Interactive comment on “On the importance of multiple-component evaluation of spatial patterns for optimization of earth system models – A case study using mHM v5.6 at catchment scale” by Julian Koch et al.

Anonymous Referee #2
Received and published: 2 January 2018

The manuscript by Koch et al. proposes a multicomponent metric for evaluation and optimization of a hydrological model which can be used for any spatial pattern comparison. The topic is of interest for GMD and the manuscript is well structured, the conclusions well supported by adequate figures. I have no major concerns about the manuscript but a couple of suggestions that may help to improve the manuscript.

We would like to thank the reviewer for his/her elaborated review of our manuscript. Overall, we are very pleased that our efforts to promote and evaluate the SPAEF metric are generally well received by the reviewer. We will follow the suggestions made by the reviewer to revise our manuscript and believe that it will strengthen the scientific quality of our work. Our replies below indicate what we intend to change in the manuscript prior to resubmission.

The two major comments are:
1) Title: The title emphasizes that it is a method for Earth system models. While the manuscript strongly focusses on hydrological models. I am not a hydrologist and I found the Introduction too focussed on hydrological models and not very interesting for Earth system modellers. The title suggests a stronger overall discussion of Earth system models, while the whole paper is mainly about hydrological models, in the introduction as well as in the discussion. I suggest to remove the reference to Earth system models in the title to not raise wrong expectations.

We totally agree to that point. Our original idea was to promote SPAEF to a broader audience since GMD covers earth modelling disciplines beyond hydrology. However we agree to the reviewer that our work is limited to the modeling of hydrological systems which we will clearly state in the revised manuscript. Other disciplines of earth system modelling may also work with spatially distributed models, but the way these models are parametrized and calibrated may differ from the hydrological community. This should also be reflected in the introduction of the revised manuscript. We also intend to change the title to: “The SPAtial EFficiency metric (SPAEF): Multiple-component evaluation of spatial patterns for optimization of hydrological models”

2) your manuscript does not mention data uncertainty, while this could/should be a major component of a comparison metric too. if the model is within the uncertainty of observations further optimization would be overfitting. As more and more datasets provide data uncertainties, the possibility to include this information can be a major advantage over other metrics.

We agree that data uncertainty should be an elementary consideration when evaluating models and that a metric should ideally reflect this. We have decided to deal implicitly with *data uncertainty in our study. This was achieved twofold, first through temporal aggregation to monthly maps of evapotranspiration and secondly through the bias insensitivity of the promoted metric. The temporal aggregation will remove noise and uncertainties in the observations may cancel out. Monthly maps of ET will be less affected by uncertain rainfall variability and the dominant pattern influenced by soil and vegetation will become more apparent. The fact that SPAEF is bias insensitive will also alleviate the effect of uncertainties in the observations. In the end, we do not assess the exact values at grid scale, instead we investigate global characteristics such as distribution and variability which are expected not to be strongly affected by data uncertainty. The correlation coefficient is part of the SPAEF formulation as well and may be more prone to data uncertainty, but again, we investigate the overall allocation of high and low values which will control the correlation and uncertainty is likely not to have a strong effect. These thoughts on data uncertainty will be added to the revised discussion section.*

Specific comments:

There are a number of grammar and spelling errors throughout the manuscript. As Copernicus offers an editing service I do not detail these errors here.

We will pay special attention to detect any grammar and spelling errors during the revision of the manuscript. The remaining ones are then hopefully corrected by the editing team.

p.1 l. 20: " to the optimizer", the optimization issue was not introduced before and is not relevant here. stand-alone metrics do not only fail to provide the necessary information to optimizers, but also an evaluation or calibration can suffer from only one quantified characteristic.

We agree and will remove the reference "to the optimizer".

p.2 l. 1-3: I don't understand, earth system models usually have 2 spatial dimensions, but I dont see why they are and obstacle for modelling efforts. Do you mean the spatial scale or resolution? Even then I am not sure whether this is the major obstacle in general. Maybe it is for hydrological models? Otherwise please add a reference. It does not get clear from this sentence why this should be the case.

With the expression "spatial dimension" of earth system models we intendent to refer to the spatial variability. We agree that this may be confusing to the reader and will change it to the term "spatial variability".

p.2. l. 6-9. These developments could be interesting if you would give more detail. It would also put your work better in the context. Do these approaches already use multicomponent metrics? what are the differences between the approaches of spatial pattern oriented model evaluation? These examples are all from the field of hydrology? No other field of research has been dealing with such metrics?

We will elaborate more on the cited literature. The main point is that several other studies have highlighted the value of spatial observations in the evaluation of distributed models, but the main idea, to use multiple-components, has not been clearly addressed before. This marks the key novelty of our work and we will make sure that this is stated clearly in the revised version of our manuscript. We will extend the citing literature with examples outside the field of hydrology.

p.2 l. 9-11: Strange. In Earth system modelling spatial and temporal scales are quite related. For instance the necessary temporal time step depends on the spatial resolution. also parameterizations might require adjustments due to changes in temporal or spatial resolutions. Maybe this is very specific for hydrological models?

There may be a misunderstanding, we do not intend to refer to spatial and temporal scales. Instead we refer to spatial and temporal processes. The term "dimension" may be misleading and we will replace it with "variability". We want to point out that different parameters control temporal and spatial variability. The reviewer may be right that this a phenomena limited to the context of hydrological modelling, where our main expertise lies. In our experience, it is a challenging exercise to try to infer a meaningful spatial distribution of parameters by calibrating a hydrological model only against streamflow observations. The problem of equifinality arises where many parameter fields yield the same hydrograph. On the other hand, calibrating a model against spatial patterns only does not necessarily yield a meaningful hydrograph. This issue was addressed by Demirel et al., 2017 who applied the same model setup to conduct several calibration experiments: One using only streamflow data and another using only spatial patterns. This study highlighted the independency of the temporal and spatial observations and, when used jointly in a combined calibration, very limited tradeoffs in performance were apparent.

Demirel, M. C., Mai, J., Mendiguren, G., Koch, J., Samaniego, L. and Stisen, S.: Combining satellite data and appropriate objective functions for improved spatial pattern performance of a distributed hydrologic model, Hydrol. Earth Syst. Sci. Discuss., 1–22, doi:10.5194/hess-2017-570, 2017.

p.2 l. 15-16: It might depend on the application of the model, sometimes the spatial pattern might even be irrelevant and a good temporal performance is sufficient. At some later point you mention that the necessary performance depends on the application of the model, but it might be useful to mention this already earlier in the introduction.

This is an excellent comment. We will make sure to state this already in the introduction.

p.3 l. 1-5: are the requirements for earth system models and hydrological models the same? you claim your studies findings are imporant for earth system models but all your requirements and testing seem very focussed on hydrological models.

As mentioned earlier, we will remove the broad scope of earth system modeling and focus on hydrological modeling in the revised manuscript.

p.3 l. 9-12: if your variable has different units, ok. but if the unit is the same you might want your model to have the same mean or at least not a large deviation. That would then require an additional metric? how would you merge it then with your multicomponent metric?

If a bias term was desired in the spatial pattern evaluation it could easily be added to the SPAEF formulation, in a similar fashion as it is done in the KGE formulation. However we do not regard this as necessary, because bias-insensitivity allows the modeler to implicitly deal with data uncertainty. SPAEF focuses on the overall pattern and comparing the simulated and observed mean may overrate the quality of the remote sensing data. Most commonly, discharge timeseries data is available for hydrological modeling studies. Such *data allows for a reliable investigation of the overall water balance; i.e. the mean simulated and observed flow can be compared. However, the discharge data does not contain any information on the internal spatial variability of hydrological processes within a catchment. Here, the remote sensing data can make a significant contribution. We cannot expect that remote sensing observations can close the water balance through model calibration, but we can improve the internal distribution of hydrological variability of a catchment. Also, the remote sensing estimates represent a series of snapshots of cloudfree days in time and provide thereby not a continuous record. This further underlines why remote sensing data are not very well suited to address model biases. This will be stated clearly in the revised discussion.*

p.3 l. 15: the possibility to include data uncertainties could be another point. Remote sensing data inlude considerable uncertainties, optimizing the model by treating the "observed data" as the truth can lead to overfitting or biased model parameters especially if the uncertainties in the data scale with another important variable or increase with increasing values of the variable.

As pointed out before, temporal aggregation and bias insensitivity are ways to implicitly deal with data uncertainty, which we will clearly state in the revised manuscript.

p.4 l.30: this seems your way to partly deal with the data uncertainty.

Correct.

p.5, l.17, "source of information" this seems to be the wrong expression, probably a single metric or a single characteristic? single source of information sounds to me like using only one dataset to compare the model with as opposed to using multiple datastreams to optimize or evalute the model.

We agree and reformulate the sentence. The term "source of information" could be changed to "single component".

p.8, l.3: why are you doing a sensitivity analysis? Is this to select a limited set of parameters for the optimization? if yes please explain.

Correct, we have conducted the sensitivity analysis to select a limited set of the most sensitive parameters. mHM has 48 parameters and the sensitivity analysis has identified the 17 most informative parameters which then were estimated in the calibration. The reasoning behind the applied sensitivity analysis will be clearly stated in the revised manuscript.

p.8, l. 22-25: This seems to be a result, please move this paragraph.

Agreed. We will move this section to the results.

p. 12. l. 14: The insensitivity to bias can also be a disadvantage, in many cases the optimized model is desired to be unbiased.

We totally agree to this comment. We recommend to use remote sensing data in combination with discharge timeseries for the calibration of spatially distributed models. The discharge data will ensure that the overall waterbalance is in place (i.e. unbiased) and the remote sensing data will constrain the catchment internal distribution of fluxes. Again, the remote *sensing data is only obtained at cloudfree days and therefore does not provide a full record. This hampers the suitability of remote sensing data to assess model biases. We will clearly point this out in the revised manuscript. Especially that the bias insensitivity in the SPAEF metric is only reasonable when being accompanied by discharge data.*

p.12, l. 15: if the units differ, it might depend how the two units relate to each to other. it certainly is ok if they linearly scale. How about a nonlinear relationship? How about a possible change in sign as for instance with celsius and kelvin? if the mean temperature in celsius would go towards zero you would get difficulties for the beta part of your metric?

This is a very interesting point which will be discussed in the revised manuscript. We will advise the reader to investigate the relationship between the variables to be compared by SPAEF. In case there is a non-linear relationship one may consider to log transform the data. The variability term in the SPAEF formulation is mean normalized which should be quite robust. The histogram term is based on the z-score transform of the data which should also work for most cases. However we will inform the reader about alternative ways of normalization which may be relevant for certain cases. In case of non-linear relationships the transformation could be especially relevant for the correlation coefficient which assumes linearity.

Reproduceability: Will you provide your model outputs, observations used and analysis scripts?

All scripts used in this study are made available via GitHub and citable via Zenodo. Model outputs and observations will be made available upon request.
https://github.com/cuneyd/spaef
https://github.com/JulKoch/SEEM
https://github.com/mhm-ufz/mhm