(in this document, **all reviewer comments are in bold**, whereas our replies are in standard font)

Comments on the paper Hess-2019-685 entitled: “A robust objective function for calibration of groundwater models in light of deficiency of model structure and observations”, by R. Schneider et al.

Reply:
We would like to thank the reviewer for their thorough review.

For better understanding of the overall changes, we want to start with the overall plan on how to revise this manuscript (parts of these changes have been inspired by the responses by the other reviewers):

1. We will add a synthetic calibration test based on the Storå model showing the ability of the CRPS compared to other metrics
2. The synthetic test will also include the mean absolute error (MAE) and mean root error (MRE) as objective functions. The CRPS, MAE, and MRE are all performing similarly well in our test case; and clearly better than the conventional MSE. Based on the higher sensitivity towards bias of the CRPS compared to the MAE and MRE, we still prefer the CRPS.
3. We will remove the Odense from most of the presentation of the “real-world” results to reduce redundancy, and only use it for a kind of “proxy-basin” validation test

All further changes we intend to make are outlined below and in the replies to the other reviewers. Overall, we are thankful for the feedback from all three reviewers and are convinced that the manuscript will benefit significantly from the revisions.

We believe that we can add significantly to the manuscript by incorporating some changes suggested by the reviewer, and improve where the reviewer pointed out some lack of clarity. Our specific replies follow below.

This work intends to show that the classical objective functions (OF) in the inversion of subsurface flow, such as the sum of squared errors (SSE) between simulated and observed heads, or the sum of absolute errors (SAE), are functions mainly dominated by a few large errors. If these errors are stemming from structural model discrepancies, then the inversion procedure would compensate on model parameters to lower the OF, but with sometimes the downside of rendering awkward or unphysical solutions. Therefore, it is proposed to rely upon an OF based on the continuous ranked probability score (CRPS) reputed less sensitive to large residuals, as it measures the squared distance between the cumulated probability density (cumulated statistical distribution) of model outputs and its equivalent in terms of local observations (usually, a Heaviside function). A few examples of this reduced sensitivity to high residuals are provided on the basis of very simple examples such as a series of five values, or a continuous Gaussian distribution. Then, a comparison of CRPS, SSE, and SAE is carried out for two inversion problems dealing with actual watershed systems.
It seems interesting employing the CRPS, usually devoted to the analysis of multiple equiprobable realizations of a single variable, in the framework of a single realization of a single variable but distributed over time and space. However, in my opinion, the study partly misses its target because the applications are a priori free from model structural errors; at least, these errors are not explicitly considered in the analysis of the inverse sought solutions.

Reply:
We would like to point out again that our suggested use of the CRPS as an objective function instead of the SSE due to its lower sensitivity to large residuals is not only owing possible structural errors, but also uncertainties in the observational data used as target in the inversion. As we for example write in lines 55 ff, we are confronted with data of unknown quality. Almost inevitably, there will be some (significant) observation errors in some of the observations, however we often cannot identify them. Hence, we argue

that there is a point in having an objective function that is less sensitive to large residuals. One alternative, that has been used sometimes (we provide some examples from literature in line 79), is to discard observations that deviate too much from the modelled values or a conceptual understanding of the modelled system. We think that this is an arbitrary, and, hence, undesired way of filtering data, potentially removing valuable information from the inversion or the later model validation. Instead, we want to use an objective function less sensitive to large outliers.

In our revised manuscript we will be more clear about the two issues we are tackling: i) structural errors, and ii) observational errors.

It is correct that the structural errors are not explicitly considered in our models; this is not/hardly possible for such large-scale practical applications. However, structural errors will still have an impact on the achievable model fit in parts of the model (for example due to a missing geological layer/lens, or a wrong boundary condition). Again, here we hope that an objective function less sensitive to large residuals will yield in a better solution of the parameter estimation.

I have a few concerns of various importance with the present writing, and some specific points (in a non-exhaustive inventory) that let me think that the contribution is not mature enough for rapid publication in HESS. My suggestion is that the paper needs major revisions, including new numerical investigations, and a further complete round of review.
Regarding the main concerns:

1- **The notion of structural error is not well defined. In a first approach, one could consider that structural errors are all the errors that do not directly target model parameter values. This could include: errors on the geometry of the modeled system, on initial and boundary conditions, and on source-sink terms. One could also consider that structural errors are those associated with features hardly inverted in view of their direct influence on the observed state variable. In that case, one could remove initial and boundary conditions, but also source-sink terms from the structural errors, as these characteristics of the model can be inverted in view, here, of hydraulic head measurements. Finally, in the specific case of the reported study, the model parameterization relies upon a parameterization of the zonation type, building a “block” system with uniform parameters within each block. A flawed delineation of these blocks could also be considered as a structural error. Even better, one could suggest that for the two actual tests cases reported by the authors, errors in the delineation of uniform blocks could be the main structural error generating high residuals on heads that will never be compensated by tuning the model parameters of each block. In the end, it seems important to better state what is meant by structural errors, then deliberately generate these errors in exploratory calculations before checking on the performance of a CRPS-based objective function.**

Reply:
We understand that there is a need to be more clear around our use of the term “structural error” – we will clarify this in the revised manuscript. Based on our experience, though, the delineation of the geological zonation together with drainage representation and unsaturated zone description should be the largest contributor to structural uncertainty. The influence of initial conditions should be negligible, as we are considering hotstart and warmup periods when running our models. Boundary conditions have an impact, however we expect it to be moderate: along land-boundaries of the model they are taken from larger models; the sea boundary conditions are straight forward (fixed head = 0m).

In general, we consider all types of structural errors the reviewer mentioned – however, most of those cannot be properly disentangled in large-scale hydrologic models in practical applications such as ours.

2- **The two actual test cases discussed in the paper are redundant, mainly because they deal with watershed systems of the same size, with the same density of streamflow routing in their surface compartment, and a very similar density of evenly spread locations monitoring the subsurface waters. Why to report on both? The authors would have been well advised to focus on a single system, and consider that an inverse solution becomes some kind of reference problem to which structural errors are added. Here, the first structural error I would give a try would be that of a flawed parameter zonation. Then, by providing us with a metric on model parameters distinguishing values inherited from the "reference" and the "flawed" problems, some proofs that CRPS outclasses SSE and SAE could be made available.**

Reply:
We do understand the reviewer's concern regarding the redundancy of the two case studies. The main reasons for including both cases is that we (i) wanted to show some level of reproducibility and robustness, and (ii) the two cases, geologically speaking, actually are different, with the Storå catchment being relatively flat and more dominated by sand in the uppermost layers, and the Odense catchment being more hilly and dominated by clay. Also in light of extending the paper by adding a synthetic experiments (based on the Storå model), we plan to follow the reviewer's suggestion and remove most of the results of the Odense model in the revised version of the manuscript. I.e. remove the Odense results and respective discussion from Table 1 and 2 and Figure 4 and 5. We plan, however, to keep Figure 7, though potentially simplify or merge it with Figure 6. The motivation will be mentioned more explicitly in the revised manuscript.

Furthermore, we will add a synthetic experiment, based on the Storå model case, to show that the CRPS performs as stated in cases with flawed observations. These experiments are performed with synthetic observations sampled at the exact same locations and points in time as the real observations. The synthetic observations are sampled from a model run with a specific parameter set, and some white noise is added to all of them to account for general measurement uncertainty. Moreover, some of the observations are perturbed further with a systematic bias. Then, the model is calibrated against the perturbed synthetic observations, starting from a different parameter set. The calibration experiments are carried out with either the CRPS, mean root error (MRE), mean absolute erroer (MAE), or SSE as an objective function. In these synthetic calibration experiments, it could be shown that the CRPS behaves as expected by us – the parameter set resulting from the CRPS-based calibration is closer to the parameter set of the synthetic truth, than the parameter set resulting from the SSE-based calibration. For more details, we would like to refer to the reply to reviewers 1 and 2 who also mentioned the need for synthetic experiments.

3- **My understanding is that in many locations within the modeled system, the authors (for the principle of parsimony?) lump the measurements of heads at various times and in various layers of the subsurface to build an averaged information. I doubt that this information has the sensitivity of a single observation to both parameter and structural errors. Let us take for example the case of a point measurement of head located not so far from a boundary condition. This condition is flawed and prescribes a Neumann-type boundary with prescribed fluxes instead of a Dirichlet-type condition with prescribed head. The Neumann flux is not sufficient in the wet periods to feed the system, but too high in the dry periods, thus rendering negative (positive) errors on the head at a short distance in the winter compensated by positive (negative) errors in the summer. As a result, the structural error is not seen by the data, as would render the true Dirichlet condition able to feed the system at will. This example is just for showing that averaging various measurements is probably not a good idea to reveal that structural errors exist. I must acknowledge that I have never seen in the literature inversions taking averaged errors over large**

periods at some locations as the basis for an OF. I guess that it is “dangerous” to proceed that way, but probably my knowledge of the literature is not sharp enough.

Reply:
We do lump the head measurements to model grid cells. This means that we lump observations from various wells, if they fall into the same model grid cell, and we lump multiple observations in time per well. However, we do not lump observations across various layers of the subsurface; observations from different model layers will fall into different model grid cells, and hence will remain individual observations. This will be made more clear in the revised manuscript to avoid misunderstandings.
We do agree with the reviewer’s concern that aggregating the information within each grid cell from different points in time risks the loss of some information, in particular the potential for a compensation of negative and positive residuals. However, this can only occur in very specific cases, where (i) a timeseries is available (ii) without any bias in the simulated groundwater heads, but only an error in the simulated groundwater head amplitudes (usually from seasonal effects).
However, we try to explain our motivation for such an aggregation:

- We do only aggregate to the smallest spatial unit the model can resolve – a single model cell.
- In many cases, we aggregate single observations (not time series) from different wells within one model grid. These observations actually often “contradict” each other, for example by showing a positive residual of a few metres in the one well, and a negative residual of a few metres in the other. Reasons for that can be anything from heterogeneities within one model cell not being described in the model, small scale topographic variability, observations errors, etc. We cannot identify the exact reason, or identify which observations are most valid. Therefore, we still want to use all available data, and assume that it is reasonable to trust the aggregated mean per model grid cell in this case.
- Typical seasonal variations are in the range of ~1m (or less), whereas our residuals are in the range of a few metres. I.e. our typical model residuals are significantly larger than seasonal variations, which reduces the information lost from aggregation in time.

In practical applications, at least with large-scale models and large datasets of varying origin, quality, and spatiotemporal resolution it seems common to, for example, aggregate all observations from one well into one residual, i.e. aggregate over time (Sonnenborg et al., 2003).

4- **The authors employ the same cumulative distribution of residuals to build their OF, irrespective of the location where the distribution is used to measure the performance of the model. This implies that the distribution of residuals should be stationary over space (which differs from the assumption of ergodicity associated with the inference of a CRPS on the basis of a single realization, but could also go with…). I doubt that in the presence of structural errors, e.g., local errors on the system geometry, or its boundary conditions, the statistical distribution of residuals would be stationary. If the authors are right, the distribution should not be stationary in being skewed toward high residual values in regions under structural errors. By the way, the CRPS should give less weight to important residuals in regions where structural errors are plaguing the convergence of the inverse problem by only tuning the model parameters.**

Reply:
Unfortunately, we are unsure whether we fully understand this comment. We will try to respond to the reviewer, but also ask them for clarification if we misunderstood.
The CRPS is calculated across all residuals in the model. It is correct that we do not assume that the residuals are distributed stationary over space (concerning ergodicity, please refer to the answer to the first comment of reviewer 1).

Concerning the last sentence in this comment: Yes, that is exactly our point: The CRPS should give less weight to large residuals in regions with structural errors (or observation errors). Hence, it is less prone to parameter compensation, potentially allowing an easier identification of such areas in the model after the calibration.

In addition to the above general comments, I have a few specific comments (a nonexhaustive list), mainly as the consequence of lack of clarity in the writing.

1- **Line 111, Eq. 1. The CRPS seems to be not well defined if it is supposed to serve as an indicator concealed in an OF. With an integral from minus infinity to plus infinity and an expected value of zero (optimal residual) the CRPS will remain the same irrespective of the location where it is applied. My understanding is that for a variable X (here a residual) and an associated bound x, the CRPS should write as the integral between minus infinity and x of (Ps(x')-Po(x'))^2dx', with Ps(x') the probability for the variable X of not exceeding the value x'. In this case, and for a residual value x at a given location, CRPS(x) measures the distance between x and zero.**

Reply:
Equation 1 gives the general, original definition of the CRPS (except for formalities identical to the formulation in the cited (Gneiting and Raftery, 2005) and (Hersbach, 2000)). We are not sure we do fully comprehend the reviewer's concern. Maybe there is some confusion arising from the fact that, usually, the CRPS is applied to the difference between Ps and Po, where Po is the true/observed value – whereas in our case, as we are dealing with residuals and not absolute values, Po is zero.
We think the reviewer could have misunderstood how we apply the CRPS (also concerning the comment above) – it is not applied to every "location", but is applied across all observations of the entire model. We are happy and confident to fully clear this up with the reviewer in the following round.

2- **Lines 135-143, Eqs 2 and 4. If the significance of the dPi is well exemplified in Fig. 1 (with differences between the left panel (CRPS) and the right panel (MSE)), the text does not mention this difference. In a CRPS dPi is the cumulated probability of not exceeding the value xi, when dPi in a MSE is the probability of x being within an interval bounded by xi-1 – xi, or something of the kind. I would change the notation to avoid misunderstandings and be clear on that in the main text.**

Reply:
Yes, that is correctly understood. This is mentioned in the figure caption, though maybe not clear enough. Will be explained better in the text as well.

3- **Line 169. What means "a description of the unsaturated zone" in MIKE SHE, a simplification, a 3-D resolution of the Richards equations? A short explanation should be given as a reminder. Integrated hydrological models coupling surface and subsurface flow have many options to handle the subsurface including the vadose and the saturated zones, and very often these options condition how two different models respond differently to the same forward problem.**

Reply:
Yes, the description of the unsaturated zone is a crucial part of such coupled models, we will add a few more details in the revised version of the manuscript. The described models are using the 2-Layer method of MIKE SHE. This is a relatively simple description of the unsaturated zone including the processes of interception, ponding and evapotranspiration while simplifying the entire unsaturated zone to two layers (DHI, 2019, p.27).

4- **Section 3 "Model and data". As told earlier, I think that presenting a single model for a single study area would be enough. In general, the overall depiction of the models in terms of hydrological context is very poor. The reader ignores what are, for example, the mean discharges of the stream at the outlet of the system, their seasonal variability, the overall variability of heads within the subsurface, what is the hydro-meteorological forcing, what are the boundary conditions, etc. Even though the main question is not to go into the detailed features of the forward problem, a few words for fixing the context would be welcome. The hydrological context could condition the applicability of the CRPS as an OF; most of inverse problems are case-study dependents.**

Reply:
We left out further descriptions of the model areas because we considered those things covered by the various publications on our national model, which the models presented in this manuscript are based on. But we certainly can understand, that further details will give a better context to most readers who are unfamiliar with Denmark, without having to refer to the references. We will add further information along the lines of the reviewer's suggestions in the revised manuscript.

5- **Line 175 -. It is stated that the hydrogeological model (the subsurface) encloses several "layers", which I think to be the representation of a geological stratification in the subsurface, with the consequence of generating vertical heterogeneity in the hydraulic parameters. A few lines later, (230 and followings) it is stated that only "six different geological units' hydraulic conductivities are sought, which would mean that within a unit (a "block" sub-system), the conductivity is uniform over the various geological layers. Why to distinguish these layers in the model geometry if they are similar in terms of hydraulic parameters?**

Reply:
Given the aggregation of some of the layers of the hydrogeological model into fewer (seven) computational layers in the model, there not only is a vertical heterogeneity, but also a horizontal heterogeneity within each computational layer.
That means, that despite only calibrating six different geological units, there is heterogeneity in across the seven computational layers. Furthermore, some of the geological units span several distinct layers, for example can we have a case where layer 1 (top) is a sand layer with conductivity K1, overlaying a clay layer (middle) with conductivity K2, overlaying itself a sand layer (bottom) with conductivity K1 again.

6- **Line 235 and followings. The so-called "benchmark" appears here as drawn out of the blue. When the reader expects that it will be discussed on the application of the CRPS, SSE, and SAE, to the actual case-studies, a "synthetic" problem is presented based on the various responses of the OFs to continuous Gaussian distributions of residuals. In addition, the "benchmark" is not well presented at all, and the reader is required to conjecture on the calculations performed in the benchmark.**

Reply:
The mentioned section (lines 235 ff. and Figure 3) is meant as an illustration of the general behavior of the different objective functions (CRPS, SSE, SAE) to different example distributions of residuals. The "benchmark" mentioned is just a normal distribution with mean 0 and standard deviation 1, and its meant as a reference to be compared to the other distributions, which are slightly deviating from it (either by introducing a bias or adding some outliers. Maybe the term "reference" is better than "benchmark"?

Furthermore, we understand that section 4.1 is not as such a result of our hydrological model calibration/the application of the CRPS as an objective function to our hydrological model. Hence, section 4.1 and the related Figure 3 could be moved to the end of section 2: After explaining the CRPS in general, its behavior to large values and its differences to the SSE, those effects can be shown based on a few example distributions (Figure 3).

> **7- Section 4.2. Even though, associated tables and figures report on the fact that CRPS outperforms the other OF, all the material is in fact a blind test as we ignore what are the structural errors in the models rendering high residuals. As told earlier, I would focus on a single test-case, I would consider a given inverse solution as a reference problem, and then I would add deliberate structural errors, for example on the delineation of the unit blocks, by overestimating or under estimating the aquifer thickness is some areas, by modifying the boundary conditions, by artificially generating a few zones of preferential infiltration, etc… Then by inverting these various configurations, a comparison of the performances of the various OFs could be carried out. In the present form of the study, the CRPS appears better as a matter of fact completely dependent of the overall settings of the forward problem, but applicability to other contexts is compromised, and a better response to structural errors (even though these errors probably exist in the tested forward problems) is not proven.**

Reply:
These are all valid concerns. We added a synthetic test (see reply to point 2). Furthermore, we want to point out again (as in our reply to the initial comment of the reviewer) that we are not only considering potential errors arising from structural errors, but also observational errors.

References

DHI: MIKE SHE, Volume 2: Reference Guide, , 2, 374 [online] Available from: https://manuals.mikepoweredbydhi.help/2019/Water_Resources/MIKE_SHE_Printed_V2.pdf, 2019.

Gneiting, T. and Raftery, A. E.: Strictly Proper Scoring Rules, Predictions, and Estimation., 2005.

Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, Weather Forecast., 15(5), 559–570, doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2, 2000.

Sonnenborg, T. O., Christensen, B. S. B., Nyegaard, P., Henriksen, H. J. and Refsgaard, J. C.: Transient modeling of regional groundwater flow using parameter estimates from steady-state automatic calibration, J. Hydrol., 273(1–4), 188–204, doi:10.1016/S0022-1694(02)00389-X, 2003.