Geosci. Model Dev. Discuss., referee comment RC2
https://doi.org/10.5194/gmd-2020-329-RC2, 2021

Comment on gmd-2020-329

Anonymous Referee #2

Referee comment on "Impact of Initialized Land Surface Temperature and Snowpack on Subseasonal to Seasonal Prediction Project, Phase I (LS4P-I): Organization and Experimental design" by Yongkang Xue et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-329-RC2, 2021

This paper describes the international / multi-organization LS4P project and provides some initial analyses of data submissions. It is essentially an "introduce the community to the project" paper, with most of the scientific findings to be documented later, as the project progresses.

While the paper is well written, in my opinion it has enough issues to rate a review of "major revision". On the one hand, I do applaud the experiment organizers for bringing together such a wide-ranging group of participants to address such an important problem. With this diverse a group and a corresponding collection of model outputs from such a diverse set of models, I don't doubt that the project will bear useful scientific fruit, and a paper like this that introduces the project to the broader community is certainly of value. On the other hand, the paper's write-up glosses over several critical aspects of seasonal and subseasonal prediction that call some of the project's long-term strategies into question, at least in terms of how they're currently described. A revised version should address these shortcomings through substantial qualification (not just a sentence or two here and there) or, better yet, through a substantial rethinking of the approaches to be applied.

Response: We thank for the reviewer's acknowledgment that we have a wide-ranging group of participants to address an important problem and that the project will bear scientific fruit. We have taken into careful considerations of the reviewer's comments/concerns on our writing and questioning of our strategies. The following is our point-by-point responses to address the issues that the reviewer raises. We believe some issues are rooted in the reviewers understanding in our experimental design, thus in our revisions, we have made a great deal of effort to improve the presentation of our experimental design section. The following responses explain and address the reviewer's questions.

1. The underlying assumption of the project appears to be that if a model does not produce an accurate temperature over the Tibetan Plateau, the fault lies with the land model (e.g., in how long the land model maintains an initial condition). The paper states this explicitly on line 465. The truth is, all models have biases in both air temperature and precipitation across the globe, and these biases could have any number of sources. A temperature bias over the Tibetan Plateau might have nothing to do with land model processes. It might instead result from deficiencies in the reproduction of the general circulation, for example, or from some problems with the radiation balance. Forcing the model to have a low temperature bias by imposing a stronger initial temperature anomaly (perhaps even an unrealistic anomaly, through eq. 1) may amount to "getting the right answer for the wrong reason", which is not a good basis for a forecast experiment. It's

quite possible that forcing a correct temperature through such an initialization when the model wants to do something else for reasons unrelated to land processes might have unexpected negative consequences – especially if the model is artificially modified in one region and not in surrounding regions. Substantial discussion regarding this is needed.

Response: we fully agree with the reviewer's comments that "all models have biases in both air temperature and precipitation across the globe, and these biases could have any number of sources," and when you correct the bias you may be "getting the right answer for the wrong reason".

But we have to clarify a few issues:
(1). Numerous modeling studies, since the very beginning of meteorological model development, have worked on correcting model-produced precipitation and temperature errors by improving some model parameterizations and through improving initial and boundary conditions. Our approaches and statements are based on a number of published papers in our field's major journals. We cannot simply speculate these peer-reviewed research's results are produced due to the wrong reason unless there is an evidence to support such statement. Otherwise, we may eliminate any modeling improvement studies because, at least at its early stage, they cannot prove that improvement is not due to a wrong reason. Only the community collective efforts with long term exercise can prove it.

(2). The reviewer speculates the problem in simulating the LST anomaly may not be due to the surface models. In general, when we try to correct a model deficiency in one variable, it is normal to check the dynamic and physical processes relevant to this variable first. For instance, for the precipitation simulation errors, we naturally check the convective and cloud process modeling and surface evaporation parametrization first. Unless there is an evidence to show these are other reasons, we cannot claim it is a wrong approach. Our statement on land model deficiencies are based on published papers and analyses from the LS4P research. On lines 126-129, we present a publication (Liu et al., 2020), which focuses on exploring the causes of model deficiency in properly producing the observed surface temperature anomaly in high mountains. That study demonstrated this deficiency IS associated with the land surface process model, including snow/albedo and soil subsurface memory effects. Recently, we had another published paper (Li et al., 2021) address this issue and further confirms Liu et al.'s conclusion. In addition, we have also analyzed the reanalysis data, which are used for model initialization, and indicated that the deficiencies in reanalysis data in high mountain areas also contributes to the simulated surface temperature bias. We believe we have caught the main (if not all of the) reasons for the model deficiency for this aspect. We have more clearly indicated our statements are based on the published papers in abstract and conclusion, and welcome more research to explore this issue in the revised version (Lines 493-496)

Liu Y., Y. Xue, Q. Li, D. Lettenmaier, and P. Zhao, 2020: Investigation of the variability of near-surface temperature anomaly and its causes over the Tibetan Plateau. J. Geophy. Res. 125, e2020JD032800. https://doi.org/10.1029/2020JD032800.
Li, Q., Xue, Y., and Liu, Y.: Impact of frozen soil processes on soil thermal characteristics at seasonal to decadal scales over the Tibetan Plateau and North China, Hydrol. Earth Syst. Sci., 25, 2089–2107, https://doi.org/10.5194/hess-25-2089-2021, 2021.

(3). However, to improve land model and reanalyses data for this aspect are not simple tasks and may take decadal effort (today's land temperature model development has a more than 70-year history), but proper S2S prediction, including drought/flood/heat wave prediction, is a WMO task requiring urgent attention owing to a significant societal demand. On line 377-379, based on the LS4P modeling group's practice, we have pointed out that preliminary research suggests "prescribing both LST and SUBT initial anomalies based on the observed T-2m anomaly and model bias is the only way for the current ESMs to accurately produce the observed May T-2m anomalies". In fact, using initialization to improve meteorological prediction is nothing new but a traditional meteorological approach. We believe when this issue gets more attention and more data are/become available, more methodologies may be developed to address this issue. But the scientific development always takes time. We have to allow any development to take step-by-step improvement.

(4). We agree with the reviewer that the statement on line 465(previous version) may be misleading and therefore we have modified the text to more properly reflect the ideas that we present in this paper (Lines 493-501).

2. The overall strategy seems to ignore the fact that forecast models generate their forecasts relative to their own climatologies. A model that is known to be biased warm may produce an anomalously cold 2003 over the Tibetan Plateau (compared to what it usually produces there), and that would be useful information even if the forecasted temperatures are still warmer than the average observed TP temperature. The point is that people know how to account for long-term model biases. They would properly consider a forecast model's result to be "2003 will be colder than usual by 5 degrees" rather than "the temperature will be 20C". The emphasis here on matching the observed temperature in absolute magnitude seems inappropriate to a discussion of forecast systems. See, e.g., the NMME forecast anomaly pages [https://www.cpc.ncep.noaa.gov/products/NMME/seasanom.shtml], which show forecasted anomalies relative to each model's climatology.

Response: The reviewer raises several issues here.

(1). Because of the model systematic bias, some groups indeed applied the anomaly prediction in their normal practice. We understand the justification for these groups, including some LS4P groups, used the anomaly predictions, but predicting the temperature in the real world is always our ultimate goal because the public needs the forecast for the real world, not the forecast relative to one group's model climatology.

(2). However, in the LS4P experimental design, as the first step, we only intend to see if there is any relationship between the observed Tibetan Plateau spring LST/SUBT anomaly and downstream summer precipitation anomalies. We mainly look at the difference between Task 3 and Task1. In this way, the model systematic bias has been eliminated so what we really look is indeed the anomaly, which is just what the reviewer tries to emphasize here. Because of this approach, our goal here does not emphasize producing the best initialization for May 2003 per say, although the methodology present here should serve this purpose. We have made major revision for Section 3.2 (3). In the modified manuscript, we have clarified our idea/approach.

We acknowledge that our schematic diagram and relevant text do not emphasize this idea clearly as pointed out in the Reviewer's third minor comment. We have modified the schematic diagram (Figure 2), figure captions, and reorganize the "Section 3.2(3) Task 3" in the revised manuscript to make the idea clearer. Original Figure 2 includes both warm and cold years. In the revised version, to make the thing simple and less confusion for readers who are unfamiliar with LS4P, we only include cold year (same as the case in the LS4P Phase I) in the text and move the schematic diagram for the warm case to appendix for readers who want to pursue this issue further for their own research. In our response to the reviewer' 3rd minor comment below, we will have some more detailed explanation.

3. Forecast systems also produce a range of forecasted values through the running of ensembles, and any one ensemble member could represent what happens in nature. The experimental analysis protocol, however, emphasizes the importance of having the *ensemble mean* match the observed anomaly. This is inappropriate. The key question is, do any of the ensemble members look like the observations? (And, in conjunction with point #2 above, the truly key question is, do any of the ensemble member anomalies *relative to the forecast model's climatology* look like the observed anomaly?) A model cannot be considered wrong if one of its ensemble members looks like the observations. Insisting that an ensemble mean match a specific year's temperature seems wrong.

Response: In this paper, we do not emphasize the model intercomparison as well as each model's evaluation because the major focus for the LS4P is whether the LST /SUBT can provide S2S predictability through the multi-model efforts, which idea has never been tested before. Since many multi-model projects, such as CMIP, WAMME, and many others, find the ensemble means normally produce better results than any individual model's performance, we use the ensemble mean and the range of individual model results to assess whether there is S2S predictability and its uncertainty. The reviewer may have different opinion on this, but this is a common approach currently used in multi-model studies in the community, such as CMIP, AMIP, WAMME and endorsed by the LS4P modeling groups. In addition, the LS4P has more than 20 ESMs and many RCMs. A comprehensive analysis of each model performance is not that closely related to our main focus at this stage.

4. The model results concerning May Tibetan Plateau temperature anomalies and June precipitation anomalies in east Asia is perhaps suggestive but far from indicative of a causal relationship. Even if the agreements in 6a/6c and 6b/6d do suggest that one pattern led to the other (it could very well be coincidental), I don't see how the Tibetan Plateau in 6a/6c can be isolated as the source of the agreement in 6b/6d. Significant qualification of this figure's implications is needed.

Response. We fully agree with the reviewer's comments and apologize not to have presented our ideas more clearly. Figures 6a/6c and 6b/6d are NOT intended to present the causal relationship. These figures only intend to explain how we develop our hypothesis.

Observational results in Figure 6A, B (from various observational data sets) along with the remarkable consistency of modeling results in Figure 6C, D (compared to Fig. 6A, B) together provided the underpinnings for the **LS4P conjecture** that if the May land surface temperature anomaly on the Tibetan Plateau does contribute to the June East Asian precipitation anomaly, then improving the May land surface temperature simulation over the Tibetan Plateau through an improved initialization should allow Earth System models to better predict June East Asian precipitation.

The scientific development normally starts from scientists' curiosity based on some preliminary discoveries. The reviewer apparently is an expert in the Earth system modeling and should understand such consistency in Figure 6 from various observational data sets and various Earth system models (with very different dynamic processes and physical parameterizations) is not by chance, and is worth to **propose a hypothesis** then explore this issue further.

By and large, Figure 6 is an important step to motivate the hypothesis. Only Task 3 (with Task 1) is designed to prove the casual relationship. We have more explicitly emphasize this point in the revised paper to avoid confusion (Lines 349-353, 624-625).

Minor comments:

-- Just to clarify: Are the warm and cold years the same for each month shown? If not, it's not clear what this figure says about the persistence of warm and cold anomalies (line 92).

Response: Yes. The years are the same for each month. We have clarified this in the Figure 1 caption of the revised paper.

-- lines 106-108. This sounds strange. Can the authors clarify the link between temperature and water amount? While I see that more water leads to a slower seasonal transition, it's not obvious that at a single point in time, temperature tells you something about water amount.

Response: This part (lines 112-115 in the revised version) has been revised in the revision. The word linking temperature and water amount is indeed potentially confusing and has been eliminated.

-- I studied Figure 2 for a long time and still can't make sense of it. Why, for a cold year initialization, does the initial condition for a model with a cold bias get set to climatology whereas that with a warm bias does not? Also, please clarify in the caption: are the biases discussed here errors for the particular year of simulation, or are they long term climatological biases? I'm guessing the former, since Task 2 would need to be done for the latter; in that case, though, the use of the term "bias" is confusing here. Bias should refer to a long-term climatological error (reflecting a model deficiency), not to the error at a specific time (which should reflect both bias and random error). Overall, Figure 2 is not helpful for explaining the approach. And again, based on my earlier comments, I'm not convinced the Tmask strategy is appropriate anyway.

Response: This paper is for the LS4P experimental design and Figure 2 shows how we tried to generate the observed T-2m anomalies using the imposed mask in initialization then checked whether this anomaly improves the S2S predictability and leads to better prediction of the observed drought and flood events. Based on the reviewer's comments, we recognize that we did not explain the idea clearly in the previous figure, figure caption, and text. Although the researchers (co-authors of this paper) working in the LS4P know the idea from Figure, but probably not the readers who are unfamiliar with the project. We have performed several iterations within co-authors to improve the figure, figure caption, and presentation in the text. The section 3.2(3) is reorganized. We hope that the revised figure and text clear-up the confusion.

In the revised figure, we have clarified which initial temperature is for Task 1 and which is for Task 3 with more explanations and indicate that the LS4P phase I's goal is to prove the causal relationship between the Tibetan Plateau spring LST/SUBT anomaly and large-scale summer precipitation anomaly. Figure 2 and Equation 1 show how to produce observed surface temperature anomaly through Task 3 initialization (relative to Task 1's initial condition). We believe these should help readers to understand the ideas better in this figure. With the revised figure, the readers can see that when we use the difference between Task 3 and Task 1, we actually try to avoid the model systematic bias which was precisely suggested by the Reviewer in the main comments.

Regarding the bias, here we did not clearly indicate whether this should be a specific year or a climatology because it depends on case and data availability. "Bias" is not always associated with climatology. As Pan et al (2001, JGR) state in their analysis that "Both GCM and RCM fields can exhibit substantial systematic differences from gridded observational data. Such discrepancies between simulated and observed fields are commonly referred to as biases", although "some differences clearly are not biases in the strict sense, but for simplicity we use the term "bias" to refer to the entire set of comparisons". Such interchanges are also used in other studies/fields. A recent paper "Precipitation Biases in the ECMWF Integrated Forecasting System" by Lavers et al (2021) discusses using the IFS control forecast from 12 June 2019 to 11 June 2020 to show that in each of the boreal winter and summer half years, the IFS has an average global wet bias". The way they use bias is similar to what we use. In remote sensing, "bias correction" terminology is also commonly used. Moreover, as discussed in the paper, from Task 2, we know the climatological T-2m bias and year 2003 T-2m bias are very consistent. Therefore, we point out this terminology issue but still keep "bias' in our paper for simplicity as did in other current practices on lines 362-365.

-- Equation 1 appears to be a means to impose an artificially large temperature anomaly at the start of a simulation so that the anomaly is maintained realistically during the forecast. As far as I can see, there's no physical basis for the equation; it's fully empirical and could lead to initial temperatures that make little physical sense (e.g., colder than the model ever gets). More qualification is needed regarding how artificial this construct is. (And again, based on my earlier comments, it may not be appropriate to fix the temperature error in this way, since it may have a source other than the land model.)
Response: The LST/SUBT approach is a new development and is at a very early stage. The importance of memory of surface temperature has still not been fully recognized by the community. Currently, no ESMs, including reanalyses, are capable to reproduce observed high mountain T-2m anomaly. Developing adequate numerical methods and physical parameterizations to permanently solve the issue may take decades or longer. In meteorology, using the initialization scheme before we develop better models and better data sets to improve the prediction is a very common approach, as done by Yeh et

al. (1984, MWR) and Yang et al (1994 MWR). Those initialization strategies were always based on some empirical relations and not a strict physically-based approach. Especially in the early stages, some approaches are highly idealized. For instance, in Koster at al. (2004), which is a rather famous study, were used an approach for which a soil moisture value is artificially imposed for every time step. But that did not prevent that paper from receiving more than 2250 citations and from becoming a classical paper in meteorology.

On other hand, our approach is not that extreme. The reviewer thought we may impose an artificially large forcing because of a tuning parameter. In fact, this is not the case. In the follow-up paper in a Climate Dynamics special issue, we will show every model's-imposed forcing. in fact, they are not extreme. The LS4P includes most of major climate centers in the world. If our approach is totally different from their normal practices, they would not endorse the LS4P and participate in this project.

If we wait until the best dynamic and physical method are developed, nothing will happen.

Reference

Koster, R. D., P. A. Dirmeyer, Z. Guo, G. Bonan, E. Chan, P. Cox, C. T., Gordon, S. Kanae, E. Kowalczyk, D. Lawrence, P. Liu, C.-H. Lu, S. Malyshev, B. McAvaney,K. Mitchell, D. Mocko, T. Oki, K.. Oleson, A. Pitman, Y. C. Sud, C. M. Taylor, D. Verseghy, R.Vasic, Y. Xue, T. Yamada, 2004: Regions of strong coupling between soil moisture and precipitation. Science, 305, 1138-1140.

Lavers, D. A., S. Harrigan, and C. Prudhomme, 2021: Precipitation Biases in the ECMWF Integrated Forecasting System, JHM, 22, 1187-1198. DOI: https://doi.org/10.1175/JHM-D-20-0308.1

Pan, Z., J. H. Christensen, R. W. Arritt, W. J. Gutowski Jr., E. S. Takle, and F. Otieno, 2001: Evaluation of uncertainties in regional climate change simulations. J. Geophys. Res., 106, 17 735–17 751, https://doi.org/10.1029/2001JD900193.

Yang, R., M.J. Fennessy and J. Shukla, 1994: The influence of initial soil wetness on medium range surface weather forecasts, Mon. Wea. Rev., 122, 471-485.

YEH, T.-C., WETHERALD, R.T. and MANABE, S. (1984). The effect of soil moisture on the short-term climate and hydrology change—A numerical experiment. Mon.Wea.Rev. 112; 474-490

-- line 211 (and elsewhere): Replace "SST" with "ocean state", since SST is only a small part of what seasonal forecast systems rely on from the ocean. Arguably, subsurface ocean temperature distributions are more relevant.
Response: Per reviewer's suggestion, we have replaced "SST" with "ocean state" in most places. However, in the historical review part, since those studies really used SST for analyses, we still keep SST there. In addition, for some discussions on the AMIP type runs, we also keep SST.

-- Clarification regarding figure 7: is this the average of the 2003 anomalies relative to the different models' climatologies, or is it the average (over all years) model T-2m and precipitation minus the average (over all years) observations? I'm guessing the latter, given the remarkable agreement with figure 6c,d. The latter can truly be considered a bias,but the term bias was used differently elsewhere in the paper.

Response: Figure 7 shows the average (over all years) model T-2m and precipitation minus the average (over all years) observations. In the "Section 3.2 (2) Task 2" of the revised paper (lines 362-365), we have clearly indicated that we use the "bias" for both climatology and 2003 for simplicity as was done in Pan et al (2001) and Lavers et al. (2021).