Response to second-round comments

In this document we reproduce the editor and reviewer comments in gray boxes, with our responses following, explaining our reasoning. Explanation of how specifically we have altered the MS in response to the comments appears in **bold**.

Editor's remarks

> Both reviewers provide a positive assessment of the revised version and require a minor revision. The revised paper will be sent to the reviewers for a final check.

We have attempted to address the remaining reviewer remarks.

Reviewer 1: Roseanna Neupauer

> [V]ery minor editorial changes.

We are gratified that we seem to have mostly satisfied this reviewer. While specific editorial changes were not specified, we have passed over the document carefully and **corrected a few minor defects throughout the text**.

Reviewer 2: Philippe Ackerer

> The paper addresses an interesting issue, not new but not solved until now. It is a paper dedicated to a new methodology using a parameterization based on Karhunen-Loève expansions. The methodology is tested with a synthetic 'model aquifer' under simplified conditions. Using synthetic data set avoids wrong interpretation of the methodology performance due to uncertainty and measurement errors that are unavoidable when using real test cases. Of course, it questioned the feasibility of the method, but this is another issue.
>
> The paper is well written, the methodology well described and the results convincingly discussed.

We appreciate the overall positive assessment.

> Some comments that can be discussed:
>
> - Since the exact heads and hydraulic conductivities are known over the all domain, it may be interesting to analyze the interactions between head and hydraulic conductivities (using a cross variogram for example). This information could be included in the discussion about the spacing/density of the measurements.

We agree this is an interesting interaction to consider. We are presently working on a follow-up study in which we examine the added value of spatial correlation information, including the sort of cross-covariance data used in cokriging approaches. We hope that it will not be a problem to save this particular analysis for that manuscript.

- I was surprised by the use of measured velocities. They cannot be measured, to my knowledge. In wells, flow rates are measured and the width of the captured zone is required to estimate the velocity. In the field, at least 3 piezometers are necessary to estimate the head gradient, but hydraulic conductivity is needed to compute an average velocity.

We are careful not to assume knowledge of the direction of the velocity vector in our cost function (3), only potentially the *magnitude* of the Darcy velocity, which is the quantity represented by $\tilde{q}$. This may be obtained passively at a monitoring well by use of, e.g., a point dilution test. We mention this on line 35 of the MS.

- The head sampling is based on a regular grid. I do not see the interest of the regular grid for a heterogeneous aquifer.

The use of a regular grid is naturally an idealization. Our working hypothesis is that effective measurement density, not detailed well configuration, controls the threshold of feature identification. On this view, use of a regular geometry is harmless, and it simplifies analysis because there is an obvious single measurement spacing scale, rather than a distribution. We could also have run the calibration trials by randomly distributing fixed numbers of measurement locations throughout the domain, and do not think it would appreciably change our average results. It would, however, have increased complexity of the analysis, as there would generally be some calibration trials featuring tight, redundant well clusters that poorly sample the domain, for which the effective number of distinct sampling locations is significantly less than the notional number of wells. **We mention our reasoning in section 3.4 of the revised manuscript.**

- Only 50 random fields are used. This is very low compared to the variance of the LnK and no reliable statistics can be drawn with such a number. What is the interest of these 50? Why not testing random fields with other properties (smaller LnK variance, lower integral scales, …).

We stress that we used *different sets* of 50 randomly-generated fields *for each* of the eight distinct calibration ensembles that we present in Figures 3 and 4, as well as the regularization-by-truncation trial summarized in Table 1: 450 in total. The same patterns linking measurement-scale-normalized feature scale, T, to average degree of improvement in identification at that scale due to calibration, r, were apparent in the various calibration ensembles for different Ψ shown in these figures. This suggests that the patterns observed are not artifacts of a specific set of fields that were generated. While more simulations would naturally increase confidence in our results by reducing noise (i.e., vertical scattering in these figures), we do not believe they would alter the average trends we have identified. **In section 3.4 of the revised manuscript, we make this point explicitly, and also clarify that *different* fields were used in each calibration ensemble. We also correct the text to state that, for each ensemble presented, *exactly* 50 fields were used.**

We do agree that different heterogeneity statistics, and potentially even different correlation structures, would be interesting to consider with this same approach in follow-on research. However, there is a limit to how many parameters can feasibly be varied in a computationally-intensive study such as this one (which consumed thousands of compute hours, as is).

- Optimization is stopped after 500 steps. Why? The stopping criterion should be based on the value of the objective function and/or it gradient. The number of required steps may be different, depending on the measurement density for example.

We observed that, for all Ψ considered, 500 iterations was sufficient to enter the "plateau" regime in which improvements to the objective function are not matched by significant further improvements in underlying K field L2 reconstruction error (as seen in Figure 2): the maximum reductions in field reconstruction error have been approximately achieved. As we are quantifying achievable K field reconstruction error, we considered it acceptable to terminate optimization at this point to manage computational cost, which was already very high for this study. In the absence of resource constraints, we would have done as recommended, optimizing all trials to fixed convergence criteria defined on the objective function. **We now mention our reason for imposing this fixed cutoff in section 3.4 of the revised manuscript.**

Typo Line 302, 411

We thank the reviewer for bringing these to our attention. **These have now been corrected.**