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

Review comments on "A robust objective function for calibration of groundwater models in light of deficiencies of model structure and observations" by Schnieder et al., by TR Ginn.

I entirely agree with the insightful assessment provided by Prof. Neuman, and here add some further thoughts.

Already in the abstract the reader becomes worried that the central hypothesis, that structural errors or severely erroneous observations that lead to parameter (value) compensation can be ameliorated by using the new objective function (OF) norm, will not be tested. To test this hypothesis the underlying forward models require known structural errors or severely erroneous observations. Much of the discussion in the introduction, and specifically the stated Aim of the paper (line 95ff) focuses on the impact of structural errors. These are termed scale, structural or boundary condition errors by the authors but which by binary categorization – they are not observational – are in my view structural errors. This could be tested by using the CRPS norm on synthetic models with strucutural errors but this was evidently not done, in lieu of testing real field scale groundwater models. In the final statement the methodology is only "assumed" (line 351) to yield indications of structural error. I am skeptical of this because the groundwater flow equation is a diffusion equation and a structural error in one subdomain may in fact impact (downstream) heads in a relatively distant subdomain. This often happens when recharge is poorly calibrated and the simulated heads or their gradients far away, e.g., near a distant but sole outlet boundary, depart dramatically from measured values. Thus in my view the ability of the CRPS norm to address structural errors is not demonstrated.

Reply:
We want to thank Prof. Ginn for his comments on our manuscript. We understand many of his concerns, and will address them for the revised version of the manuscript.

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.

In general, we feel that it is relevant to point out (as we also did in our reply to Prof. Neuman) that our suggested use of the CRPS as an objective function originates from issues seen in practical applications (and less from theoretical considerations). In practical, real-world large-scale hydrological modelling, researchers commonly have to employ pragmatic solutions to tackle issues arising from the mentioned potential structural issues and observational errors. This even sometimes leads to the somewhat arbitrary exclusion of data because they in some way "do not seem to fit" to the model (or the conceptual understanding) – see lines 77ff. in the manuscript. This is where we hope that the CRPS (or the use of other, similar objective functions) can contribute to, for example, avoid arbitrary exclusion of data and, in general, provide us with

a more robust objective function (less prone to parameter compensation) in cases where we can almost be certain that parts of the model structure/data are flawed without being able to fully detect the specific issues. We will make this background more clear in the revised version of the manuscript.

Moving to Prof. Ginn's comments – we do agree that synthetic experiments are relevant to better show the claimed benefits of using the CRPS as an objective function. We already have performed some such experiments (without including them in the first submission), and are confident that we can show its benefits:

For a synthetic test, we took a certain realization (referring to a specific set of parameters) of the Storå model as a reference model. From a run of this reference model, we sampled synthetic observations at the exact same locations and times where real observations were available. White noise with 0.5m or 1.0m standard deviation was added to the synthetic observations.
Some observations are further perturbed. (those perturbations could account for larger, but undetected observation errors or a structural error in the model leading to a local inability of the model to reproduce observed groundwater heads) In this example, all observations within a bounding box, within the uppermost five layers of the model (quaternary layers), with an observed water level of at least 5.0m below the surface were perturbed by adding 4.0m to their observed value (to their synthetic truth). These observations are displayed in Figure 1.

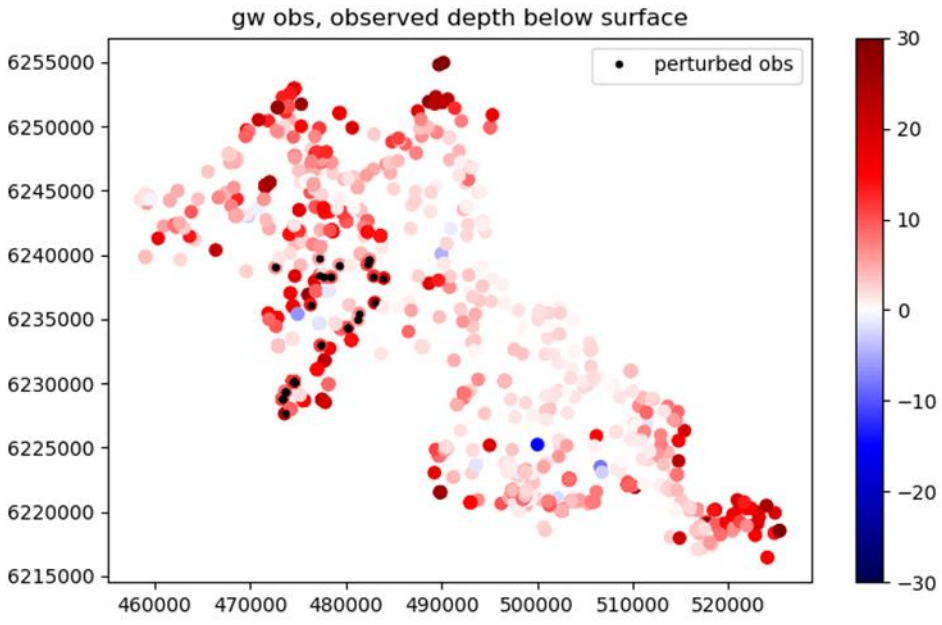

Figure 1. All groundwater head observations in the Storå catchment. The synthetic observations are sampled from a reference model run; the observations marked with a black dot are perturbed.

Then, starting from a different parameter set, the model is calibrated using the synthetic observations including the perturbations, with either the CRPS, MSE, MAE, or MRE as an objective function. The idea is, that the model calibrated using the CRPS as an objective function should be less affected by the few outliers in the synthetic observations. The synthetic test allows the conclusion that this actually is the case. For example, as can be seen in Figure 2, the parameter values resulting from an CRPS-based calibration are closer to the parameters of the synthetic truth than the parameters resulting from a SSE-based calibration, with the MAE and MRE-based calibrations lying in-between.

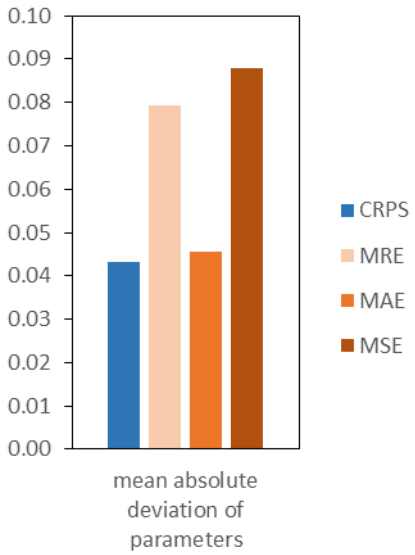

Figure 2. Mean absolute deviations from true parameter values (i.e. the reference model's ones), using the perturbed dataset and different objective functions. The mean values are weighted by the relative sensitivity across the six parameters.

Another indicator for our claim of the CRPS better coping with outliers than the SSE can be seen when comparing the model results of the reference model with each of the calibration results. This is done for the difference between average simulated groundwater head across the calibration period in the reference model compared to the models calibrated using the different objective functions, averaged across all computational layers in Figure 3. It can be seen clearly, that the calibrated model using the CRPS as an objective function is much closer to the reference model than the calibrated model using the SSE as an objective function.
The models using the MRE and MAE as an objective function perform similarly well as the CRPS. However, due to the higher sensitivity of the CRPS to biases, we prefer this objective function (also compare Figure 4 below).
We will add a discussion of the potential alternatives MRE and MAE to the revised version of the manuscript.

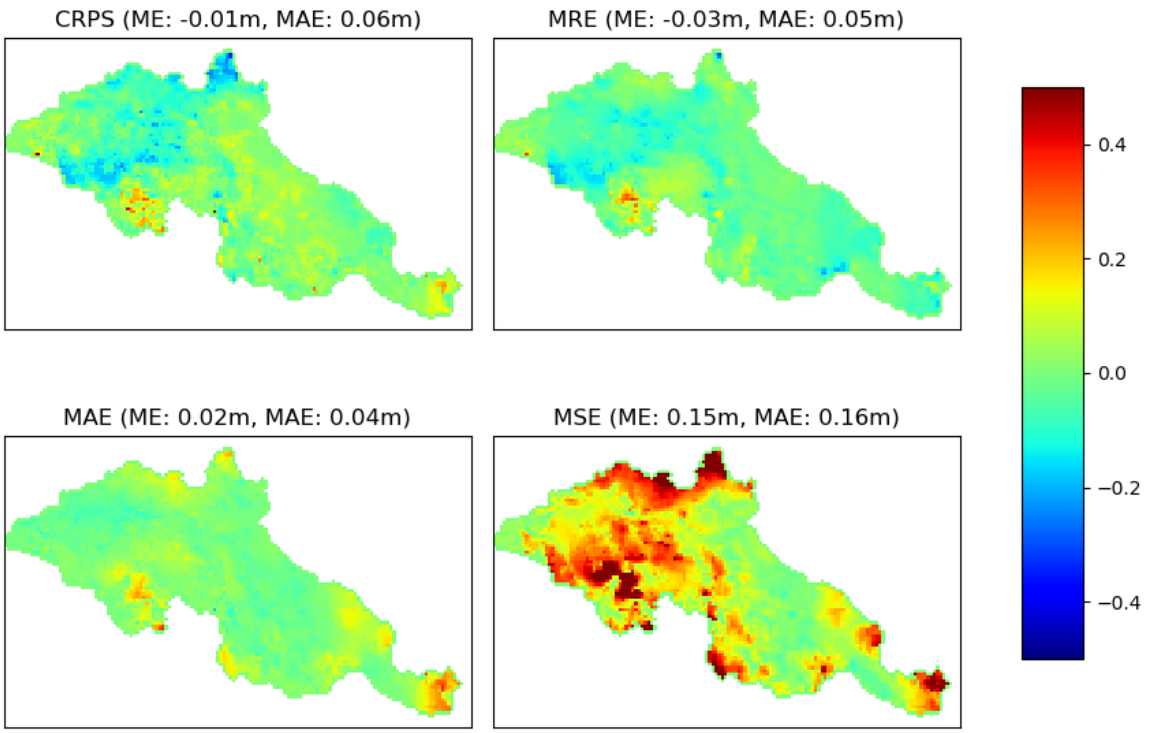

Figure 3. The deviation of the average simulated groundwater heads [m] of the models calibrated against the perturbed observations compared to the reference model as the mean across all model layers. The ME and MAE given in each title give the average deviations across all model grid cells.

It is true that, as Prof. Ginn points out, structural errors in a groundwater model in some cases only can be seen at detached/downstream locations. A complete detachment of the location of the structural error in a model (i.e. false representation of geology) and the respective response in simulated values (e.g. simulated groundwater heads showing bias), however, only occurs in some specific cases (e.g. where lateral flow is the most important flow path). In those cases, it will always be hard to point out *where* exactly there is an issue with model structure, but still there is an indication that there is an issue with model structure *somewhere*. In many cases, the effects of faults in model structure will be less detached from their impacts. We will discuss such limitations in the revised version of the manuscript.

The conceptual foundation for the method is probabilistic and as well noted by Prof. Neuman requires an ergodic argument. E.g., one instance on line 113, "…timestep." should in my opinion say "… timestep and spatial location." which creates conceptual problems in the subsequent extention to treating individual data locations in a single realization as sources for a probabilistic ensemble. Another example is (line 126)"… and the ECDF of residuals at every single observation point" which I do not understand, unless the authors mean "… and the ECDF of the collective set of residuals in the model." It may be possible to repackage the CRPS norm as a heuristic to avoid this often insurmountable challenge. Figure 1 seems to lead to a practical (heuristic) definition of dP as a vector of normalized cumulative cardinal number (or rank) of errors, where the cardinal counting is done from the largest +/- error, and x is actually the similarly ranked differences in errors again counting from the largest +- error. If I understand how it works the CRPS norm in this example is

$$dx * dP^2 = |\varepsilon_1 - \varepsilon_2|\left(\frac{1}{5}\right)^2 + |\varepsilon_2|\left(\frac{2}{5}\right)^2 + |\varepsilon_3|\left(\frac{3}{5}\right)^2 + |\varepsilon_4 - \varepsilon_3|\left(\frac{2}{5}\right)^2 |\varepsilon_5 - \varepsilon_4|\left(\frac{1}{5}\right)^2$$

where the first two terms are counting from the left and the last three are counting from the right because the first two are underestimates and the last three are overestimates. This clever device seems to weigh not errors but differences between errors that are adjacent in magnitude, with weight increasing with proximity of rank order to the observed value. It could be posed as a potentially promising alternative to the MSE (L2 norm) and MAE (L1 norm) and however should be raced also against a norm which magnifies the smaller errors (e.g., an L(1/2) norm).

Reply:
We do not assume that our calibration error is ergodic – the mean residual of a certain observation (equivalent to the mean residual of a certain ensemble member in "conventional" use of the CRPS) is not equal to the mean residual of all observations. However, we do not believe that it is a necessary criterion for using the CRPS in the way we intend, purely as a value for an objective function. There is limited literature discussing ergodicity as a strict requirement of applying the CRPS. However, we found one example arguing that the CRPS can also be used in combination with non-ergodic Schlather models (Yuen, 2015, p.14).

Yes, by our statement in line 126 we mean the ECDF of the collective set of residuals in the model. We will clarify this in the revised manuscript.

We agree with the reviewer's understanding of the CRPS, as shown for the example case with five values in Figure 1 and are glad to hear that he as well considers it a potentially promising alternative to the commonly used MSE or MAE.

We find the reviewer's idea with the L(1/2) norm (mean root error, MRE) interesting, and will include it in the synthetic example, see above. We still want to focus on the CRPS in this manuscript, given its higher sensitivity to a general bias, as can be seen in Figure 4 below. The RME and CRPS may be similarly sensitive to outliers (see the bottom right plot). However, as can be seen in the bottom left plot, the MRE is less sensitive to a systematic bias than the CRPS. In our models, where we are keen to achieve general water balance error, we Figure 4. Sensitivity of the CRPS in comparison to MSE, MAE, and RME towards bias and outliers. The top left plot shows the reference. The other plot titles report values of each metric relative to their respective reference values.prefer the behavior of the CRPS. A discussion of this will be added to the revised manuscript alongside Figure 4, which will be replacing Figure 3 in the manuscript.

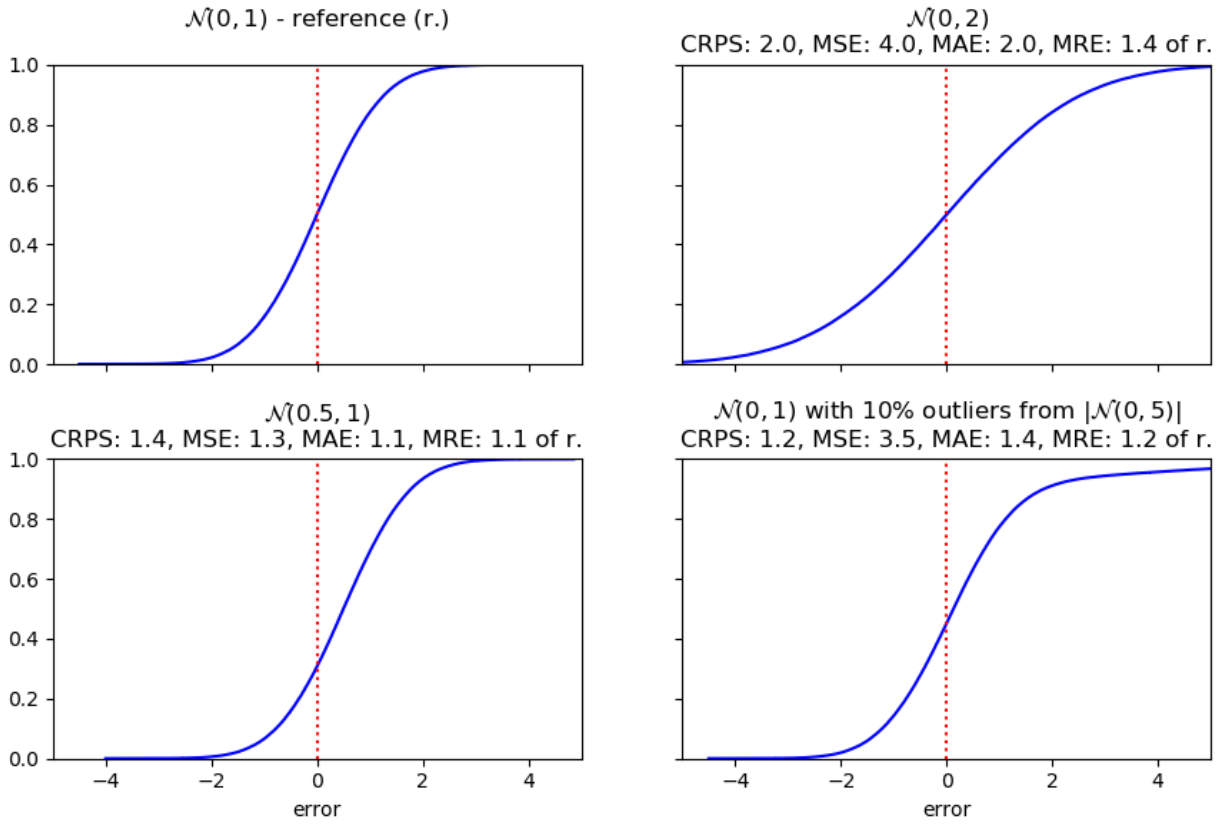

Figure 4. Sensitivity of the CRPS in comparison to MSE, MAE, and RME towards bias and outliers. The top left plot shows the reference. The other plot titles report values of each metric relative to their respective reference values.

A few lesser issues appear in the discussion of the nature of head data and of model failure modes. In lines 50-60 or thereabouts, and elsewhere, the focus is on the number of head data ("large enough set"). I believe that it is more often the distribution of the head data that is the more salient aspect that matters to inversion. Head data are often not uniformly distributed among or representative of the whole domain and/or the various important conductive units, especially in regions of strongly varying elevations, where most wells are in the valleys. The issue of model failure (lines 70-75; also 332) is attributed to the case when untrue parameter values (resulting from a structural problem) are obtained. It should be noted that these parameter values are already effective by construction, and their untrue values even if far removed from what a local pumping test would tell are only a problem if the model is incapable of simulating past or predicting future behavior, which is often found when the water changes direction, that is, when recharge or hydraulic boundary conditions change. At line 332 the term "nonoptimal" begs this very question. If the calibration minimizes the chosen OF norm then the parameter values are indeed optimal, at least to the mathematical inverse problem.

Reply:
Yes, we agree with the reviewer that often, the available head measurements are not distributed evenly across the model domain or the different geologic units. In Danish landscapes, however, with their rather gentle topography, at least there is no large imbalance in the distribution of head observations between valleys and ridges.
It is correctly noted, that we use the term "non-optimal" a bit loosely – we will clarify this in the revised version of the manuscript, making clear that what we want to avoid is parameter compensation.

Furthermore, we want to add that parameter compensation is not limited to model structural issues, but can also occur to compensate for observational errors.

References:

Yuen, R. A.: Topics on estimation, prediction and bounding risk for multivariate extremes, The University of Michigan. [online] Available from: https://deepblue.lib.umich.edu/bitstream/handle/2027.42/111408/bobyuen_1.pdf, 2015.