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

Comments by Shlomo P. Neuman on
A robust objective function for calibration of groundwater models in light of deficiencies of model structure and observations
by Raphael Schneider, Hans Jørgen Henriksen, and Simon Stisen

The authors propose using Continuous Ranked Probability Score (CRSP) as a criterion for calibrating groundwater models against hydraulic head data. They reason that large residual calibration errors, which often result in part from structural model errors, would dominate CRSP calibration results to a lesser degree than they do when one uses standard criteria such as mean square error (MSE)or mean absolute error(MAE); CSRP would assign greater weight to the majority of smaller residuals than to a few larger residuals at the edges of their cumulative distribution .Whereas CSRP is designed to work with ensembles of predictions, the authors suggest applying it to a single realization of calibration errors across a model space-time horizon. To test their idea, the authors apply CSRP to two regional scale coupled surface-groundwater models to conclude that their proposed criterion results in lesser calibration bias than do MSE or MAE.

Reply:
We thank Prof. Neuman for his insightful and critical 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.

In general, we feel it is relevant to stress that our suggested approach of using an alternative objective function to the commonly used MSE originates from dealing with issues arising in the practical application of large-scale hydrological models, and less from theoretical considerations on inverse problems etc. Such practical issues, as (i) the mentioned unavoidable structural errors in large-scale models due to a lack of detailed knowledge of the geology, process simplifications, or matters of scale, (ii) observational datasets of unknown and varying quality as well as scarce and uneven distribution in time and space, and (iii) the inability to quantify and disentangle different sources of uncertainty in many real-world problems, lead to often unsound ways of dealing with data that “do not seem to fit”, i.e. a somewhat arbitrary outlier filtering. We want to provide the practitioner an objective function that better allows to also ingest some “flawed” observations – when it is not possible to detect which observations actually are flawed. Moreover, such an objective function will better allow to identify areas where our model or data is flawed, as we mention in the manuscript in lines 280ff., and then subsequently perform further investigations into the cause of the discrepancies. We feel that, in general, the sensitivity of the squared error-based objectives in inversions is acknowledged (think e.g. also of more advanced techniques of handling this issue, such as iteratively reweighted regression), but too often – at least in practical applications – is ignored, partly due to the above outlined challenges of unknown data quality and modelling requiring stark simplifications of reality. We will make this background more clear in the revised version of the manuscript.

Though, of course we agree that also applied modelling approaches need to be examined based on a sound theoretical foundation. We are convinced that the manuscript will benefit greatly by addressing these concerns, amongst others by adding synthetic examples to further showcase the claimed benefits of the CRPS-based objective function.

I find the idea of using CRSP as a calibration criterion interesting but consider the authors' attempt to demonstrate its utility unconvincing. My reasons are as follows:

1. **Applying the probabilistic CRSP criterion to a single realization of calibration error requires an assumption of ergodicity. There is no discussion of this potential restriction in the manuscript.**

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).

2. **Groundwater flow models differ fundamentally from most surface water models in that parameters entering the former (hydraulic conductivity or transmissivity, specific storage or drainable porosity) tend to have reasonably well-defined physical meanings and can often be estimated, independently of the calibrated model, through methods such as pumping tests and geostatistical interpolation. This makes it possible, and often necessary, to regularize the model calibration process with the aid of parameter plausibility criteria based either on such independent prior parameter estimates or on functional criteria such as smoothness. One purpose of such regularization criteria is to ensure that large calibration errors do not dominate the parameter estimation process. Would CRSP be still necessary, and/or useful, in this context? The manuscript does not address this question.**

Reply:
Yes, we still consider the CRPS relevant here. For example:

a) Distributed groundwater models' parameters have physical meaning, but because of issues with model grid scale, heterogeneity, simplification of processes (e.g. of the unsaturated zone, artificial drain, etc), errors in the hydrogeological model used to outline conductivity zones, etc. parameter values usually cannot be estimated directly. Despite not using any regularization or parameter bounds, our models' parameters fall into reasonable/expectable ranges after the calibration (e.g. horizontal conductivities in sand layers range from ~$1*10^{-3}$ m/s to ~$1*10^{-4}$ m/s, and from ~$3*10^{-6}$ m/s to ~$6*10^{-8}$ m/s in clay layers)
b) We assume Prof. Neuman is referring mainly to highly parameterized approaches (inversion of geophysical data or highly parameterized hydrological models solved e.g. by pilot points), as he mentions regularization and smoothness (that is, smoothness of parameter fields?). We are using a unit-based approach, which does not require the use of regularization or similar. Furthermore, despite not using tight parameter bounds limiting parameters to "reasonable" values, the inversion process in the vast majority of cases ends up with "reasonable" parameter values – see point a). In pilot-point approaches, with regularization of the parameter field, one still could assume that the

CRPS is beneficial, as it still reduces the impact of single (wrong) observations on the parameter field (even if only local). However, testing this is outside the scope of this manuscript.

We will add some of these clarifications to the introduction and discussion part of the revised manuscript.

3. **The two case studies fail to provide information about the reliability of parameters estimated using either CSRP, MSE or MAE. To validly compare these three criteria, one would need to test them on synthetic systems having known structures, parameters and forcing terms that are corrupted by known random and/or systematic errors of realistic kinds and magnitudes. One would further need to explore CSRP in the context of regularization criteria such as those commonly used in groundwater model calibration. Only then would it make sense to demonstrate the utility of CSRP on partially defined field problems such as those in the two case studies described.**

Reply:
Concerning the reliability of the estimated parameter values we want to refer to our reply a) to point 2: With the given model structure and scale, all parameter values are effective parameters, and cannot be directly measured in the field or related to literature values. However, we still have expected ranges for parameter values. In general, the estimated parameter values in the two presented real-world models fall into expected/plausible ranges, though without allowing for the conclusion that the CRPS-based objective function yields narrower confidence intervals or more “reasonable” parameter values.

Concerning synthetic tests: When developing and testing the idea of using the CRPS as an objective function, we also tested it in synthetic environments. We fully understand Prof. Neuman’s concern; therefore, we will add some of the results of the synthetic tests to the revised manuscript, as outlined in the following paragraphs. Similar changes were also requested by reviewer 2, Prof. Ginn, who also was interested in a comparison of the CRPS against the mean root error (MRE) and mean absolute error (MAE) – hence, those two were also included in the synthetic tests.
(Our synthetic tests, however, are based on the same type of practically applied large-scale hydrological models as presented in the manuscript. That is, our models are parameterized based on (few) geological units, where the hydrological units are distributed based on conceptual hydrogeological models. Such model setups are common in research and practical applications. Therefore, a test of regularization is beyond the scope of our manuscript. We will add it as part of limitations/outlook to the revised manuscript)
For a synthetic test, we took a certain realization (referring to a specific set of parameters) of the Storå model as the 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.

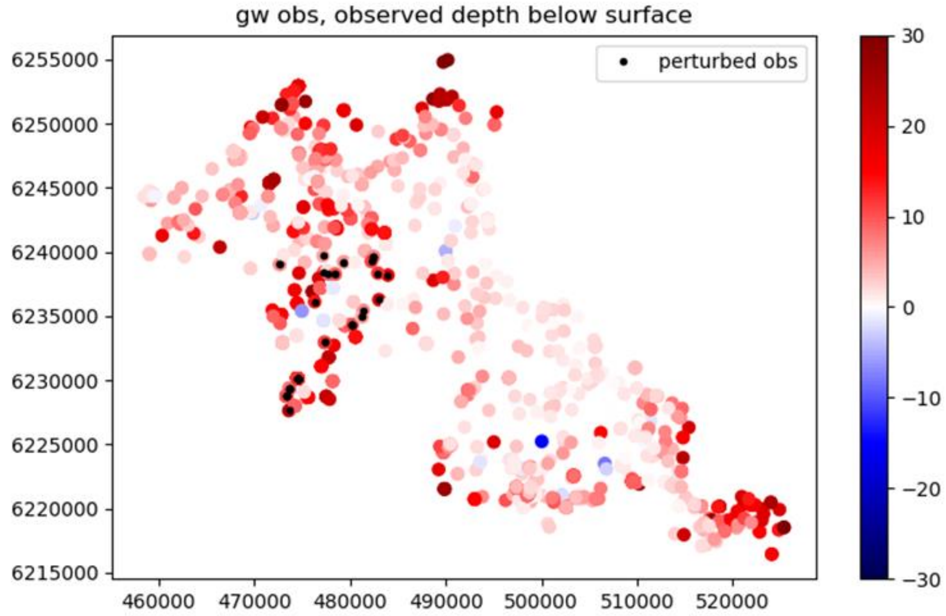

Figure 1. All groundwater head observations in the Storå catchment. The synthetic observations are sampled from a reference model run; the observations marked with a black dot are perturbed.

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.

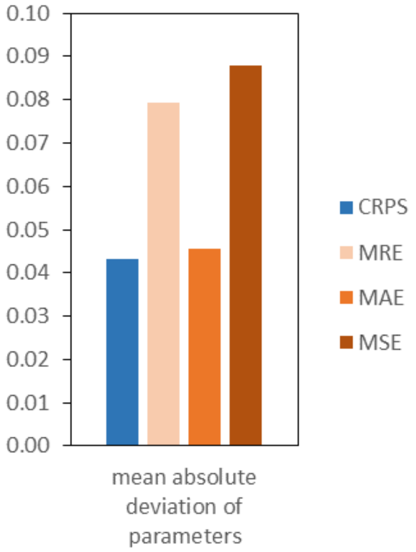

Figure 2. Mean absolute deviations from true parameter values (i.e. the reference model's ones), using the perturbed dataset and different objective functions. The mean values are weighted by the relative sensitivity across the six parameters.

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 **Error! Reference source not found.** in the reply to reviewer 2).

We will add a discussion of the potential alternatives MRE and MAE to the revised version of the manuscript.

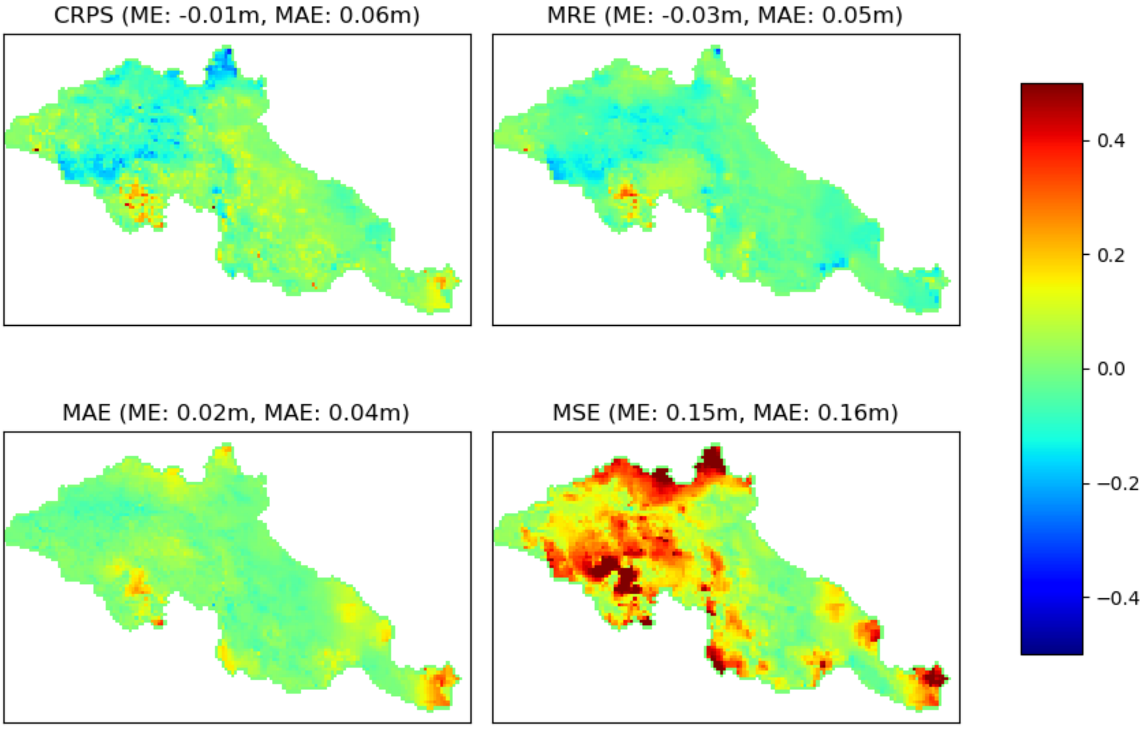
CRPS (ME: -0.01m, MAE: 0.06m)
MRE (ME: -0.03m, MAE: 0.05m)
MAE (ME: 0.02m, MAE: 0.04m)
MSE (ME: 0.15m, MAE: 0.16m)
0.4
0.2
0.0
−0.2
−0.4

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.

References

Sanchez-Vila, X., Guadagnini, A. and Carrera, J.: Representative hydraulic conductivities in saturated grqundwater flow, Rev. Geophys., 44(3), 1–46, doi:10.1029/2005RG000169, 2006.

Wang, Y. L., Yeh, T. C. J., Wen, J. C., Gao, X., Zhang, Z. and Huang, S. Y.: Resolution and Ergodicity Issues of River Stage Tomography With Different Excitations, Water Resour. Res., 55(6), 4974–4993, doi:10.1029/2018WR023204, 2019.

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.