Throughout this response, the reviewer's text is presented in black, our response in blue, and the proposed revisions in green.

This study deploys the Method of Morris to evaluate the sensitivity of release and storage time series to adjustments in the parameters of generic reservoir operating rules. The analysis is performed using a high resolution hydrological model (WBM) and focuses on a cascade of reservoirs located in the Upper Snake River Basin. The paper is well written, with clearly defined methodology and easy-to-follow results section. The experiments conducted are ill-suited to the aims of the study, leaving too many confounding factors. Specifically, I disagree with the authors key claim that the approach exposes coordination between reservoirs missing from the generic operating schemes (see specific comments below). While I do not doubt that coordination does occur in this particular cascade system, I cannot see how this sensitivity analysis exposes that coordination unequivocally.

We want to thank the reviewer for their thoughtful and constructive comments. We believe they provide a basis for us to clarify and be more explicit with the aims and assumptions of our work, both with textual explanations and supplementary data. The concern that our work does not distinguish when non-coordination is the most likely cause of the simulated drought and flood errors is also well-taken. In a revised version, we will show how coordinated operations would have avoided some (most) of the water extreme related impacts simulated in Sections 4.3 and 4.4, even in a model where (as in any hydrological model) there are other sources of error.

Our detailed response to the key concern expressed by the reviewer (that the errors seen in the paper may not be due to the absence of the coordination in the reservoir release rule) considers multiple aspects in this order:

1) Coordination is not represented in reservoir operation rules as presented in the current literature on large-scale hydrological models. This is not something this paper tries to demonstrate but a fact supported by the literature review made in the introduction.

2) Instead, this paper tries to diagnose some possible consequences of this non representation of coordination. Thus, our study is not focused on identifying the dominant parameters in the rule (for a reservoir in isolation, those can be deduced from the equations). Our Method of Morris analysis is a quantitative diagnostic for better understanding how the influence of the reservoir rules' parameters evolves both through time and through the multi-reservoir cascade. This diagnosis is helpful for addressing two key questions (i) does the assumption that reservoirs can be parameterized separately (made in studies that propose new and more sophisticated release rules) hold? And (ii) what are the potential consequences of this assumption for flood and drought assessments by increasingly detailed hydrological models in a changing world?

3) Whereas question (i) can be addressed by looking at simulation results alone (see Section 4.2), question (ii) is better examined by using a benchmark where we know for a fact that efficient multi-reservoir coordination has historically minimized flood and drought risk: the USRB. We provided evidence of that coordination when comparing historical and simulated storage and release trajectories. This is apparent when narrating the drought

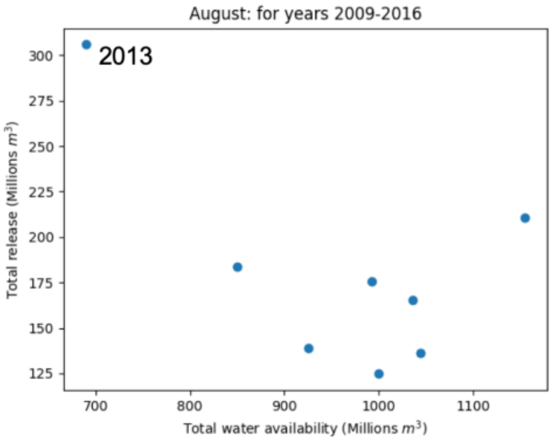
event in Section 4.3: p19 lines 2-4, 8-11 and 15-18, and when narrating the flood event in Section 4.4: p21 lines 25-34 and p23 lines 16-20. We however do agree with the reviewer that we did not provide direct evidence that coordination led to release behaviors that could not be accounted for with improved parameterization of the reservoir release rules. Please refer to point (A) below for a detailed explanation of how we propose to address this in the revision.

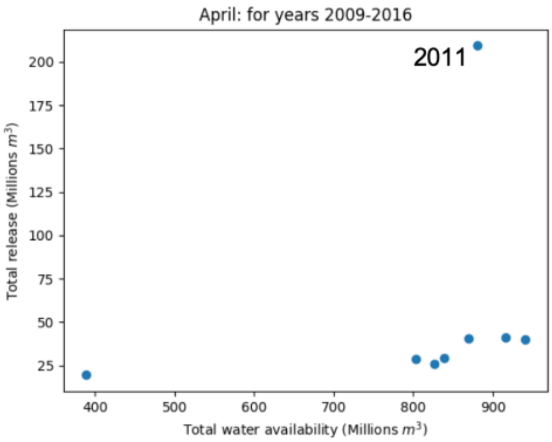
4) More specifically, our comparative analyses between historical and simulated release decisions in Sections 4.3 and 4.4 show qualitatively different behavior when contrasting historical data with simulation of the same period. We agree with the reviewer's concern that due to potential other errors in the model, our time-variant diagnosis approach does not in itself rule out that the observed differences may have little to do with the coordination issue. Please refer to point (B) below for a detailed explanation of how we propose to address this in the revision.

5) Please also note that as this is a diagnostic study, our goal is not to fully explain the errors or to correct them. There is to our knowledge no study of a real operational context using a hydrological modeling experiment where one would absolutely know every source of error or potential confounding factor. That being said, we are grateful to the reviewer for remarks that underlined that we need to clarify the presence or absence of confounding factors.

In a revised version, we will:

A) provide more quantitative classification and proof of coordination in the historical record, for both the high flow and low flow event. We will do so by taking the most upstream reservoir, Jackson Lake, and showing that local explanatory variables (storage and incoming monthly inflows) are not enough to explain observed releases. We will then insert and explain the following figures below. For the drought in August 2013 (Section 4.3; left panel) as for the flood in April 2011 (Section 4.4; right panel), the release response (y-axis) to storage + monthly inflows (that's total water availability, x-axis), is unusual. In the 2013 drought, the reservoir has already been emptied to rescue American Falls levels downstream but releases are still the highest in 8 years, whereas in the 2011 flood, releases are around 8 times as high whereas water availability is similar for 6 of the other 7 years. In both cases the release difference with other years cannot be explained by immediately available variables.





This proof of coordination will be inserted in Section 4.1 to add to the comparison of simulated and historical operations.

B) In Sections 4.3 and 4.4, we will demonstrate where observed historical coordination would have been enough to avert the simulated water extremes, notwithstanding other sources of error in the model (which naturally exist). To do this, we'll build a simple offline surrogate of the reservoir cascade and conduct the following 'what-if' experiments where we look at the consequences of storage dynamics based on combining observed releases (mirroring coordination) with simulated inflows (incorporating other sources of error in the model):

(i) for the drought event (Section 4.3) we'll apply the observed releases from Palisades assuming a long (7-day) delay between Palisades releases and American Falls inflows, and demonstrate that American Falls would not have emptied (i.e., lost its capacity to regulate irrigation) in that case. Similarly we'll do the same with Jackson Lake operations and examine how they avoid emptying Palisades in 2013. Note that the 7-day routing delay is a conservative assumption: if higher releases with long delays are enough for the water to reach the downstream reservoir in time, then the shorter actual delays would be enough as well.

(ii) for the flood event (Section 4.4) we'll focus on Jackson Lake first, and show how the observed release behavior would have avoided for the reservoir to lose its regulation capacity by becoming full, even with the simulated inflows. Then we'll translate this different release behavior to Palisades assuming a less than 1-day routing time (the conservative assumption here is that the flood event propagates very quickly) and demonstrate how observed releases would have avoided filling Palisades completely even with simulated inflows. We'll then translate the consequences of this to American Falls.

Please note that for both (i) and (ii) preliminary experiments have been conducted to guarantee that the results will be what we say they are; but we did not have time to produce clean and easy-to-follow figures for this discussion comment.

C) we will add a supplementary Section 3.1 to develop the methodology beyond explaining the rollout of the method of Morris, moving current Sections 3.1 and 3.2 to 3.2 and 3.3 respectively. This section 3.1 will detail:

1) that the analysis will rely on the comparison of coordinated observed operations with non-coordinated simulations (while explaining how we will look for a proof of this coordination)
2) the reasons for using the method of Morris, namely the identification of temporal "signatures", i.e. combinations of variables that are important at the same time, and the study of how these temporal "signatures" may travel through time and through the cascade, or get inverted
3) The supplementary online testing to demonstrate the impact of coordinated operations on avoiding simulated flood and droughts (as described in (B))

D) update the abstract to reflect the above points and clarify the methods and goals of the study.

I also have some concerns with the study design, and in particular the use of the high-resolution hydrological model, which introduces a severe and unnecessary computational constraint while also supplying the reservoirs with (likely) inaccurate historical inflows. The latter leaves the reader unsure as to whether the difference between simulated and observed storage/release is more a function of erroneous inflow than inaccurate operating decisions. This is particularly important when demonstrating the inadequacy of the reservoirs rules in representing flood and drought (it could simply be that the upstream hydrology from the model is delivering the wrong volumes of water). Since the actual observed inflows are available for this system, it would seem far more prudent to develop a simple, offline cascade model. I suspect such a radical change to the experimentation at this stage would be unrealistic, so I would instead encourage the authors to change the aims and storyline offered here. The simulations are sound and there is a great need for the community to learn more about the nature and performance of generic reservoir schemes. I would support a revision if either (a) the authors can convincingly rebut my concerns listed below, or (b) a new angle is developed with a more defensible conclusion.

Our introduction highlights that large scale hydrological models are being employed with increasingly ambitious assessment goals (p3 lines 2-14). These large-scale assessments of hydrological risks are not carried out with surrogates of full-scale models. In this regard, WBM and the reservoir rule system represent a state-of-the-art representative of this class of assessment model that is distinguished in its representation of human infrastructure systems. Our study has a clear and direct design: evaluating model outcomes in the same conditions as large-scale hydrological vulnerability assessments are carried out. This will be explicitly clarified in the methodology.
Moreover, it is non-trivial to connect complex large regional infrastructures to the broader natural components of the water balance. For instance, inflows to other reservoirs need to reflect the consequences of simulated upstream releases (including what that means for flow routing along the river course where there are lateral inflows). There is also the question of which other parts of the model should be included: for instance there are potential consequences for water withdrawals in case of low-flow conditions. This would have

required testing and arbitrarily excluding components of the water balance from the online surrogate to avoid making it overly complex.

This being said, we acknowledge that supplementary experiments involving offline surrogates would complement the main experiment (conducted with a full-scale model) to help clarify the contribution of coordination in averting impacts from water extremes. Therefore, we therefore thanks the reviewer for their suggestion here.

Specific comments to authors:

Section 3.2. The justification for the 10% decision is unconvincing. Suppose a dam has an average inflow of 100 cumecs. An Rmin of 0.1 would be 10 cumecs and would vary by a maximum of 1 cumec. So the left-hand side of the operating curve hardly moves. In contrast, if your Rmax is 5 (=500 cumecs) then you'd vary this parameter by plus or minus 50 cumecs. So the right hand side of the curve will shift wildly in comparison. This is surely why Rmin appears to be unimportant in your sensitivity analysis; you've barely moved it. I appreciate that there are physical reasons why Rmin would be expected to be somewhat less variable than Rmax in absolute terms (although both are highly uncertain), but the uniform 10% assumption is not ideal either, and it invalidates your later statements about which parameters constitute the signature of parametric influence (page 15).

This criticism of the chosen multipliers would be exactly on point if the goal of this sensitivity analysis was to ascertain which parameters in the reservoir rule are dominant. Instead (and as noted above), this analysis diagnoses how the parameter signature (i.e., the set of parameters that dominate and in which direction they influence outputs) evolves throughout the cascade and with time. The core contribution is not what single parameter controls are dominant, but that the dynamics in sensitivities are complex, highly dynamic, and non-separable. This is the core concern expressed in our Introduction, that these parametric behaviors are in direct contradiction to the current reservoir rule forms representation of reservoirs as being individually separable and independent parameterization problems without any information on coordination..

We acknowledge that the rationale for using the Method of Morris has not been explained with the requisite precision in the previous submission, so it will be explicit in Section 3 in a revised version.

Section 4.2 through 4.4. The results described here show that the operations of upstream dams can affect the decisions made at downstream dams. This occurs because any change upstream has an effect on the inflows into a downstream dam, thus affecting its storage levels and therefore its releases (which are a function of storage). Sensitivity analysis exposes how the decisions taken at one dam are affected by the rules deployed at other dams (often with some intriguing complexities). This is insufficient, however, to claim that the models must be missing *coordination* (which presumably means dam operator A looking at the storage levels of dams B and C to inform his or her decision). Similarly, the results that describe inadequate representation of flood and drought mitigation *could* be caused by uncoordinated reservoir schemes. But inadequate mitigation could also be simply because the operating schemes at individual dams are insufficient. It's possible that more realistic operating schemes at the individual dams (i.e., not necessarily incorporating coordination, but with more realistic structure and/or parameterization) would provide the correct mitigation responses. It's also possible that the failure of these models to represent flood and drought mitigation is partly caused by bias in your inflow data (which is not shown or compared to observed at the upstream dam). Given these possibilities, the proposed framework fails to demonstrate unequivocally that a missing piece of the reservoir model is coordination.

This comment echoes the reviewer's main concern above and further illustrates it. We thank the reviewer for taking the time to formulate it. It has been helpful in formulating our response to the reviewer's main concern.

Other minor comments:

Figure 1. Difficult to interpret due to color scheme used.

We agree and will amend the figure to improve the color scheme

Table 1. Please provide some additional justification for these parameters. Has there been any validation done?

The general form of the reservoir rule was first presented by Proussevitch et al. (2013) and validated using the GRanD database (Lehner and Liemann, 2011). Variants of this rule have been used with a daily time step on the Niger river basin (Oyerinde et al., 2016), and with large-scale assessments using WBM (Grogan et al, 2015; 2017; Zaveri et al, 2016; Liu et al., 2017). The fine-tuning of the parameters when establishing this version of the rule was made using a set of 22 large North-American and Eurasian reservoirs in offline mode., including the two largest reservoirs in the USRB (Palisades and American Falls, daily release NSE 0.70 and 0.60 respectively). Similar to what happens when a reservoir rule that classifies reservoirs by purpose is used in a large-scale model, we did not fine-tune the rule to each reservoir. This allows us to use the reservoir rule in conditions that are similar to what is done in most state-of-the-art hydrological models.

We will clarify the above in Section 2.4: the reviewer is right to point out that readers should know where the figures in Table 1 come from.

Why would the Rref for a water supply reservoir be 0.1? This parameter should vary widely across water supply reservoirs as a function of demand relative to inflow (and 0.1 is very low).

Thanks for this question that warrants a clarification. In this work, "water supply" means that water is meant to be withdrawn directly from the reservoir, instead of being diverted downstream after release. Therefore, the reservoir must remain as full as possible.

We will provide this clarification in a revised version.

Generally I don’t see why the purpose of the reservoir would control these parameters. Rmin (and indeed the whole left-hand section of the operating curve) would likely be strongly determined by environmental flow requirements, which are independent of reservoir purpose. I would also suggest discussing the omission of seasonality in the operational parameters (which could apply to non-irrigation reservoirs). If there is no good justification for these rules then this is ok (it’s not the purpose of your experiment to defend the status quo)âˇAˇ Tin this instance just make a statement to inform the reader.

The idea that reservoir purpose controls release rule parameters is reproduced here because it is ingrained in the literature (see p2 lines 23-35 and p7 lines 11-14). We agree with the reviewer that no parametrization is perfect, be it in our paper or in the wider literature. However, and as the reviewer is right to point out, our goal is not improve existing parameterizations. Instead, it is to diagnose them using a reservoir rule with several state-of-the-art characteristics.

P6 line 1 - reservoir data *were* derived…
P6 line 1 - not clear who performed these manual updates

We will correct the typo and specify that authors made the manual updates

P6 line 7 - How accurate are these demand data compared to USGS estimates, which are US specific.

The GPW is based on US census data, and is processed to a gridded resolution on par with our model grid, whereas USGS estimates are county level. We verified the difference between the two demand estimates and found it to be below 50%, in a basin where industrial and domestic demands are two orders of magnitude smaller than irrigation demands. Therefore, we deemed the GPW data to be acceptable for this study.

Section 2.1. I hadn’t studied the Method of Morris previously, but this is an excellent description that educates the reader.
We appreciate this kind comment.

P16 line 7 - The high variability in release (relative to observed) is surely just caused by the right-hand side slope of the operating curve being too steep, or the Sref being too low. Why should this point to incorrect representation of coordination?

We are not sure whose reservoir’s Sref or right-hand slope the reviewer is referring to. We are also not sure what they mean by “incorrect”: in Section 4.2 we do not assess whether the representation is correct. Instead, as stated clearly at the top of the Section (p 15 lines 32-33), “Our analysis in Figure 6 focuses on flows from a single year to clarify the complex interactions between upstream releases and Minidoka operations”

In other words, in Section 4.2 and Figure 6 we use results from the method of Morris to demonstrate that release rule parameters upstream influence inflows, storage and release decision at Minidoka. In other words, we show that the assumption that this reservoir can be parameterized separately from the others does not hold here.

This means that the reservoirs should be parameterized jointly. And that would account for coordination. We will insert a sentence in Section 4.2 to make this point explicit.

P16 line 27 - this may have been a low flow event, but your figure for Jackson shows clearly that the drawdown was in large part caused by sustained high release through 2013.

There is no contradiction between the two statements. In fact, as can be shown in the Figure in this document (left panel), there was little water and high releases. This is explained by coordination with downstream reservoirs.

We will insert that figure to clarify this.

It would be helpful to see inflow, storage, and release graphs for the full time series for all reservoirs (perhaps in supplemental material).

We are happy to present this data in supplementary material, using similar graphs as in the paper.