosciences. February 2010, Volume 42, Issue 2, pp 187–198

Samiei-Esfahany S., H. Bähr (2015): Research and Development Project for Geodetic Deformation Monitoring – Contribution to the research project: "Long-term study on anomalous time-dependent subsidence in the Wadden Sea Region". NAM report EP201505216980.
http://feitenencijfers.namplatform.nl/download/rapport/6af45eca-b02f-4844-8adf-7835b5e0ee55?open=true

Samiei-Esfahany S., H. Bähr (2017): Erratum to the report: "Research and Development project for Geodetic Deformation Monitoring" – Revision of recommendations from the project: "Long-term study on anomalous time-dependent subsidence in the Wadden sea region". NAM report EP201701215912.

Snyder C, Bengtsson T, Bickel P, Anderson J. (2008) Obstacles to high-dimensional particle filtering. Monthly Weather Review 136: 4629-4640.

Teunissen P. J. G. (2000a): Adjustment Theory. VSSD, Delft.

Teunissen P. J. G. (2000b): Testing Theory. VSSD, Delft.

TNO (2011) Toetsing van de belasting op de gebruiksruimte in de kombergingsgebieden Pinkegat en Zoutkamperlaag door bodemdaling ten gevolge van gaswinning onder de Wadden Sea. TNO-060-ur-2011-02035/C
http://nlog.nl/bodemdaling