Review of: "**Assessment of Precipitation Error Propagation in Multi-Model Global Water Resources Reanalysis**"

Authors: *Md Abul Ehsan Bhuiyan, Efthymios. I. Nikolopoulos, Emmanouil. N. Anagnostou, Clement Albergel, Emanuel Dutra, Gabriel Fink, Alberto Martinez de la Torre, Simon Munier, and Jan Polcher*

General Comment:

This paper compares outputs of four land surface models (LSMs) and a global hydrologic model (GHM) in the Iberian Peninsula forced by different precipitation (P) products for a period of 11 years. Precipitation products include satellite, reanalysis, and combined (stochastically generated) products. The SAFRAN precipitation products, which merge reanalysis and gauge observations, and the hydrologic simulations obtained with these precipitation inputs are assumed as reference.
The authors perform a set of analyses to evaluate how the uncertainties due to precipitation products and model structure affect three hydrologic variables, including surface runoff, subsurface runoff, ET.

The topics of the paper are interesting for the audience of HESS. The paper is, for the most part, well written. Thus, I am supportive of its publication. However, in my opinion, there are a few unclear parts in the text and analyses that require to be addressed first.

Major concerns:

1. It is not clear how the metrics used in the analyses are applied in terms of space and time aggregation. This should be clearly specified for each metric in the methodology section. For example, what is/are the time step/s of RD? How is this metric used in figure 5? Is figure 5 presenting the distribution of the RD's in all pixels (i.e., how is space considered)? Similar questions arise for the boxplots, the Taylor diagrams, and the CV. Please, clarify.

2. The time resolution of the satellite-based P products is 3 hours. How about the other two products? This, in combination with the resolution of the hydrologic simulations, affects the interpretation of the ability to simulate the hydrologic processes (notably, surface runoff).

3. The authors should provide in the Methodology section three details on the hydrologic simulations and their evaluation:

(i) What is the time resolution adopted for each model?
(ii) Was the model calibrated (I guess only one was) and, if not, which set of parameters was used?
(iii) State that: (i) simulations are evaluated for long-term averages of annual, daily and, in some cases, 3-hour variables (see comment 1); and (ii) no seasonal analysis is performed.

4. The first result that I was expecting to see is the comparison of the bias between SAFRAN and the other P products (a figure like figure 2 but for P). This would give immediately an idea of what to expect for runoff and other hydrologic variables.

5. Related to the previous point: In my opinion, time series (at monthly resolution?) of spatially-averaged P, Qs, Qsb and E would be quite useful to have an idea of how the models vary among each other, across years, and within each year.

6. The analyses of the ensemble spread is not properly introduced in Section 3.4. What are the ensemble members referring to? Also, the definition of the metrics and associated symbols is not clear. Things become a bit clearer in section 4.4. However, I think that Sections 3.4, 4.4 and Figure 13 should be eliminated, since, as it stands, this analysis is superficial and does not add much to the message of the paper.

7. The interpretation on page 8, lines 21-23 is counterintuitive or I did not have enough information to understand it (see comment 3). To me, if an LSM is run at 3-hour resolution with a P product that has the resolution of 3 hours, there are higher chances that infiltration-excess runoff will be generated. This is because P products should be able to capture storms localized in time. In contrast, if an LSM is run at 3-hour resolution with a P product that has the original resolution of 24 hours and a uniform P intensity is assumed to create inputs at 3-hour resolutions, then the chances are lower.
On the other hand, if an LSM runs at 24-hour resolution and it has not been calibrated with P products at 3-hour resolution, then we can have unexpected effects on the generated runoff. In this case, I am not able to say a-priori what we should expect. Thus, the biases that the authors have found may be an effect of the calibrated parameters, rather than the model structure. I suggest the authors to clarify this part and elaborate more.

Minor concerns:

P 4, lines 12-16. This sentence seems too long.

P 5, lines 14-17: Please revise the sentences on climate and “topography in the Pyrenees”. It doesn’t make sense to me.

Section 2.1: Can the authors provide some quantitative information on the SAFRAN performances against rain gauges?

P6, line 4: consider using the acronym LSM for land surface model. Otherwise, don’t capitalize “L”.

P 15, line 22: I could not verify in the figure that NCRMSE > 0.75 for surface runoff simulated with 3B42 in all (or most) cases. Can the authors check this again and explain?

P. 16, lines 8-10: What are the implications of this? Is it expected? I could not figure this out by myself without knowing for which time scale the CV was computed.

P. 16, lines 10-13: I could not verify this interpretation in the figure. The median of the boxplots for SURFEX are for the most part larger than 1. Please, clarify.

Section 4.2: I suggest moving the sentence on page 16, lines 14-20 after line 8, as I believe that the comment on precipitation should be provided first.

Figure 7 should be improved. There are labels in the y axes only in some panels.