Reviewer #2

#2.1. The manuscript “Uncertainty sources in a large ensemble of hydrological projections: Regional Climate Models and Internal Variability matter” by Evin et al. investigates the outcomes of several climate-to-hydrology modeling chains over the area of France. The perspective is to quantify the uncertainty introduced at any level of the chain by the various combinations of models that can be used at any level of the chain. This study is of great interest because it highlights the value and limitations of future hydrological projections, providing advice on which hydrological indicators can be predicted more robustly and where within the study area.

We thank the reviewer for this positive feedback.

#2.2. The work is a great effort of synthesis but leaves incomplete a fundamental point that I believe should be extended and integrated in the work. In the Discussion section (and recalled also in the Conclusions) the authors note that, in some cases, some models/combinations of models are discordant, then contributing significantly to the overall uncertainty. They suggest (page 26) to “look carefully at these deviating models and understand the reasons behind the atypical behavior”; then, if the model deviates for “wrong reasons (numerical artifacts, bugs, model inadequacy)” should be discarded, otherwise the analyst should investigate the reason of the dissimilarity further. In this case, the authors acknowledge the possibility that the model is superior to others. This point is extremely important, and the authors also acknowledge it at the end of page 26 and in the Conclusions section.

However, I would have expected a more thorough investigation of this point, i.e. a deeper analysis of “what to do” in these cases. Clearly, it is not possible to provide general instructions valid for any case, but the study provides a very good base to identify a few specific cases to develop in more detail. For instance, one could look at a specific catchment/climate regime and decide whether one model/model chain should be discarded or, on the contrary, should be preferred because more reliable in a specific context (see e.g. the cases reported in section 5.4.3). I understand that this is not an easy task, but it is necessary to make the paper much more than a descriptive product.

We thank the reviewer for this comment and we agree that this is a crucial point that was a bit eluded in the manuscript. This point deserves a dedicated subsection in the discussion and will be added in the revised version. We also agree that a detailed and motivated selection for a specific case would be interesting. However, this is not an easy task and this would probably require an additional section. As the manuscript is already rather long, we leave this detailed illustration for a further analysis. We will strengthen the point that a good model for the recent climate is not necessarily a good one for modified climates and conversely a rather poor model for the recent climate can suggest important hydrological changes that can not be identified with highly calibrated conceptual models.

For your information, note that the Explore2 project produced two reports (Sauquet and Héraut, 2023, Sauquet et al., 2025, in French) dedicated to the diagnostic assessments that have been performed for all catchments. They provide an overview of the qualities and performance of surface hydrology and hydrogeological models when driven by the SAFRAN reanalysis. Sauquet and Héraut (2023) also provide recommendations to the end users to help them select or filter some model chains for a specific catchment. Uncertainty assessments shown in the manuscript are thus complemented by a series of thorough diagnostics on the reference period. The purpose of these diagnostics is not to pit the models against one another; rather, it aims to support the selection of a group of models—when several are available—to be prioritized in a prospective exercise. In addition, Héraut et al. (2024) describes some choices which have been to identify atypical simulations: 1. Using the mean annual flow QA on the reference period: a simulation is not plausible if its QA is outside the range [0.5*med(QA), 2*med(QA)], where med(QA) is the multi-model median of all QA values; 2. Using the QA anomaly between the end-of-century QA and the QA for the reference period ΔQA: a simulation is not plausible if ΔQA is outside the range [med(ΔQA) +/- 3*σ(ΔQA)], where med(ΔQA) and σ(ΔQA) are the multi-model median and standard deviation of all ΔQA values, respectively. If, for one catchment, 50% of the projections corresponding to an hydrological model are not plausible, then, it can be considered that hydrological projections are flawed for this model (imperfect representation of the hydrological processes, wrong drained area) and they are all rejected. This loose set of criteria mainly aims at automatically filtering simulations which exhibit obvious shortcomings.

The recommendation made to the end users of the simulations is to examine summary sheets produced at different levels of aggregation, from local (station level) to regional (hydrographic region level), and make their own choices. In our opinion, it is not possible to recommend a set of models a priori without knowing how the data will be used. Indeed, the choice must be guided by users' needs—for example, whether only streamflow is required, or streamflow together with other variables, or streamflow at a prescribed set of simulation points available. The choice is constrained by the models' ability to provide all desired variables at the points of interest.

As an illustration, this guideline has been followed for a study dedicated to the future water resources of the Isere department, an area located in the French Alps (see report in French: https://www.isere.fr/sites/default/files/2025-05/livrets-methodologiques.pdf). A compromise between models' performances and diversity was made and the set of selected hydrological models were MORDOR-SD, GRSD, SMASH, SIM2 et J2000 for high flows, MORDOR-SD, GRSD, SMASH et SIM2 for mean flows, and MORDOR-SD, GRSD, SMASH and J2000 for low flows.

It is worth pointing out that the discrepancies observed between reference data and simulated data are not solely attributable to the models. The analysis relied on reference datasets. Despite the care taken during the selection of reference points and the filtering of outlier data, reference datasets may still contain errors or residual influences. In addition, forcings in mountainous areas (which are inherently heterogeneous) may lack precision or, more generally, may fail to capture convective precipitation.

Before any use of hydrological projections, it should be recalled that strong model performance under present-day conditions does not necessarily imply reliability under climate change. This limitation is particularly pronounced for empirically based or conceptual models, which rely on empirical relationships and parameterizations calibrated to recent climate conditions. Such models are often finely tuned to reproduce observed behavior as closely as possible. This is, for example, the case for conceptual hydrological models in which snowmelt is represented using temperature-index approaches (e.g., degree-day methods) and evapotranspiration losses are estimated from potential evapotranspiration formulations. As a result, these carefully calibrated models often exhibit high performance and may outperform physically based models that represent processes more explicitly. However, the ability to reproduce past observations does not guarantee that a model will adequately represent processes under altered hydroclimatic conditions. Model evaluation is therefore a critical issue in model selection. While it must assess the capacity to reproduce observations, it should also examine—when possible—the temporal transferability of models. This second aspect of evaluation is difficult, and often impossible, but remains essential. A particularly critical issue concerns evapotranspiration losses. The feedbacks of increasing atmospheric CO2 on plant phenology and, consequently, on evapotranspiration are typically not represented in hydrological models. These feedbacks may lead to substantially different future water balances. Physically based models that allow exploration of such interactions therefore deserve consideration, even if their performance is lower than that of highly calibrated empirical models. Ongoing climate and hydrological changes may provide opportunities to assess model behavior under modified conditions, although targeted and enhanced observations will likely be required to support such evaluations.

In an ideal world, the research community should examine the divergences between models driven by the same climate projections and investigate their origins (e.g. by running additional experiments). For example, in Explore2, the hydrological model ORCHIDEE shows important discrepancies with the other hydrological models for many catchments. In continuation of Explore2, developers of the model ORCHIDEE have clearly identified what parts of the land surface processes could be better represented (groundwater module for drainage, snow representation in mountainous catchments, Huand et al., 2024). On the other hand, developers of conceptual models have also many perspectives to improve the transferability of their modelling framework for future climates, in particular regarding the evolution of the vegetation (land use, vegetation type, interactions between vegetation and the carbon cycle).

Héraut, Louis; Vidal, Jean-Philippe; Évin, Guillaume; Sauquet, Éric,"Notice de lecture des fiches de résultats des modèles hydrologiques de surface", Recherche Data Gouv, 2024. INRAE. https://doi.org/10.57745/6YNIUF

Huang, Peng, Agnès Ducharne, Lucia Rinchiuso, et al. « Multi-Objective Calibration and Evaluation of the ORCHIDEE Land Surface Model over France at High Resolution ». *Hydrology and Earth System Sciences* 28, nº 19 (2024): 4455-76. https://doi.org/10.5194/hess-28-4455-2024.

Sauquet, Éric, Louis Héraut, Jérémie Bonneau, Alix Reverdy, Laurent Strohmenger, and Jean-Philippe Vidal. *Diagnostic des modèles hydrologiques : Des données aux résultats*. Recherche Data Gouv, 2025. INRAE. https://doi.org/10.57745/S6PQXD

Sauquet, E, and L. Héraut. « Notice de lecture des fiches « diagnostic » des modèles hydrologiques ». Recherche Data Gouv, 2024, INRAE. https://doi.org/10.57745/MDHS0D.