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.

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

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

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).

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.