Dear Reviewer,

thank you for your comments and for the detailed suggestions. Below we provide our responses and how the manuscript will be modified accordingly.

> ***Recommendation:***
> *I think the paper requires major revisions. I like the idea of introducing a comprehensively validated, modeled high-resolution land-atmosphere dataset, but regarding the present version of the manuscript and dataset I have several concerns which need to be addressed.*

Thank you for your positive feedback regarding the overall purpose of the study and of the dataset. We regret to know that some parts of the manuscript were not clear, but we believe that comments and suggestions were very useful to address these issues and strengthen the discussion. Specific responses and how the manuscript will be improved are listed below.

> ***General remarks:***
> *1.) "The applicability and purpose of the dataset is not entirely clear to me. The authors only briefly comment on this in the conclusions section, stating it could be used for exploring land-atmosphere feedbacks, investigating potential model simplifications, and data assimilation testing. Honestly, I do not see how the dataset can be useful in such analyses as (i) land-atmosphere interactions are known to be hardly robustly captured by models in general, and (ii) the testing of model simplifications and data assimilation would require different versions of the dataset with respective different model configurations in my opinion. In this context, I ask the authors to expand and clarify their discussion on the applicability of this dataset."*

Indeed, we agree with the reviewer that the purpose of this dataset is a key characteristic that should be clear from the manuscript. Within our project, we are working on developing and testing data assimilation approaches. More specifically, we aim to disentangle the value of different observation types in fully coupled models. We used the fully coupled model to develop a virtual reality. We faced two important issues. First, the reliability of the created virtual reality. Second, the huge computational effort needed to run a simulation with the desired spatial and temporal resolution. We decided to work therefore on this study to 1) provide a comprehensive assessment of the reliability of the states and fluxes simulated by the fully coupled models, and 2) make not only model code, forcings and parameters available but also the complete model output to guarantee the reproducibility of the entire model. This should allow the use of this model setup by other groups to do targeted tests, evaluating for example the impact of a different spatial model resolution, or a specific simplification (for example, neglecting lateral flows in the subsurface). This is challenging because we aimed at mimicking the real world without doing a formal model calibration. We believe that the comparison with real observations within the area support the capability of the fully coupled model to reproduce the water and energy cycles and should provide the confidence to use the dataset.

The dataset will be used for data assimilation experiment in our project, but is available for anybody interested in data assimilation and modelling experiments. We will extract (virtual) observations from the virtual reality and perturb the observations with a measurement error. The effectiveness of any data assimilation algorithm can be quantified making it a powerful tool for the development of such algorithms. In addition, thanks to the availability of a fully coupled system, we will be able to check the effect of assimilating observations in the different terrestrial compartments (e.g., assimilating groundwater level data and check effect on land-atmosphere fluxes like evapotranspiration).

In addition to this primary purpose, this dataset can be exploited for model simplification, to test reconstruction methods or define monitoring networks. Examples from our work are listed:

Baroni, Gabriele, Bernd Schalge, Oldrich Rakovec, Rohini Kumar, Lennart Schüler, Luis Samaniego, Clemens Simmer, and Sabine Attinger. "A Comprehensive Distributed Hydrological Modelling Inter-Comparison to Support Processes Representation and Data Collection Strategies." *Water Resources Research*, January 17, 2019. https://doi.org/10.1029/2018WR023941.

Lv, Shaoning, Bernd Schalge, Pablo Saavedra Garfias, and Clemens Simmer. "Required Sampling Density of Ground-Based Soil Moisture and Brightness Temperature Observations for Calibration and Validation of L-Band Satellite Observations Based on a Virtual Reality." *Hydrology and Earth System Sciences* 24, no. 4 (April 17, 2020): 1957–73. https://doi.org/10.5194/hess-24-1957-2020.

Haese, B., S. Hörning, C. Chwala, A. Bárdossy, B. Schalge, and H. Kunstmann. "Stochastic Reconstruction and Interpolation of Precipitation Fields Using Combined Information of Commercial Microwave Links and Rain Gauges." *Water Resources Research* 53, no. 12 (2017): 10740–56. https://doi.org/10.1002/2017WR021015.

In Baroni et al. (2019) a similar dataset created with TerrSysMP is compared to a distributed, conceptual-based hydrological model. The comparison was able to identify where and when simulations yielded similar results (and as such the simplified model could be used instead of the complex one) and where differences emerged (and as such identifying the need of model improvements and measurements). In Lv et al. (2020) the virtual reality was used to identify optimal sampling strategies for capturing soil moisture dynamics at a desired accuracy and precision. This is a basis for designing cost-efficient monitoring schemes. Haese et al. (2017) introduced a method to reconstruct precipitation fields combining precipitation measured by rain gauges and commercial microwave links. The synthetic dataset allowed the comparison of the reconstructed fields and its estimated error against the virtual truth. These are just few examples of how the data set can be used to test different methods or modeling strategies.

In addition, it is important to point out that indeed land-atmosphere feedbacks are difficult to capture by models. The advantage of our coupled modelling approach is that we replace the hydrology of the land surface model CLM3.5 with an approach where lateral subsurface flows (soil and, especially important, groundwater) are considered as well, and also lateral flows along the land surface (streams). As such, it is a state of the art model for representing land-atmosphere interactions, especially for the water and energy cycles. We agree that using CLM3.5 instead of CLM5.0 implies that the representation of vegetation and biogeochemical cycles in land-atmosphere interaction it is not the best we can do at the moment. However, a coupled land-atmosphere model using CLM5.0 and at the same time a full subsurface hydrology including lateral flows is not available yet.

We will add all these clarifications as to how we want to use this dataset ourselves and how other people can use it in the new version of the manuscript. We will clarify this in the introduction of the manuscript.

2.) "CLM3.5 is used here as land surface model. This is outdated. By now, CLM5 is clearly more advanced in terms of simulating processes related to vegetation and hydrological dynamics at the land surface (Lawrence et al. 2019)."

We agree that CLM has seen several improvements in the last years. Accordingly, an updated version of TerrSysMP with CLM5 is also in development and being tested. However, it is important to note that the upgrade to CLM5 would have only small impact on the dataset presented and discussed in this study. The main reason is that the hydrology including overland flow and lateral water flows in the subsurface including the representation of groundwater are being replaced by ParFlow. So the simplicity of CLM3.5 in that regard is not an issue. In addition, other CLM5 improvements, such as C and N cycles are not the focus of our dataset. For these reasons, while we agree that the use of CLM3.5 could be seen as outdated, the actual dataset created with the coupled modelling system is not, given the mechanistic representation of the terrestrial water cycle. We will better clarify this in the revised version of the manuscript.

3.) "Throughout the manuscript there are various inaccurate statements limiting the reproducibility of the model simulation (e.g. 'increased by about', 'increased by approximately', 'set considerably higher', 'needed to be increased from its standard values', various pedotransfer functions used without explaining criteria, see also respective specific comments below). More information needs to be provided in each of these cases to ensure the reproducibility of the entire analysis, either in the manuscript or in an appendix."

We will improve the description with more details in the appropriate sections in a revised version of the paper. We want to point out that the forcing and namelist files are part of the dataset as supplementary files, so the exact setup of our models can be re-created easily. A better link to them also within the text will be provided in the new version of the manuscript.

4.) "There are several arbitrary choices made throughout the study which need to be (better) motivated. This includes modifications of the modeling setup of which the purpose is not clear or the magnitude is arbitrary (e.g. 20% increase of sand fraction, ignoring of karst layers, conceptualization of alluvial layers as gravel and bedrock layers including the assumption(s) of values for various involved parameters, modifications of the LAI data). When making these modifications to adapt the model behavior in particular respects (more sandy soils to enhance infiltration) it should be kept in mind that even if the particular purpose is fulfilled, the land-atmosphere system is highly interconnected such that unforeseen side effects can occur. Further, the arbitrary choices include the approach(es) used to validate the modeled dataset (e.g. spatial averaging of model data across 25 grid cells for validation of atmospheric boundary layer characteristics, seemingly random time intervals of the soil moisture, evapotranspiration and runoff validations)."

Our aim was to generate a virtual reality for the atmosphere-land surface-subsurface system which is close to the reality. Therefore, a detailed comparison with measurements was made. It is usual in land surface and subsurface modelling that systematic deviations occur between model simulations and measurement data, which are related to specific parameter settings. It is therefore not surprising that we also find these systematic deviations, and in order to reduce these systematic deviations we made some additional adjustments:

- 20% increase of sand fraction: preliminary simulations with the model showed a wet bias related with too high soil moisture and groundwater levels by using the soil parametrizations derived by the original soil map. This wet bias has been detected in many other studies using Parflow or similar hydrological models (Shrestha et al., 2015, 2018) and it is generally attributed to the need of effective parametrizations that account for the spatial resolution of the model (see also discussion in the manuscript on that topic). Usually, few specific hydraulic parameters like the hydraulic conductivity or the Van Genuchten n value are modified to account for that. However, we found that these changes would create some inconsistencies in the modelling settings. This is specifically true in our coupled modelling framework due to a strong dependency between thermal and soil hydraulic properties. For this reason, instead of increasing a single specific parameter like e.g., hydraulic conductivity to account for effective model resolution, we decided to change the texture percentages. Accordingly, by applying pedotransfer functions to estimate hydraulic and thermal parameters, all the soil parameters are modified consistently. In the specific, the increase of the sand percentage of 20% has been applied to increase the soil infiltration capacity but to minimize the change in the soil classes detected in the original soil map.
- Elimination of karst. Karst cannot be represented well in standard hydrogeological models. We prefer not to claim that we can model karstic areas well with this approach.
- Conceptualization of alluvial layers as gravel and bedrock layers. Even though the Neckar river valley is (with some exceptions) quite narrow and has rather small alluvial layers, we found that alluvial layers needed to be introduced to increase subsurface water flow and reduce soil moisture contents in river valleys. This can be considered an additional calibration step to simulate groundwater levels and soil moisture contents closer to reality.
- Modification of LAI-data: this was done to remove bias from remotely sensed MODIS-data especially for winter time. This larger bias of MODIS-data is for example related to snow covered needleleaf forests.
- Spatial averaging of model data across 25 grid cells for validation of atmospheric boundary layer characteristics. This is done because we only considered the most dominant land-use type for a grid cell. However, in reality we would have a mix of several land use types and by choosing 25 grid cells we can consider this in our analysis.
- Time intervals for soil moisture, evapotranspiration and runoff validations. For each of the mentioned variables we want to show the most important aspect of the modelled behavior and we selected the periods accordingly. For river discharge showing we selected three years with relevant differences in the weather conditions (wet and dry periods). Some specific results are shown for one year as example to visualize the river discharge under changing seasonal conditions. Please also note that there is an error in the text where the time period mentioned is not the same as the one shown in the figure, this will be corrected. As for ET we show a basic analysis of the relation between ET and groundwater level and then show a selection of representative gridpoints to explain the observed behavior. ET and near-surface soil moisture react very fast to rainfall events and thus a smaller time interval is needed to analyze this. In addition, the chosen soil moisture time interval is related to the frequency of satellite overpasses.

These clarifications will be made in the different sections of the new version of the manuscript.

5.) "Soil types are an important ingredient for hydro-climatological model simulations. The downscaling-based derivation of soil types in this study is (i) difficult to understand and (ii) contains several assumptions which are not motivated, among which are the amount of considered 1995 locations, the 20% increase of sand fraction (see above), the choice of an exponential model, the choice of conditional co-simulation versus kriging, the focus on first three soil horizons (first means uppermost I guess?). I wonder what is the impact of the choices made here on the final dataset?"

We thank the reviewer to give us the opportunity to provide more details regarding the generation of the soil types. Indeed, the method to downscale the original soil map underwent several tests. Some of them have been described in (Baroni et al., 2017). We have referred to that paper also within the manuscript, but we summarize below the main steps for the sake of clarity.

The original soil map with necessary data for running the model is quite coarse with the areas of most of the soil polygons above 20 km2 (Figure 1).

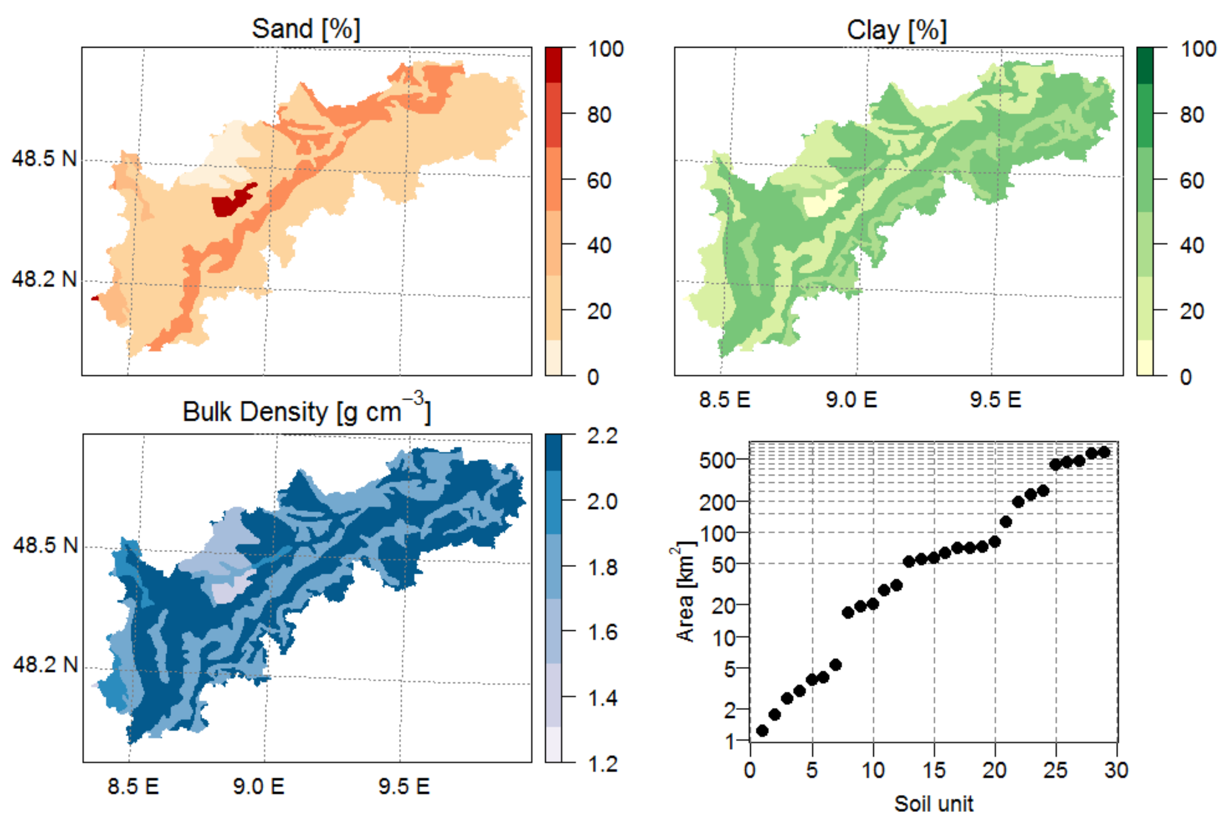

Figure 1: original soil map of the area. Please note that these figures do not represent the entire domain of the VR1 but they are reported only to show the main characteristics.

Based on that, different methods have been tested and compared to the new so-called conditional point method (CPM). The results are depicted in Figure 2.

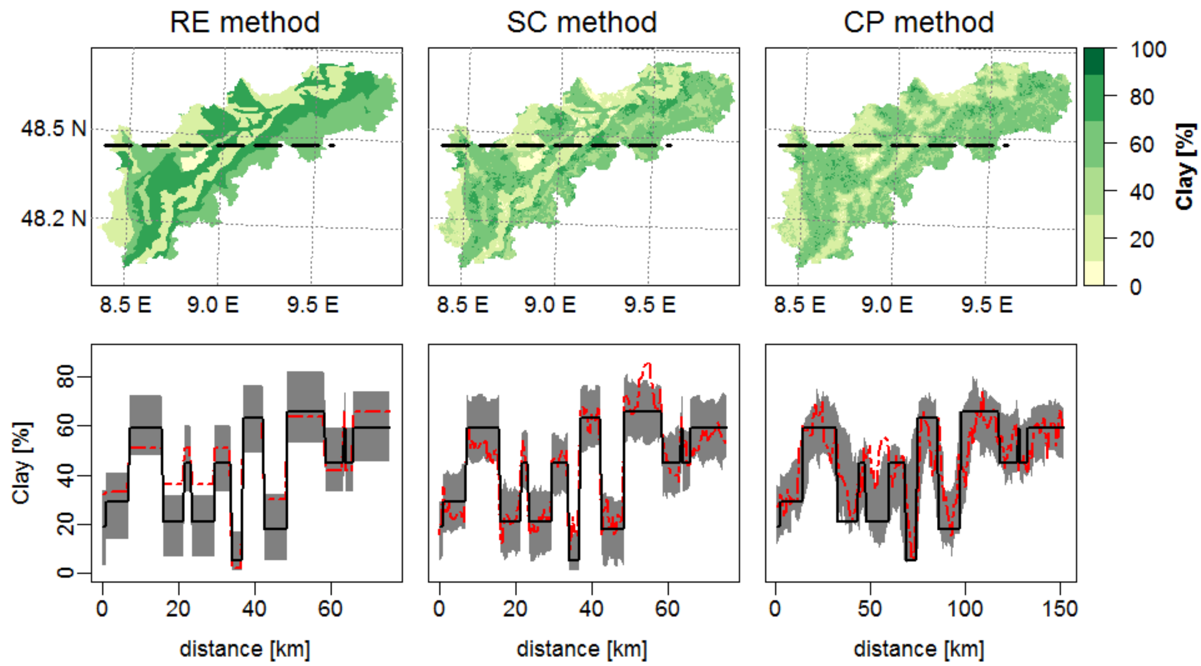

Figure 2: soil clay map created with different methods. From left: random error (RE) method where nominal value of each soil type has been perturbed, spatially correlated (SC) method where a spatial random variable has been superimposed to the original soil map (black line) and, conditional point (CP) method where conditional points are extracted and used in the conditional simulation.

The results showed that the CP method introduces uncertainty only at small spatial scales while the longer spatial patterns are preserved without sharp changes between the soil units. For this reason, this method has also been selected within the present study. Please also note that we did not consider the kriging approach as kriging removes small-scale variability which was considered an important feature to be considered in the dataset and coupled modelling.

The conditional point CP method is described in the next Figure 3.

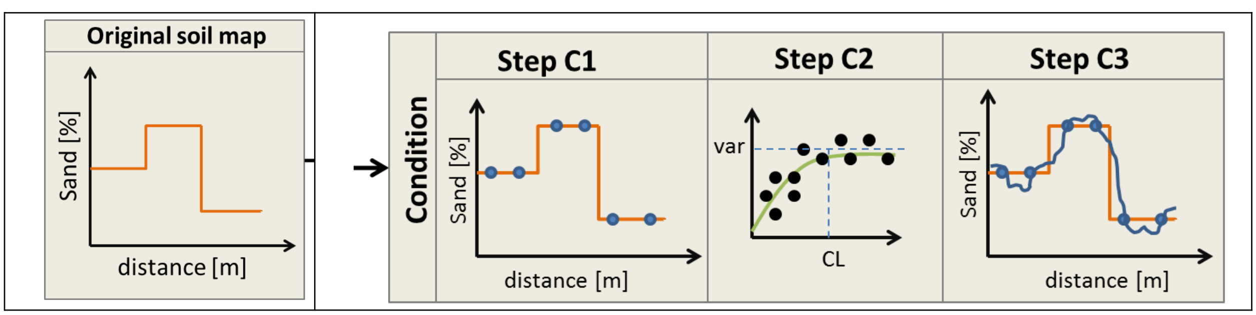

Figure 3: sketch of the method to downscale the soil map (a transect is depicted as example)

Starting from the original soil map, (step C1) soil samples are selected, (step C2) a variogram model is fitted and (step C3) a conditional simulation is created. In total 1995 locations have been used to mimic a realistic number of soil samples that can be collected for creating a soil map. More specifically a density of 1 sample per 5 km^2 is used. Some tests conducted for different sampling densities did not change the results significantly. The procedure will be discussed in greater detail in the new version of the manuscript. In contrast, the selection of the variogram model did not underwent a detailed analysis but was selected to provide a stronger short scale variability in contrast to the Gaussian variogram. This will be also acknowledged in the new version of the manuscript. Finally, the use of the first upper soil horizons has been considered because they represent the main soil horizons of the root zone (horizons A, B and C). Additionally, the use of only the first three horizons allowed us to have a variable soil depth under which we imposed the rocks. The original soil map provides more soil layers which would have resulted in a uniform soil depth of 2m.

As previously discussed, all these features have been selected after several tests conducted with the fully coupled model or with simpler configurations (not coupled, short period, limited areas), finally resulting in more realistic dynamics. We did not completely evaluate the impact of all these decisions on this dataset. This would require a detailed sensitivity analysis that is beyond the scope of the present study and also not feasible from the computational point of view. Still, we thank the reviewer for the comments and will clarify these aspects of the simulations in the new version of the manuscript.

> *6.) "The validation of the model simulation in terms of evapotranspiration is very limited. While it is reassuring that the ET and groundwater dynamics are broadly coupled according to expectations this is not a quantitative assessment. The modeled ET could instead (or additionally) be compared with state-of-the-art evapotranspiration datasets such as GLEAM (Martens et al. 2017) or FLUXCOM (Jung et al. 2019) at larger spatial scales."*

We have limited the scope of the ET analysis since a general analysis of how TerrSysMP performs with respect to ET was already done (Shrestha et al., 2017 https://doi.org/10.1016/j.jhydrol.2018.01.024). Since our land surface is different from the real land surface (the land-use is similar, but restricted to a dominant land use type per gridcell and LAI is fixed without interannual variability) a comparison to any real world dataset would be biased and it would be very hard to figure out if the differences seen in the analysis are due to the difference in land-use, neglecting inter-annual variability or differences related to the model biophysics. In addition, the datasets GLEAM and FLUXCOM are not measured values, but also affected by strong model assumptions (GLEAM) and/or interpretation (e.g., energy balance errors at EC-sites) and interpolation (for FLUXCOM) of data.

> *7.) "I like the comprehensive validation of the dataset in terms of several variables - an overview table summarizing the determined strengths and weaknesses would be helpful for users I think."*

Thanks for the suggestion. We will add such a table to the manuscript.

> *8.) "There are too many figures in my opinion, diluting the main messages. Figures 3-5, 7, 14, 17 could be moved to supplementary, and Figures 9 & 10 could be combined."*

We will reduce the number of figures and move to supplementary. Please note that there are also many specific comments for the figures, automatically leading to a reduction in the number of figures as suggested.

Specific comments:

“line 75-78: you do not aim to reproduce to observed catchment dynamics but still validate the model in some respects - this seems contradictory to me; what is the aim here if not validating the model against observations? how useful is a modeled dataset for the community if is not resembling observations?”

The aim of the validation was to show that the model system, despite its simplifications, is still able to simulate all the important core processes adequately. Therefore, we compared many of our output variables to observations to see if we had large biases or other undesirable behavior by the choices of our setup. However, a full reproduction of all measurement data is not possible as this requires the calibration of an integrated atmosphere-land surface-subsurface model, which still has not been documented in the literature. In summary, we want to check that systematic biases are small, but are concerned with random deviations between model and measurements.

“lines 139 & 145: the chosen time period and simulation catchment/area are not motivated”

We have chosen the Neckar because it has varying topography and land-use and typical central-European climate. A further motivation to choose the Neckar catchment was that groundwater levels are relatively high so that groundwater-atmosphere feedbacks are more likely and prominent in this catchment. We wanted to have a long simulation time period while still having boundary data from a convection-permitting model since our domain is rather small and a huge boundary zone would have been needed if forcing from a coarser model was used. 2007 was thus the first year we had access to such a forcing and was chosen as the start of the simulation for that reason alone. In the revised version of the paper we will clarify these points better.

“line 172: please give more information on the ’software restriction’”

This restriction was actually a combination of software and hardware. The machine we used was optimized for using many relatively slow cores. As such the only way to run the simulation was to use many cores at a time. As the number of grid cells increases so does the number of cores and at a certain point the model would no longer run. We never found the root cause but the result was a limit to the domain size in terms of grid cells. Disabling I/O did increase the limit before the model would stop running but doing so is obviously not feasible. We strongly believe that this issue was related to the specific setup of the computer we used and is likely not to occur on more modern machines with higher per-core performance.

“lines 194-197: in the abstract of the Tian et al. 2004 paper I found “On average, the model [...] overestimates FPAR over most areas in the Northern Hemisphere com- pared to MODIS observations during all seasons except northern middle latitude summer.“ “The MODIS LAI is generally consistent with the model during the snow-free periods...“ which makes me wonder why the authors modify LAI in summer? Further this could create jumps in the LAI time series from May-June and August-September. More importantly, you state here that LAI is used “for the year 2008“. Does this mean there is no interannual vegetation variability? This would affect evapotranspiration and thereby many related variables and would need to be stated as a serious shortcoming.“

Since we do not use MODIS land-use but CORINE, we had to re-map the MODIS observations to the CORINE land-use map. As such you would have a range of values for each land use type. We wanted to keep things simple and used just one value for each land use type at a given timestep. So, for the summer we chose values from the upper end of that range while only larger changes were made to LAI for needle leaf trees in winter. There are no jumps in the dataset. All years have the same LAI-cycle, so there is no interannual variability. We will acknowledge this in the revised version of the paper.

> "lines 239-240: 'to avoid the manifold hydrological challenges related to its modeling', please be more specific here, also please comment on the impact of this simplification of the approach on the final dataset"

The most obvious impact is the one of subsurface cave systems. These can be vast while being subgrid scale at the same time. They may only have an impact locally, or be connected to a larger system of caves that can re-distribute groundwater going as far as re-distributing smaller rivers (the Danube sinkhole is one example). They also can go deeper underground than what we are modelling. However, some regions may not be affected by the underground caves if they are not connected to the surface at that point while other regions are greatly impacted. Detailed maps of the cave system are not available. Even if such maps would be available, it would be an enormous effort to implement them or find parameterizations which could replace the karst system. Finally, the impact on the near-surface soil hydrology will only matter locally and for much smaller areas. We argue that including karst areas is beyond the scope of this study given the huge amount of time needed to represent karst systems adequately in the subsurface. In addition, a deeper vertical extent of the model domain would be needed as well. We will provide a clearer formulation in the revised version of the paper.

> "line 245: if these alluvial bodies are so relevant, why does this study use datasets which do not include them?"

The alluvial bodies of the Neckar are very small since the Neckar is most of time confined to a rather narrow valley. With our rather coarse resolution we are barely able to capture these valleys accurately. The soil map, which we used as basis for our soil hydraulic properties, did not contain these alluvial bodies, as these bodies are not soil. On the other hand, the geological map is much coarser and does not include these fine structures, although they are from the subsurface flow perspective important. We first tried if we could disregard the alluvial bodies altogether, similarly to the karst features, but found out that this has a large impact even for rather small streams which made it necessary to include them and due to the small size of the valleys our chosen extent of three grid cells is already generous in most locations. We will make in the manuscript some additional clarifications regarding the alluvial bodies.

> "line 249: evapotranspiration errors in models can be significant and might be underestimated here"

Previous studies (Shrestha et al, 2017, https://doi.org/10.1016/j.jhydrol.2018.01.024) have shown that for a climate as ours transpiration is energy limited. Therefore, any difference to real values would be due to LAI differences. The study also shows that we are likely overestimating ground evaporation due to the resolution effect discussed. That means that our value of 30% subsurface flow is actually a conservative estimate.

> "section 4: please discuss for the performed validation analyses how the determined performance of the study dataset compares generally with the performance of other regional climate models in similar hydro-climatic regions"

This is beyond the scope of this study. We provide all states and fluxes from the subsurface to the atmosphere and a major contribution is that these states and fluxes are also provided for the subsurface, including groundwater levels, groundwater flows, lateral water flows in soils, and river discharge. We wanted that this virtual reality is realistic and not far away from reality, for all those compartments of the terrestrial system. The aim was not a comparison with other RCM´s.

> "line 331: the potential (dis)agreement of simulated and actual land cover can be checked using high-resolution land cover datasets such as provided by ESA CCI

While we can check this, the mismatch will remain (and is mostly related to resolution effects and our choice to only use one land-use type per grid cell) and needs to be considered when comparing our results to actual observations. Attributing differences is much more difficult which is why we take the approach to average them out as much as possible.

> line 343: how are the temperature standard deviations determined?"

This standard deviation can be considered the mean absolute difference between the observations and our modelled results. This is done for each time of day separately.

> "line 367: why not using the ESA CCI soil moisture dataset derived from observations of various satellites for this validation?"

The remotely sensed soil moisture cannot really serve for validation. The simulated soil moisture content will be as reliable as the remotely sensed soil moisture content, so we prefer not to use ESA CCI for validation.

> "lines 394-395: I do not really understand why this daily matching is applied here? Also it is not clear how this is done."

The soil moisture overestimation by the virtual reality is largely responsible for the significant bias found when comparing synthetic with real observations of brightness temperature (Figure 15 orange-line). This soil moisture overestimation is not constant along the year, therefore a daily soil moisture correction factor was found in order to apply a correction factor to the soil moisture input for the radiative transfer simulation. As indicated in line 392, the correction factor is found by matching the cumulative distribution functions of soil moisture from virtual reality and satellite retrievals, this correction factor is then used to simulate the adjusted brightness temperature which is shown by black lines and shaded IQR area in Figure 15 a and b.

> "lines 447-450: I do not understand how the "fluctuations" are "scaled". Do you divide by the inter-annual standard deviation to obtain normalized anomalies (or z-scores)? If so, please name it this way as the term "fluctuations" is rather unclear."

The reviewer makes a fair point that this section was rather unclear. For Figure 19b, we simply compare the groundwater observations when the mean over the plotted period is removed, hence all lines in the plot have a zero mean. Apart from that, the figure is not scaled or shifted. For Figure 19b, scaling means that both the magnitude and the mean is manually shifted so that both curves can be compared in the same plot. Therefore, the y-axis on the figure lacks units and we simply talk about a trend. Hence, no formal scaling technique has been used. In a revised manuscript, this section will be better explained.

> “lines 451-452: How is this trend computed?”

For the simulation this is the area-averaged changes of groundwater level over the full model domain, while for the observation wells it is the arithmetic average over all wells. As for the obvious difference between the two and the hence required scaling to compare them, please see the answer to the question above.

> “lines 502-512: As there are multiple concrete ideas to improve the model setup and consequently the dataset, why not implementing them before publishing this dataset? comment if this will be game-changers"

Most of these changes are related to resolution. Going from 400m to 200m or 100m for the land surface and subsurface would obviously improve several aspects of the simulation, specifically for river discharge. However, such a high resolution is currently hardly feasible and computing resources would not be available to repeat the simulations at such a high resolution. We already went at the limits of what is currently feasible. Other improvements can be applied with less effort but still would require the entire simulation to be re-run, since they were developed after the simulation was done (as a way to improve the results we saw actually). While these changes would improve the simulation, we do not expect results to improve a lot. The only exception may be river discharge during flood events. Again, re-running this very large simulation (including a new spin-up) would consume a very large amount of compute resources. In the future with much improved computational power it would be of interest to do an updated version of such a run using ICON instead of COSMO, CLM5.0 instead of CLM3.5 and a GPU-based ParFlow version in order to see improvements for all terrestrial compartments. However, this new version of TerrSysMP (ICON-CLM5.0-ParFlow-GPU) is currently under construction and not ready in the next six months.

In addition to the specific comments above there are several more comments of minor deficiencies that can easily be fixed in an updated version of the manuscript. We did not include all of these here specifically.