The CRPS – used as a robust objective function for groundwater model calibration in light of observation and model structural uncertainty

Raphael Schneider*, Hans Jørgen Henriksen, Simon Stisen

Geologic Survey of Denmark and Greenland (GEUS), Department of Hydrology

*rs@geus.dk

EGU2020-6966

Let’s talk about objective functions…

Hydrologic models commonly require parameter estimation through optimization, which involves fitting simulated results to some observations

The **objective functions** used here are usually **squared-error-based.**

Squared-error-based performance criteria, however, are (overly?) **sensitive to extreme values/outliers**

forecasts are quite poor. However, a significant problem with these measures is that they heavily penalize large errors for individual forecasts. In this case, the errors for just two forecasts (1962 and 1993) account for almost 90% of the MSE. These errors are caused by a similar situation. Both years experienced very wet summers, followed by a drier than average September; with high moisture conditions at the start of September, most

Bradley et al., 2004

CONCLUSIONS

This paper has illustrated with an example that the value of the persistence index – a quadratic criterion commonly used to assess the performance of forecasting hydrological models – computed over a continuous data series is mostly influenced by a very small number of time steps characterised by a high runoff variation.

Berthet et al., 2010

tively. Because of the squared differences, however, E is overly sensitive to extreme values, as is R^2.

Legates and McCabe, 1999

MSE, which calculates the arithmetic mean of squared errors, was used in the first M-Competition [12]. However, its use was widely criticized later as inappropriate [6, 27]. MSE is more vulnerable to outliers since it gives extra weight to large errors. Also, the squared errors are on different scale from the original data. Thus, RMSE, which is the squre root of MSE, is often preferred to MSE as it is on the same scale as the data. However, RMSE is also sensitive to forecasting outliers [28].

Chen et al., 2017

The sensitivity of the RMSE to outliers is the most common concern with the use of this metric. In fact, the existence of outliers and their probability of occurrence is well described by the normal distribution underlying the use of the RMSE. Table 1 shows that with enough samples ($n \geq 100$), including those outliers, one can closely re-construct the error distribution. In practice, it might be justifiable to throw out the outliers that are several orders larger than the other samples when calculating the RMSE, especially if the number of samples is limited. If the model biases are severe, one may also need to remove the systematic errors before calculating the RMSEs.

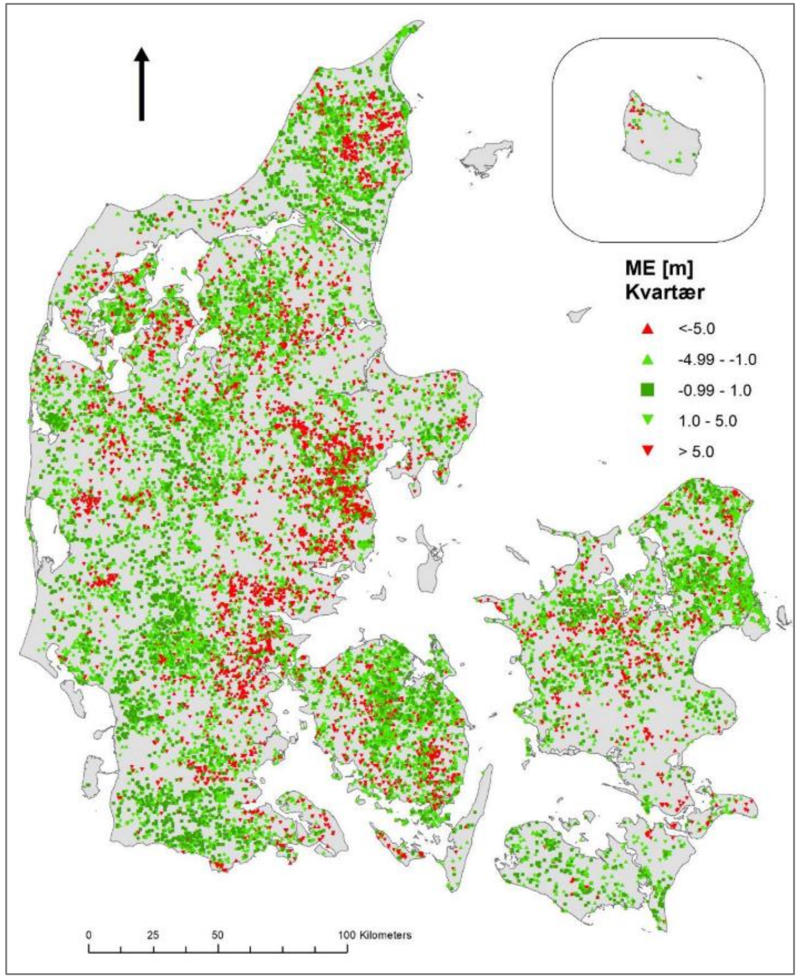
Chai and Draxler, 2014

…and about deficiencies in our observations and models

This sensitivity of squared-error-based performance criteria to extreme values is problematic, especially in practical applications of large-scale models:

1. **Model deficiencies**
   Inevitable model structural errors, e.g. in the hydrogeologic model
2. **Data deficiencies**
   Can outliers in observations be identified reliably? Can we assess observational uncertainty?
   No, often when dealing with large datasets of e.g. groundwater heads, data origin and quality is unknow

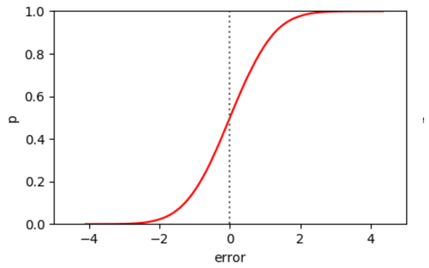
→need for a **robust** objective function



Working with large-scale models: the DK-model, covering all of Denmark with the available head observations, from the public national well database (Højberg et al., 2013; Stisen et al., 2019)

Using the CRPS as *robust* objective function

reference: N(0,1)



Originally, the Continuous Ranked Probability Score CRPS is an evaluation tool for probabilistic forecasts (Gneiting et al., 2005)

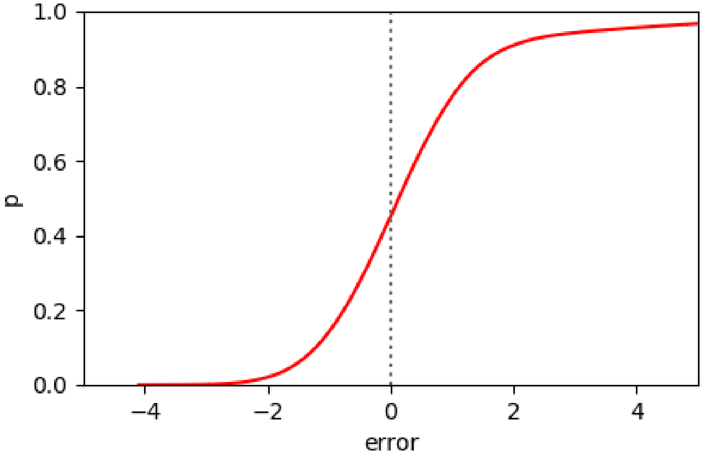
We suggest using it as an objective function, instead of the commonly used squared-error-based metrics, because of its

- insensitivity to outliers/extreme values
- sensitivity to bias

Both are desirable properties e.g. in our applications, large-scale groundwater models

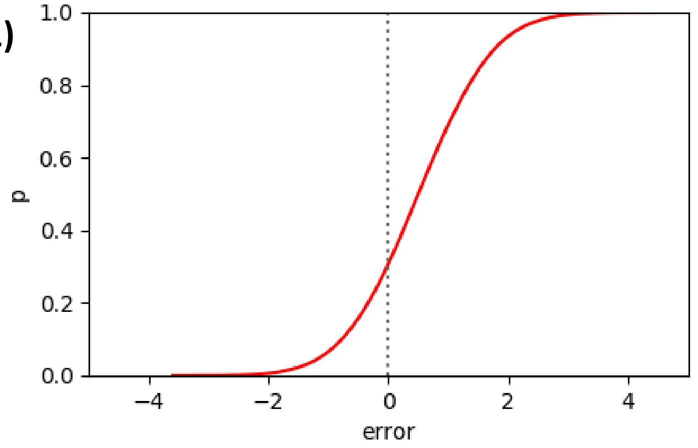
adding outliers: N(0,1) with 10% from |N(0,5)|

→ CRPS: 1.2*ref
MSE: 3.5*ref
MAE: 1.4*ref
MRE: 1.2*ref



introducing bias: N(0.5,1)

→ CRPS: 1.4*ref
MSE: 1.3*ref
MAE: 1.1*ref
MRE: 1.1*ref

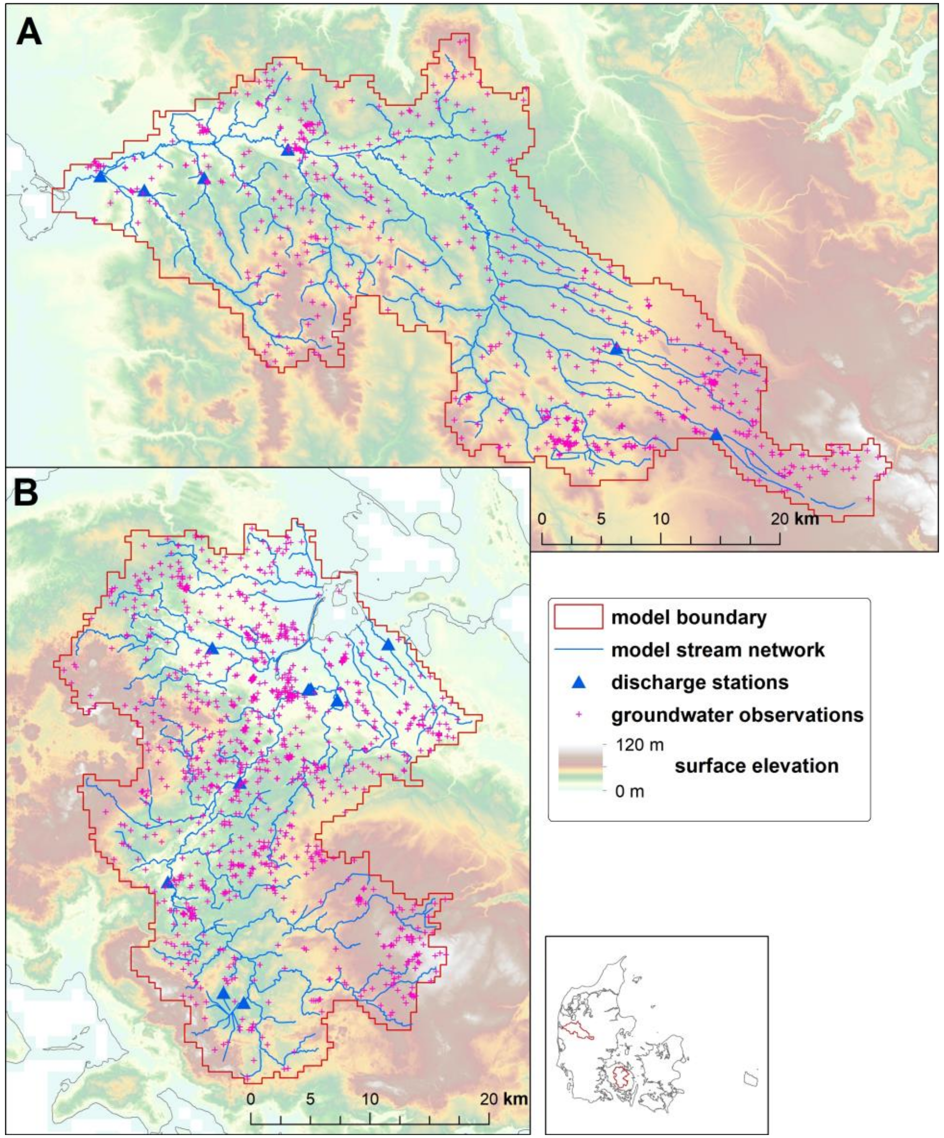


The CRPS applied…

…to the calibration of two Danish regional-scale distributed groundwater-surface water models, set up in MIKE SHE (details in Schneider et al., 2020, HESS Discussions*)

A: Storå catchment, B: Odense catchment

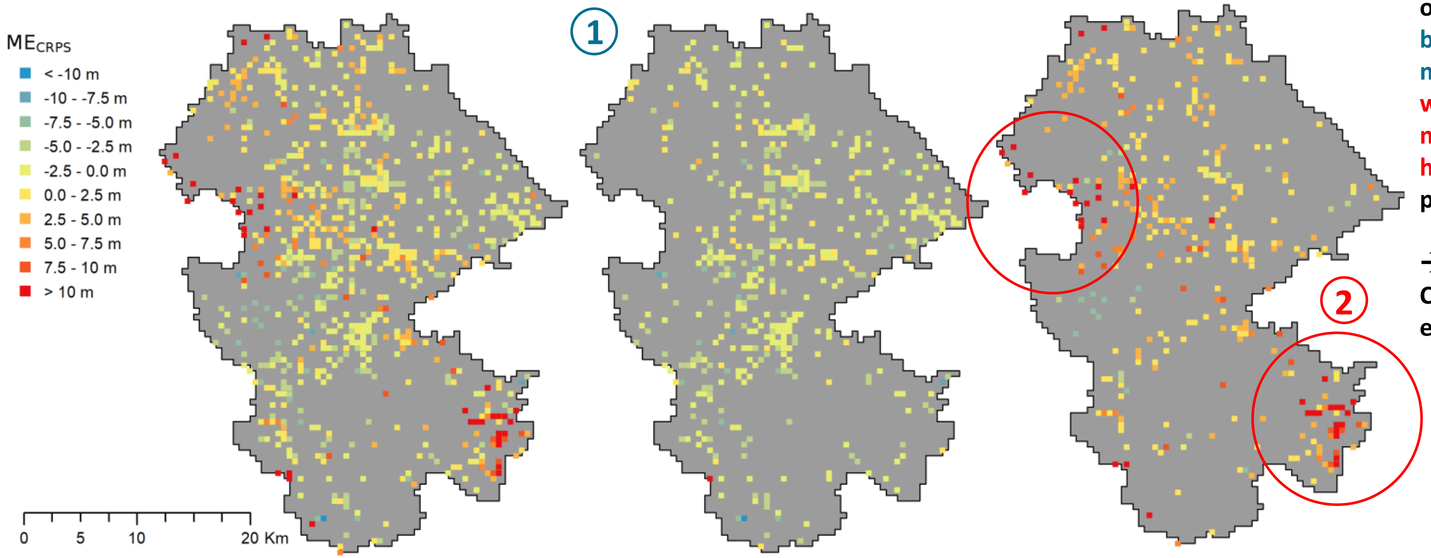
- based on the national water resource model for Denmark (DK-model, http://dk.vandmodel.dk/in-english/)
- area each ~1,000 km^2, 500 m model resolution
- **unit-based** parameterization of geology
- transient model, **daily timesteps** from 2000 to 2008
- calibrated against
  - observed **groundwater heads**
    A: 5,218 obs. in 890 wells; B: 44,723 obs. in 1820 wells
  - observed daily **stream discharge**
    A: 6 stations; B: 9 stations



* https://www.hydrol-earth-syst-sci-discuss.net/hess-2019-685/

The CRPS applied…

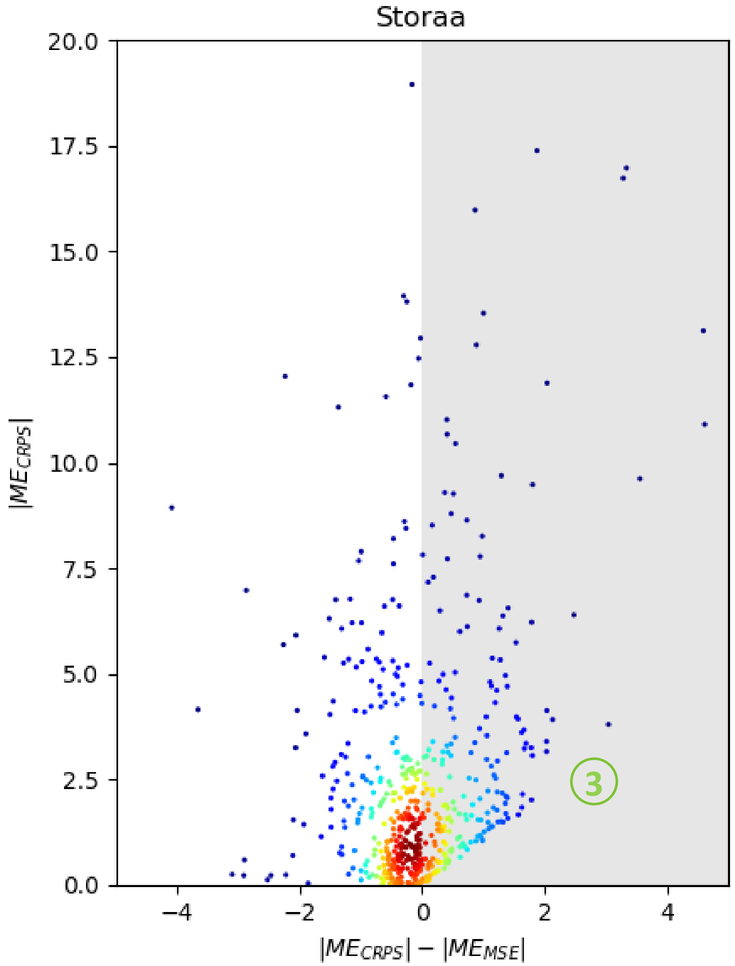
…to the calibration of two Danish regional-scale distributed groundwater-surface water models, set up in MIKE SHE (details in Schneider et al., 2020, HESS Discussions*)



Residual per grid cell for Odense. **Left**: All ME per grid from the CRPS calibration. **Middle**: only showing ME per grid, where the CRPS performs better than the MSE calibration. **Right**: only showing ME per grid, where the CRPS calibration performs worse than the MSE calibration. Again, it can be seen that the **majority of the grid cells performs better after the CRPS calibration.** **Among the worse performing grids, some patterns become obvious → issues with model structure, boundary conditions or similar?** 1 2

With using the CRPS instead of the MSE, we achieve a better model fit for the majority of observations, while keeping some larger misfits more obvious that can hint to data or model issues, potentially allowing a revisit

→the benefits of the robust CRPS shown for a synthetic example: see slide 10



y-axis: Absolute residuals per model grid cell after calibration against CRPS. x-axis: Difference in absolute residuals per model grid cell between calibration against CRPS and calibration against MSE. The majority of the grid cells lie in the white half of the plot, i.e. they have a improved fit after the CRPS calibration over the MSE calibration. Those **grid cells that have worse fit after CRPS calibration (grey half), have predominantly larger residuals.** 3

* https://www.hydrol-earth-syst-sci-discuss.net/hess-2019-685/

Conclusions

In general, the focus on **squared-error**-based objective functions can be **problematic**, especially in practical applications of large-scale models **with inevitable, hard to quantify deficiencies in model and observation data** – even though we generally are aware of the significance of the calibration targets and data in our modelling work.

We want to

1) highlight the issues with squared-error based metrics in these contexts
2) suggest CRPS as an alternative, given its relative insensitivity to outliers, as it
    - → reduces need for outlier filtering, which potentially is a subjective task
    - → reduces risk of parameter compensation for data and model deficiencies
    - → allows better identification of model structural issues or systematic errors in observations after calibration
    - → synthetic experiments could show that CRPS calibration results (i) in parameter values closer to truth, and (ii) less disturbed model results

Background slides

The CRPS – used as a robust objective function for groundwater model calibration in light of observation and model structural uncertainty

Raphael Schneider*, Hans Jørgen Henriksen, Simon Stisen
Geologic Survey of Denmark and Greenland (GEUS), Department of Hydrology
*rs@geus.dk

EGU2020-6966

GEUS

04-05-2020

Motivation

As mentioned above, in many cases (especially larger scale practical applications of hydrological models), we are dealing with inevitable deficiencies in data and models that are hard to quantify.

This is where we want to suggest the use of a **robust** objective function, which does not show the sensitivity to large values of squared-error-based metrics.

Our hypothesis is that a **robust** objective function is a practical way to

- reduce parameter compensations for model structural deficiencies
- allow better post-calibration identification of model areas with model structure or data deficiencies
- allow better ingestion of (large) datasets with unknown and varying quality
- reduce the need for sometimes subjective (false?) outlier filtering

> Different definitions are possible for outliers. In our approach, we postulated that an outlier is a catchment that behaves so differently from its geographical neighbours (even after accounting for its own physical characteristics through equation (1)) that it affects the efficiency of the regionalization process. However, the same catchment would not necessarily have been considered an outlier if located in a different neighbourhood (contrary to

Boldetti et al., 2010

> available for this site. By reviewing the extensive database developed by *Fenelon* [2005], we were able to filter out unreliable measurements and only use head measurements which are trustworthy and thus unaffected by wellbore and/or drilling affects. A total of 583 head measurements collected at 60 different wells were selected for inclusion in this

Keating et al., 2010

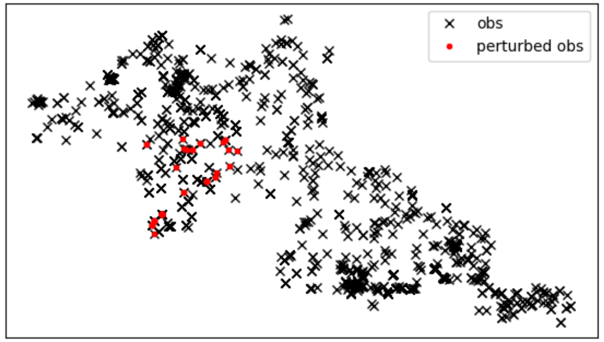
> drastic human interference or error. It should be highlighted that hydrographs classified as "irregular" by visual inspection should not be regarded as unsuitable; on the contrary, they provide a wealth of information on system behaviour. Yet, in the context of the present paper, we disregarded irregular time series as each record receives its own group.

Haaf and Barthel, 2018

The CRPS applied in a synthetic example

Synthetic example, based on the Storå model, with same spatio-temporal distribution as real observations

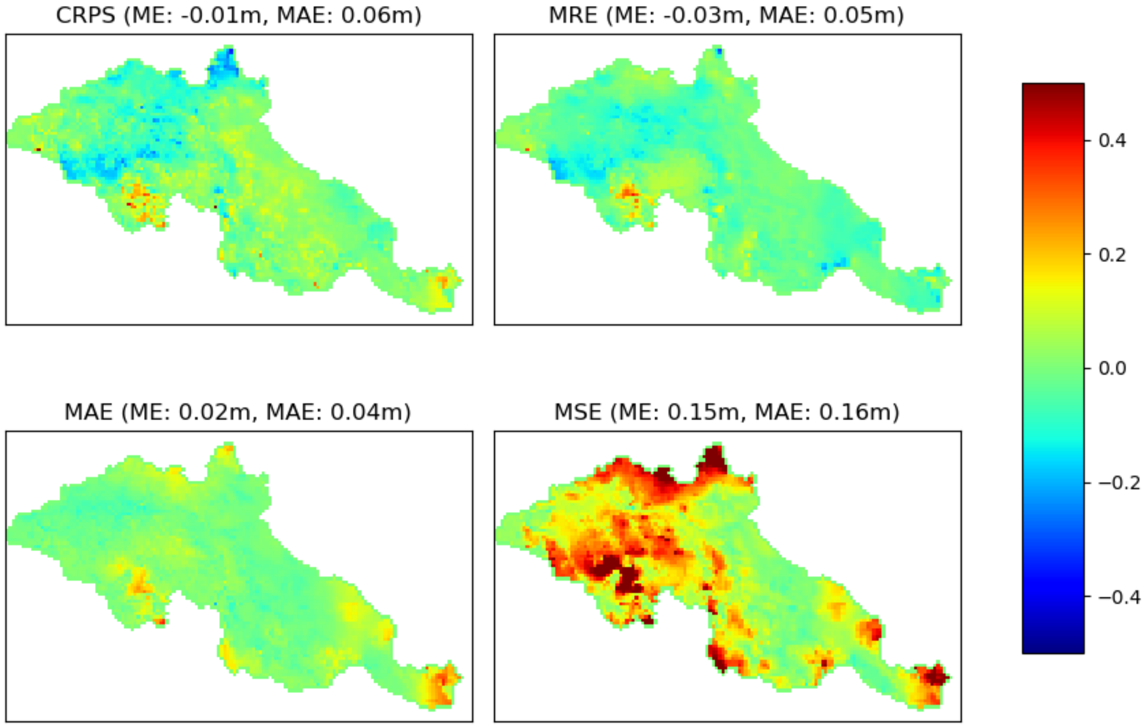
- synthetic observations taken from reference run
- some synthetic observations perturbed
- model calibrated against perturbed data, using different objective functions



→CRPS clearly less affected by the perturbed observations than conventional MSE

→CRPS performs similar to MAE and MRE; however, we prefer CRPS due to its higher sensitivity to bias

(similar behavior when looking at how close calibrated parameters are to true parameters)



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.

GEUS

Abstract

The Continuous Ranked Probability Score (CRPS) is a popular evaluation tool for probabilistic forecasts. We suggest using it, outside its original scope, as an objective function in the calibration of large-scale groundwater models, due to its robustness to large residuals in the calibration data.

Groundwater models commonly require their parameters to be estimated in an optimization where some objective function measuring the model's performance is to be minimized. Many performance metrics are squared error-based, which are known to be sensitive to large values or outliers. Consequently, an optimization algorithm using squared error-based metrics will focus on reducing the very largest residuals of the model. In many cases, for example when working with large-scale groundwater models in combination with calibration data from large datasets of groundwater heads with varying and unknown quality, there are two issues with that focus on the largest residuals: Such outliers are often i) related to observational uncertainty or ii) model structural uncertainty and model scale. Hence, fitting groundwater models to such deficiencies can be undesired, and calibration often results in parameter compensation for such deficiencies.

Therefore, we suggest the use of a CRPS-based objective function that is less sensitive to (the few) large residuals, and instead is more sensitive to fitting the majority of observations with least bias. We apply the novel CRPS-based objective function to the calibration of large-scale coupled surface-groundwater models and compare to conventional squared error-based objective functions. These calibration tests show that the CRPS-based objective function successfully limits the influence of the largest residuals and reduces overall bias. Moreover, it allows for better identification of areas where the model fails to simulate groundwater heads appropriately (e.g. due to model structural errors), that is, where model structure should be investigated.

Many real-world large-scale hydrological models face similar optimizations problems related to uncertain model structures and large, uncertain calibration datasets where observation uncertainty is hard to quantify. The CRPS-based objective function is an attempt to practically address the shortcomings of squared error minimization in model optimization, and is expected to also be of relevance outside our context of groundwater models.

References

Berthet, L., Andréassian, V., Perrin, C. and Loumagne, C.: How significant are quadratic criteria? Part 2. On the relative contribution of large flood events to the value of a quadratic criterion, Hydrol. Sci. J., 55(6), 1063–1073, doi:10.1080/02626667.2010.505891, 2010.

Boldetti, G., Riffard, M., Andréassian, V. and Oudin, L.: Data-set cleansing practices and hydrological regionalization: is there any valuable information among outliers?, Hydrol. Sci. J., 55(6), 941–951, doi:10.1080/02626667.2010.505171, 2010.

Bradley, A. A., Schwartz, S. S. and Hashino, T.: Distributions-Oriented Verification of Ensemble Streamflow Predictions, J. Hydrometeorol., 5(3), 532–545, doi:10.1175/1525-7541(2004)005<0532:DVOESP>2.0.CO;2, 2004.

Chen, C., Twycross, J. and Garibaldi, J. M.: A new accuracy measure based on bounded relative error for time series forecasting, PLoS One, 12(3), 1–23, doi:10.1371/journal.pone.0174202, 2017.

Chai, T. and Draxler, R. R.: Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature, Geosci. Model Dev., 7(3), 1247–1250, doi:10.5194/gmd-7-1247-2014, 2014.

Gneiting, T., Raftery, A. E., Westveld, A. H. and Goldman, T.: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation, Mon. Weather Rev., 133, 1098–1118, doi:10.1175/MWR2904.1, 2005.

Haaf, E. and Barthel, R.: An inter-comparison of similarity-based methods for organisation and classification of groundwater hydrographs, J. Hydrol., 559, 222–237, doi:10.1016/j.jhydrol.2018.02.035, 2018.

Højberg, A. L., Troldborg, L., Stisen, S., Christensen, B. B. S. and Henriksen, H. J.: Stakeholder driven update and improvement of a national water resources model, Environ. Model. Softw., 40, 202–213, doi:10.1016/j.envsoft.2012.09.010, 2013.

Keating, E. H., Doherty, J., Vrugt, J. A. and Kang, Q.: Optimization and uncertainty assessment of strongly nonlinear groundwater models with high parameter dimensionality, Water Resour. Res., 46(10), 2009WR008584, doi:10.1029/2009WR008584, 2010.

Legates, D. R. and McCabe, G. J.: Evaluating the use of “goodness-of-fit” Measures in hydrologic and hydroclimatic model validation, Water Resour. Res., 35(1), 233–241, doi:10.1029/1998WR900018, 1999.

Schneider, R., Henriksen, H. J., and Stisen, S.: A robust objective function for calibration of groundwater models in light of deficiencies of model structure and observations, Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-685, in review, 2020.

Stisen, S., Ondracek, M., Troldborg, L., Schneider, R. J. M. and van Til, M. J.: National Vandressource Model - Modelopstilling og kalibrering af DK-model 2019., 2019.