We would like to express our deep appreciation to the two reviewers (Dr. Doherty and the anonymous reviewer) for the thoughtful comments and insightful suggestions. Working to resolve these comments helps us add lots of interesting science and add value to the manuscript.

The primary objective of this study, as pointed out by Reviewer #2, is to build a bridge for linking the parallel PEST and WRF-hydro on the basis of HPC clusters and explore the computational benefits of this bridge. We do not attempt to extensively assess each individual tool or address questions in each individual domain, such as optimizing the objective functions in PEST or calibrating WRF-Hydro to achieve the best set of model parameters. However, we appreciate the opportunity every much during the revision of this manuscript by learning more about PEST especially the method of regularization for calibrating environmental models. We are also very glad that the reviewers found the bridge we built useful for helping WRF-Hydro users with the long and tedious model calibration.

In the revised version of this manuscript, several major changes are made based on both reviewers' comments/suggestions. They are listed below:

1. We re-do the WRF-Hydro calibration using SVD-based regularization method in PEST.
2. We consider prior information for the calibrated parameters.
3. We also consider different weight for the stations that are calibrated, based on their inversed mean of discharges.
4. To test the computational benefits of the bridge, we design five experiments by assigning different amount of computing resource for parallel PEST and for parallel WRF-hydro.
5. To constrain the problem size due to the limits of computing resource, we reduce the number of calibrated parameters to 22 according to the model sensitiveness of this particular study.

Please find our one-on-one response below to each reviewer's comment. A complete list of the changes made for the revised manuscript can be found in the "track changes" version of the manuscript. A clean version of the revised manuscript is also attached at the end.

Sincerely,
Jiali Wang
jialiwang@anl.gov

Reviewer #1

Interactive comment on "A parallel workflow implementation for PEST version 13.6 in high-performance computing for WRF-Hydro version 5.0: a case study over the Midwestern United States" by Jiali Wang et al.

Doherty (Referee)
johndoherty@ozemail.com.au
Received and published: 6 December 2018

The authors describe modifications that they made to PEST to enhance its use on a HPC. They then describe use of their modified PEST in calibration of a complex surface water model. While I found the paper interesting, I found that it was lacking in information in some respects. For example nothing is said about the interface that they constructed between parallel PEST and the run management software that they employed.

Response:

Thank you for your comment. The interface is the most important thing we built in this study, and testing the operational feasibility and computational benefit of this interface are the main objectives of this manuscript. Hence it definitely should be described as you suggested. We add this paragraph in **Section 3.2 PEST files and settings**:

"The interface we have built between parallel PEST and the management software (SLURM here) is, in general, used for (1) setting the number of workers and the nodes for each worker to conduct a model run (WRF-Hydro here); (2) finding the nodes that are available; (3) setting up the working directory for the workers; (4) identifying the nodes that work for each worker; (5) passing the global files (same for all the working directory) to all the workers (these files include the lookup table files that are not to be calibrated, the namelist files for both LSM and hydrological sector, and restart files that generated by the previous simulations, or spin-up period); and (6) submitting the job for the entire calibration process, including parallel PEST and parallel WRF-hydro. This job can be submitted as a cold-start run or as a restart. The main difference for this interface on different management software is that different management software has its own way to submit jobs and identify available nodes. This difference requires some changes in the script we developed."

Nor was any reference made to PEST settings.

Response:

Thanks for the comment. In our original version of manuscript, we used estimation mode for PEST, and considered equal weight for all four calibrated stations. There was no singular value decomposition (SVD) nor regularization used.

In our revised manuscript, we conduct the calibration using SVD-based regularization, we assign prior information for all the calibrated parameters, and we also consider different weights for the calibrated stations. We add the PEST setting in **Section 3.2 PEST files and settings**:

"This study presents calibration results from PEST using the SVD-based regularization in regularization mode to ensure numerical stability (Tonkin and Doherty, 2005). We focus on calibrating 22 parameters (see Table 1 and detail description in Sec. 3.3) using 96 observation points and 22 items of prior information for the calibrated parameters. In each item of prior information, a value equal to its default value provided by the WRF-Hydro v5.0 (or the log of its default value) is assigned for each adjustable parameter, assuming

that default values are the preferred values. All prior information equations are assigned a weight of 1.0. We assigned five different regularization groups to the prior information: Manning's roughness coefficients specified by Strahler stream order in CHANPARM.TBL to one group; the parameters in HYDRO.TBL (Manning's roughness coefficients for overland flow as a function of vegetation types) to another group; and three global parameters for the Noah-MP (xslop1, refdk, and refkdt) in GENPARM.TBL to the remaining three groups. The 96 observation points are given different weights based on the inversed mean of their observed discharge during the studied period (see the detailed description in Sec. 3.3 and Sec. 4.1). For a detailed description of these settings see the PEST User Manual (Doherty, 2015)."

While I agree with the authors that use of inversion methods that can parallelize model runs and handle the estimation of many parameters employed by a complex model is a much-needed addition to the arsenal of surface water modelling, I think that many more advances could be made than the authors have made. In particular, there was no mention of the use of Tikhonov regularization to accommodate parameter nonuniqueness at the same time as it promulgates uniqueness through obtaining a set of parameters that "make sense" from an expert knowledge point of view. This, I think, is one of the strongest arguments for use of gradient-based, highly parameterized methods in regional surface or land use model calibration, that is the ability to not just accommodate nonuniqueness, but to turn the "wiggle room" engendered by nonuniqueness into formulation of an inverse problem that can actually make regionalization and transportability of parameters a reality.

Response:

Thanks for your comment, and we understand this is one of the major concerns for how PEST was used in our original manuscript to calibration a hydrological model and more broadly, environmental models. During the revision of this study, we conduct all the WRF-hydro calibration using SVD-based regularization. We also use prior information for the parameters that we calibrated, as you can see from our previous response about PEST settings. In **Section 5 Summary and discussion**, we also commented that, "In this study, we consider using the prior or regularization information only for the parameters that we calibrate. As is the case with solving inverse problems, prior information is added to improve the smoothness of the solutions. In order to build a more comprehensive calibration, an important aspect that can be considered is to enrich the prior with the available historical data. For example, in this particular case, one can use the historical observation data (e.g., April and May from the past few years) to enrich the prior information for the parameters. Hence, the regularization objective function in PEST will constitute not only the discrepancies between parameters and their "current estimates" but also the discrepancies between WRF-Hydro simulations and preferred values (which is the observed time series of historical discharge). Additionally, one can use the pilot points technique described by Doherty (2005) in conjunction with parameter estimation to add more flexibility to the calibration process. This will be potentially beneficial in improving the predictions"

In addition, to emphasize the importance of regularization for hydrological and environmental model calibration, we mentioned regularization in introduction: "PEST has four modes of operation. One of the modes is regularization mode, which supports the use of Tikhonov regularization and is found better for serving environmental models, because if implemented properly, it supports model predictions of minimum error variance, it is numerically stable and it embraces rather than eschews the heterogeneity of natural systems. Singular value decomposition (SVD) can be used as a regularization device to guarantee numerical stability of the calibration problem".

We also find that regularization mode does generate a better hydrograph shape compared to using estimation mode:

In **Section 4.1 WRF-Hydro calibration and validation**, we mentioned "Compared with the results of calibration using the estimation mode (no regularization) in PEST, the SVD-based regularization generates slightly better hydrograph shape with 24-hour later discharge peaks that are closer to the observations."

In **Section 4.3 Evaluation of spatial transferability of the calibrated parameters**, we mentioned "We also find that using the SVD-based regularization for the PEST calibration captures the timing of discharge peak better than using the estimation mode, which is one-day earlier than the observations reaching the discharge peak."

The authors use a simple objective function. This may be ok for some inverse problems. However as they point out, some of the smaller flows (in terms of location in space and location in a single flow time series) are not as well fitted as they could be. Perhaps weights should be a function of flow – and of location. Perhaps other important aspects of the flow time series should be made more visible to PEST through formulation of separate, targetted objective function components to ensure that these aspects of the time series are also well fit.

Response:

We understand this is another major concern about how PEST was setup for this study. We agree that, due to the fact that the calibrated station we chose are on different size in terms of water volume, to better handle the smaller river station (Station 1), considering different weight is fundamental, as also pointed out by Reviewer #2.

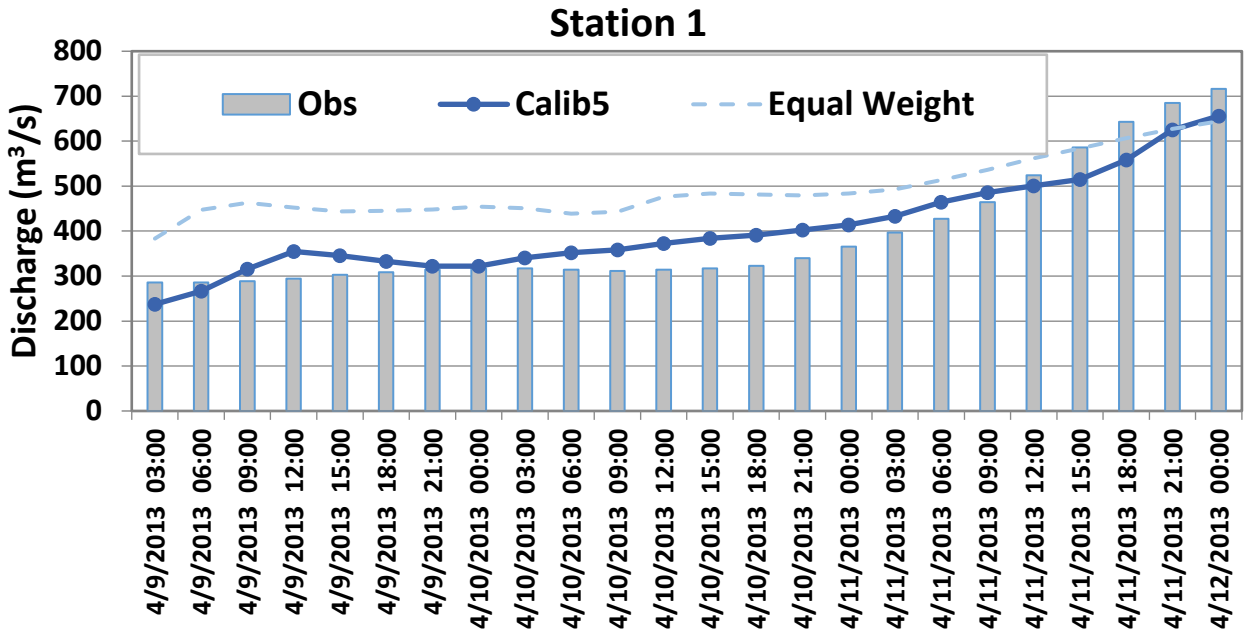
Therefore, during the revision of this study, we conduct all the PEST calibration of WRF-hydro considering a higher weight for Station 1 than for the other three stations. We add description about weight in the revised manuscript in **Section 3.2 PEST files and settings**, as you can find from previous response. We also add description in **Section 3.3 Calibration experiments:**

"We find that by using the same absolute weight for all four stations, the calibration helps three stations (Station 2, 3, and 4) with large water volumes to generate more reasonable results than do the default parameters; however, the results for Station 1, which has a relatively small volume of water, is not always better than the discharge that is modeled by using default parameters. Thus, we assign a weight of 15.0 for Station 1 versus a weight of 1.0 for the other three stations according to the inversed mean of observed discharge over these four stations in April 2013. The ratio of the weights between Station 1 and the other three stations stays similar even if the means are calculated based on different time periods."

We describe the results in **Section 4.1 WRF-Hydro calibration and validation** by comparing PEST calibration using equal weight and a higher weight for Station 1:

"Compared with the results calibrated by using equal weights for all the stations, by giving a higher weight to Station 1 the model bias over Station 1 is significantly reduced, with a higher NSE (0.87 with higher weight versus 0.14 with equal weight) and lower RMSE (48.1 versus 123.6)."

Here is a figure (Figure I) showing how the weight helps the result of Station 1. "Calib5" represent the results using higher weight for Station 1, while "Equal Weight" the results using equal weight for all the four stations.



The authors make a big deal out of their modifications to parallel PEST so that it is HPC-friendly. Actually, I think that the BEOPEST version of PEST has similar capabilities. The original version of BEOPEST used both MPI and TCP/IP for communication between master and slaves (now called manager and workers). Now only TCP/IP is used. One of the reasons that BEOPEST's capabilities exceed those of parallel PEST in the HPC environment (actually on any network) is that the manager does not need to write model input files and read model output files across the network. This makes run management must faster, more secure, and able to take place in a greater variety of network environments.

Response:

Thanks for the comment. Since our major objective is to build the bridge (or interface) between the parallel PEST and WRF-Hydro on the basis of HPCs, we expect the interface still working if one wants to use BEOPEST instead of parallel PEST to calibrated WRF-hydro, assuming there is also a version of BEOPEST in Linux environment (this is not clear to us by reading the manual). The only main changes for our script would be copying the template files and instruction file into each working directory. The command to execute BEOPEST is also different, which is *beopest* instead of *ppest* for parallel PEST. We add this paragraph in **Section 5 Summary and discussion**:

"While this study ported the parallel PEST to HPC system and linked it to WRF-Hydro, we note that BEOPEST is available in the PEST family. BEOPEST has the same functionality as parallel PEST but uses a different approach for communication between master and workers. Working with HPC-enabled BEOPEST may save total time cost since BEOPEST uses the Transmission Control Protocol (TCP) and the Internet Protocol (IP) instead of message files (reading input and writing output between master and works) for communication. We expect it to be relatively straightforward to use BEOPEST to calibrate WRF-hydro on HPCs since the interface remains similar, except one needs to copy the template and instruction files in addition to the global files (see Section 3.1) into each working folder."

In summary, I think that what the authors have done is good. However I also think that the potential for regional surface water model calibration and uncertainty analysis in a HPC environment still remains

largely untapped. Some of this potential will be realized with use of singular value decomposition to ensure numerical stability when inverse problems are ill-posed, use of Tikhonov regularisation to ensure parameter sensibility and transportability under the same conditions, and more creative formulation of the objective function than the authors have done.

Response:

We thank Reviewer #1 again for all your valuable comments. Although optimizing the objective functions in PEST is beyond the scope of this study, we do have some thoughts for future studies. We add this in **Section 5 Summary and discussion**:

"In this study, we consider using the prior or regularization information only for the parameters that we calibrate. As is the case with solving inverse problems, prior information is added to improve the smoothness of the solutions. In order to build a more comprehensive calibration, an important aspect that can be considered is to enrich the prior with the available historical data. For example, in this particular case, one can use the historical observation data (e.g., April and May from the past few years) to enrich the prior information for the parameters. Hence, the regularization objective function in PEST will constitute not only the discrepancies between parameters and their "current estimates" but also the discrepancies between WRF-Hydro simulations and preferred values (which is the observed time series of historical discharge). Additionally, one can use the pilot points technique described by Doherty (2005) in conjunction with parameter estimation to add more flexibility to the calibration process. This will be potentially beneficial in improving the predictions."

Reviewer #2

Interactive comment on "A parallel workflow implementation for PEST version 13.6 in high-performance computing for WRF-Hydro version 5.0: a case study over the Midwestern United States" by Jiali Wang et al. Anonymous Referee #2

Received and published: 16 January 2019

The paper of Wang et al. deals with a potentially interesting implementation of the parallel version of the PEST software. PEST is a powerful and very useful tool for hydrologists, helping them during long and "exhausting" calibration sessions. Therefore, introducing the portability of parallel PEST to HPCs is good news and, specifically for the present paper, the main theme to highlight.

Response:

We are so glad that the reviewer finds this study helpful for hydrologist with the most tedious part of model development and application — calibration. We also wish to thank the reviewer for your great insight about the highlight of this study, which add important value to this manuscript.

Nevertheless, in my opinion the way the paper is structured mainly highlights, instead of the advantages of the novelty, the performances of the PEST calibration, which is something widely and well assessed by the hydrology research community. Almost all figures and tables deal with PEST results. Furthermore, the calibration procedure presented is questionable from different points of view (some of which are exposed later).

Response:

Thanks and we agree your comment. In our revised manuscript, we delete the section of 7-day calibration (figures and tables) considering it is unnecessary to support our main objective. We also delete or shorten the descriptions about WRF-Hydro and PEST, which readers can easily learn from the User Guides/Manuals. We thus focus on only two points: one is to show the operational feasibility, and the other is to explore the computational benefits of the HPC-enabled parallel PEST linked to WRF-Hydro. To demonstrate the first point, we calibrate WRF-Hydro for 3 days using SVD-based regularization method in PEST, and considering different weight for the four calibrated stations. For the second point, we design new experiments, using different computing resources for PEST workers and for WRF-hydro. Details can be found below in our response to your later comments.

The most interesting/innovative Section of the paper is Section 5.1, but the analysis of scale-up capabilities should be described with much more detail. Concerning the main outcomes of the paper highlighted in the summary, points from 2 to 5 are quite obvious (they deal with the recognized skills of the PEST software), while point 1 should be expanded: what does a factor of 30 "with respect to a serial calibration" exactly mean? In my opinion it's not a rigorous statement. What do the authors exactly mean with "serial"? Even though PEST calibration is serial, WRF-Hydro can run in a parallel fashion, and the speed of the calibration process would depend on the number of nodes used for the hydrological simulation. A possible idea is to provide hints about the trade-off between the number of nodes/CPUs used for running the parallel model (i.e., WRF-Hydro in this case) and the number of nodes/CPUs used for running PEST in a parallel fashion. I guess it depends somehow also on the dimensions of the domain (and no information is given here about the number of cells in which the basin is discretized, so the reader has no idea about the actual computational burden).

Response:

Thanks for your comment and suggestions. We re-write **Section 5 Summery and discussion** to only summarize the findings of the study, and then raise some key points that beyond the scope of this study but may inform future studies. In other words, we delete majority of the summary that appears in points 2 to 5 in our original manuscript.

In the revised manuscript we expand the discussion of scale-up capabilities of parallel PEST linked to WRF-hydro by designing more experiments using different computing resource for PEST workers and for WRF-hydro. We add **Section 4.2 computational benefits of parallel PEST on HPCs**. Some key notes about the experiments and our findings are quoted below:

"In this study we test the computational performance of HPC-enabled parallel PEST using different number of workers (6, 12, and 23) for the 22-parameter calibration. As shown in Table 3, we conducted five experiments: Test 1 uses 23 workers, Test 2 uses 12 workers, and Test 3 uses 6 workers. All three tests use two nodes for each worker to run WRF-Hydro in parallel."

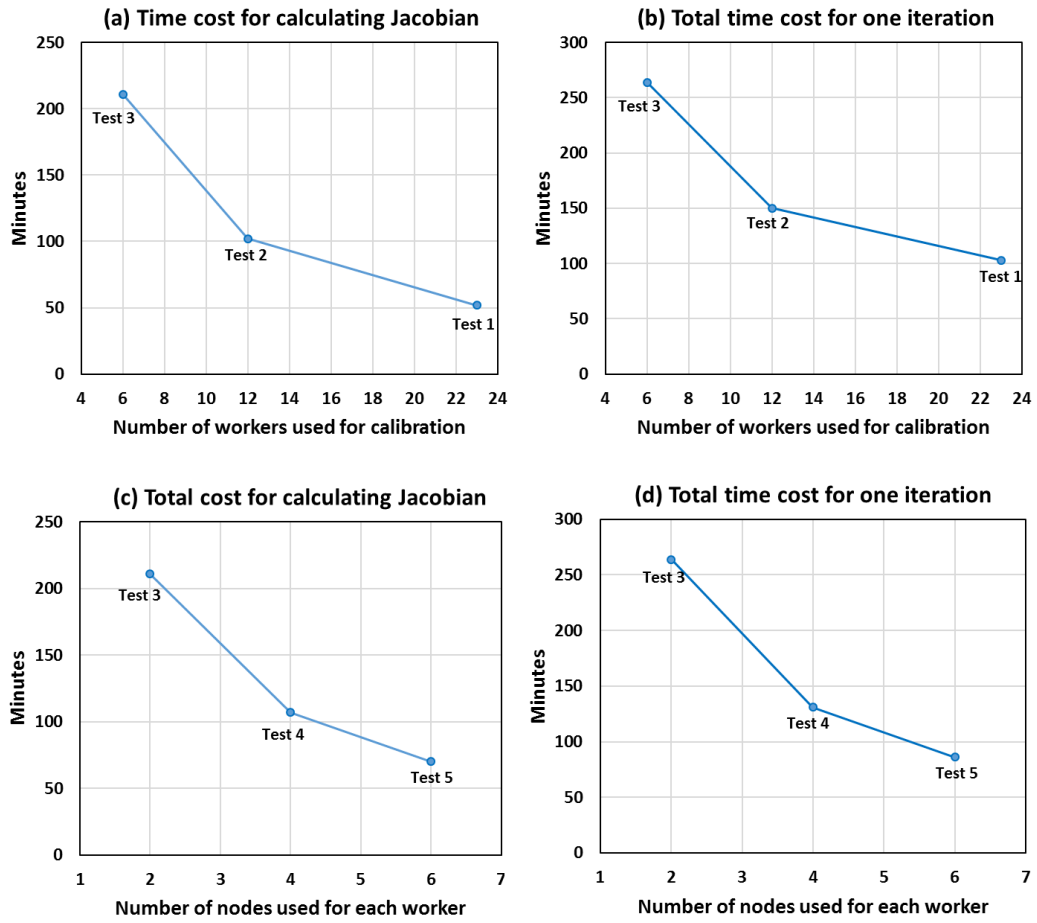
"In order to test the trade-offs between the computing nodes used for running parallel WRF-Hydro and the workers used for running parallel PEST, Tests 4 and 5 use different number of nodes for each worker to run WRF-Hydro in parallel. Explicitly, Test 4 uses four nodes per worker, and Test 5 uses six nodes per worker. Both tests use six workers for running the parallel PEST."

"when we assign 12 (and 6) workers to parallel PEST, the time cost for calculating the Jacobian matrix is increased by a factor of 2 (and 4) compared with the time cost of using 23 workers. The time cost for the parameter upgrade stays similar for the three experiments because only one cycle of WRF-hydro simulation is conducted to test the Marquardt lambdas. As a result, the total time cost for Test 2 is ~1.5 times more than that for Test 1, and the total time cost for Test 3 is ~1.5 times more than that for Test 2 (Fig. 4b). By extrapolating the speedup curve shown in Fig. 4a and Fig. 4b, we expect the total time cost to be ~1516 minutes when using only one worker (or sequential mode), which is about 15 times slower compared with running the PEST in parallel mode using 23 workers."

"On the other hand, for the same case study and using the same number of nodes for running parallel WRF-Hydro, we can estimate the computing speedup by assuming an increase in the number of calibrated parameters to 50. This would be the case, for example, to evaluate model sensitiveness to the physics in Noah-MP or the spatial variabilities of certain parameters. We then expect to use 51 workers to achieve the best computing performance for parallel PEST. This would then be 28–30 times faster than running PEST using one worker (or in sequential mode). Similarly, if 100 parameters were used for the calibration for the same case study, a factor of up to 60 speedup in the calibration process would be achieved by running HPC-enabled parallel PEST."

"In addition, by increasing the number of nodes for each worker to conduct WRF-Hydro (Tests 3, 4, and 5), the time cost for the entire calibration process is significantly reduced (Figs. 4c and 4d). Specifically, the WRF-hydro scales up well when using four and six nodes compared with using two nodes per worker for running the WRF-Hydro. Both the time spent on calculating the Jacobian matrix and the time spent on testing the parameter upgrades are decreased by 49% and 67%, respectively, when using four and six nodes. Therefore, the total time spent is also decreased when using more nodes for each worker (see Table 3). Increasing the number of nodes to eight for each worker will most likely further decrease the time cost by 70–75% compared with using only two nodes per worker. Moreover, if one has a larger study area such as the entire contiguous United States, we expect the WRF-Hydro to have an even better scale-up capability (e.g., on dozens of nodes) than this study."

Here is a figure to show the computational benefit using parallel PEST to calibrate WRF-hydro.



Another important point, that should be better discussed, is the missed capability of the implemented version of PEST to deal with the calibration of spatially distributed parameters. This is important because it's reasonable to expect parallel PEST executions with WRF-Hydro for wide domains, and wide domains often need spatial differentiation of spatially distributed parameters, like, e.g., OVROUGHRTFAC, RETDEPRTFAC or other spatially distributed parameters available with WRF-Hydro v5.0.

Response:

We do acknowledge the importance of regionalization of parameter calibration, which definitely deserves future studies especially for large domains. We have been trying to add the interface on top of what we have now, to consider the spatial distributed files/parameters. For example, one can add regional OVROUGHRTFACs (e.g., their lower/upper bounds, and default values) in the PEST control file based on catchments/basins/regular regions etc. the potential challenge is that, the selection of the locations and sizes of catchment may introduce significant uncertainties to the calibration results. Thus it requires systematic and comprehensive investigation and understanding of the study area. We add a paragraph in **Section 5 Summary and discussion** about this:

"We only calibrate the parameters in lookup tables. Using a single value to represent a physics may work for a small domain but could be problematic for a large domain, especially we expect the HPC-enabled parallel PEST to execute with WRF-Hydro for large domains, which often need parameter regionalization. For example in WRF-Hydro v5, there are many spatially distributed parameters available such as the overland flow roughness scaling factor (OVROUGHRTFAC), the factor of maximum retention depth (RETDEPRTFAC), and the soil related parameters (when compiled with SPATIAL_SOIL=1). Calibrating these spatial parameters based on grid scale (e.g., catchments) rather than a single value will give the model more flexibility and thus can better fit the observations (Wagener and Wheater, 2006; Hundecha and Bardossy, 2004). In practice, for example, one can include regional OVROUGHRTFACs (e.g., their lower/upper bounds, and default values) in the PEST control file based on catchments. However, the selection of the locations and sizes of catchment may introduce significant uncertainties to the calibration results, which requires systematic and comprehensive investigation and understanding of the study area."

By the way, another limitation is that, at least as I understand, the calibration is available only against observed streamflow. Of course, this is the first option but not the unique one (one can decide to calibrate also against, e.g., soil moisture or latent heat flux data).

Response:

For the calibration exercise we did in this study, we use frxst_pts_out.txt as an instruction file which serves the format of output files of each working directory. In this file there is only discharge and water level available, so we calibrate the model using discharge data. It is feasible, however, to calibrate other variables as long as the observation data is available. For example, one can either find the closest point from the gridded dataset to the observation location and then compare that point to observations; or one can change the WRF-Hydro I/O code to output other variables in the frxst_pts_out.txt file, so they can still use the same interface we build here to calibrate other variables in addition to the discharge. We have added this regard in **Section 5 Summary and discussion**.

Finally, another important point is to (at least) discuss the problem of equifinality, which is incidentally (but not explicitly) dealt with in P11 L29 – P12 L5.

Response:

Thanks for your comment. It's actually interesting to think about this together with your other comment that one can also calibrate other variables rather than discharge. Since equifinality is an important source of model uncertainty, to reduce the model uncertainty, one may calibrate the model using multiple variables instead of one variables. This way the calibrate can constrain the model drift and may reduce the model uncertainty of prediction of certain variables (e.g. discharge and soil moisture). We add a paragraph in the **Section 5 Summary and discussion**:

"The optimal parameter set obtained from this study is from the 5th iteration of parallel PEST by testing five Marquardt lambdas. Testing different number of lambdas or calibrating different number of parameters may generate a different set of optimal parameters. These parameter sets can all make physical sense and be equally good for reproducing observed discharges. This problem is named equifinality (Beven and Freer, 2001; Savenije, 2001), which is an important source of model uncertainty. To reduce the model uncertainty through reducing the equifinality, hydrologists carry out additional modelling objective for model evaluation to find more useful parameter sets (Mo and Beven, 2004; Gallart et al., 2007). Alternatively, inspired by No. 3 discussed above, one can calibrate the WRF-hydro model based on more than one variables, such as discharge and soil moisture (or heat flux or water table depth) to reduce the number of optimal parameter sets, and thus reduce the model uncertainty of predictions for these variables."

Summarizing, though I acknowledge that the research presented is potentially interesting and innovative, I suggest to re-think the paper highlighting much more the computational benefits provided and reviewing the calibration performed in the case study.

Response:

We thank Reviewer #2 again for all your great insights. Our responses and revisions can be found above and below, as well as in the revised manuscript with tracked changes.

Following, a (not comprehensive) list of doubts regarding the calibration procedure and other minor comments and typos. I hope my comments can help improving the research.

Doubts about the calibration procedure:

Even though I acknowledge that authors decided to "focus less on extensively assessing the performance of the WRF-Hydro model", several aspects of the calibration procedure are very questionable.

1. no information about spin-up. This is extremely important, especially for such a short range calibration (only few days). The model should be run in advance (at least one month, I would say) in order to let several variables (e.g., moisture fields) have a realistic spatial distribution.

Response:

We did run the model for 3 months for spin-up. We apologize for not including it in the model description. Here is what we add in the revised manuscript:

"We start the WRF-Hydro simulation on Jan 1 2013 and run the model for more than 3 months to reach equilibrium. This 3-month period is considered as spin-up time and is excluded from model calibration and evaluation. We calibrate the river discharge calculated by the WRF-Hydro model from 00UTC April 9 to 00UTC April 12 2013, considering it is long enough to achieve our objective. We then evaluate the model performance against U.S. Geological Survey (USGS) observed river discharge from 00UTC April 12 to 00UTC 25, 2013."

2. the authors state that: April 8-11 moderate rain, April 12-14 no rain, April 15-18 rain, peak flow April 19. 3-day calibration is: April 9-11 (to be precise, April 12 at midnight), then validation is April 13-23 (April 12 is missed). 7-days calibration is April 9-15, validation is April 17-23. To me, it does not make too much sense that 4 more days are added when only the last one is rainy. It would be much better to calibrate the model with respect to a previous flood event, as it is usual. After all, observing graphs in figures 3 and 5 one after another just shows that increasing the number of days used for calibration improves the performances (but this is rather obvious), even though not yet enough.

Response:

As we mentioned earlier, in the revised manuscript we only focus on model calibration during April 9-11, and validation from April 12-24. We delete the 7-day calibration/validation results considering it is not necessary nor helpful to demonstrate our main objective. We add these sentences to emphasize this regard:

"The primary objective of this study is to build a bridge for linking the parallel PEST and WRF-hydro on the basis of HPC clusters and to explore the computational benefits of this bridge. We do not attempt to

extensively assess each individual tool or address questions in each individual domain, such as optimizing the objective functions in PEST or calibrating WRF-Hydro for a long time period considering all the relevant parameters to achieve an optimal parameter set. The calibration period thus is limited to only three days, which we believe long enough to achieve our objective and to understand WRF-Hydro's sensitivity to the calibrated parameters."

We also add text in **4.1 WRF-Hydro calibration and validation** to explain the reason for the bias in hydrograph shape, such as the early peak and the faster decrease of river discharge:

"For Station 1, the WRF-Hydro almost captures the timing of the peak of discharge, although it still underestimates the water volume by ~25%. The reason is that this study uses a direct pass-through baseflow module, which does not account for slow discharge and long-term storage of the baseflow. Therefore, the largest contribution to river discharge is from precipitation, and groundwater does not contribute much discharge to the channels in a long-term view, as is also true for the other three large river stations. Different from Station 1, for the other three large river stations, the WRF-Hydro modeled discharge increases soon after the peak of precipitation and reaches a peak on April 21, 2013, which is much earlier than the observed peak of river discharge (near April 24). The reason is that the water contributions for these stations are from a larger river basin (Mississippi River) than we included in our current study area. Thus, when a heavy precipitation event occurs over the entire river basin, there will be a significant lag time (especially at the lower part of the basin) between the peak of precipitation amount and the peak of river discharge. For example, the precipitation over the upper part of Mississippi River Basin (MRB) has a peak amount on April 18–19, but the river discharge did not reach its peak until April 24. Because our studied area covers only half of the MRB, the modeled river discharge has a shorter delay period after the peak of precipitation than does the observed river discharge. Enlarging the study area to include the entire MRB may improve this situation. Alternatively, calibrating and validating local rivers that are included in the current study area may also reduce the bias in hydrograph shape compared to calibrating and validating large rivers. On the other hand, the WRF-Hydro simulated river discharge decreases soon after it reaches the peak and much earlier than the observed discharge. The reason is again that the direct pass-through baseflow employed by this study does not account for slow discharge and long-term storage of the baseflow. As a result, the contribution from the baseflow to the river discharge in model simulations does not stay as long as in real situations. In the observations, the river discharge decreases from the peak at a speed of ~500 m3/sec per day, while the modeled river discharge decreases from the peak at a speed of ~1667 m3/sec per day. Using exponential storage-discharge function for the baseflow may improve this situation."

3. In order to deal with the observed streamflow in Section 1, it is fundamental to work with weights.

Response:

Thanks for your comment. This is also a major concern of Reviewer #1. For all the experiments we present in the revised manuscript, we consider a higher weight for Station 1. We add description about weight in the revised manuscript in **Section 3.2 PEST files and settings**, and in **3.3 Calibration experiments:**

"The 96 observation points are given different weights based on the inversed mean of their observed discharge during the studied period (see the detailed description in Sec. 3.3 and Sec. 4.1)."

"We find that by using the same absolute weight for all four stations, the calibration helps three stations (Station 2, 3, and 4) with large water volumes to generate more reasonable results than do the default parameters; however, the results for Station 1, which has a relatively small volume of water, is not always better than the discharge that is modeled by using default parameters. Thus, we assign a weight of 15.0 for

Station 1 versus a weight of 1.0 for the other three stations according to the inversed mean of observed discharge over these four stations in April 2013. The ratio of the weights between Station 1 and the other three stations stays similar even if the means are calculated based on different time periods."

We describe the results in **Section 4.1** by comparing PEST calibration using equal weight and a higher weight for Station 1:

"Compared with the results calibrated by using equal weights for all the stations, by giving a higher weight to Station 1 the model bias over Station 1 is significantly reduced, with a higher NSE (0.87 with higher weight versus 0.14 with equal weight) and lower RMSE (48.1 versus 123.6)."

A figure is shown above in the response to Reviewer #1.

Minor comments, grammar and typos

P6 LL6-17: not clear if in this case overland flow is switched on. It should.

Response:

Yes, it is. We change the sentence to emphasize this regard: "Overland flow, saturated subsurface flow, gridded channel routing, and a conceptual baseflow are active in this study"

"If overland flow is active as it is in this study, it passes water directly to the channel model."

P6 L19: probably "tools"

Response: corrected

P7 L18: GENPARM.TBL

Response: fixed

P8 LL11-12: master, not mater. The full stop is missing.

Response: fixed

P8 L30: As it is a common problem, it is usually solved 'simply' reallocating manually the stations. It's a pity to miss streamflow data for this reason

Response:

We regret we didn't do that, and it should be done in future applications. To clarify and to emphasize this regard, we change the sentence to:

"We calibrated WRF-Hydro using four USGS sites (referred to as Station 1, Station 2, Station 3, and Station 4 hereafter), as shown in Fig. 1. More USGS sites could be included if one manually reallocated the stations that were not properly assigned to the desired location on the channel network by the GIS tool."

P9 L24 and following: I suggest to explicitly declare also the meaning of the ovn parameter

Response: added the vegetation type in Table 1 for each ovn parameter*

P12 L17: 50%, maybe

Response: Apologize for the wrong number. It should be 66.7%

Figs.2 and 3: April 12 is missing. It should be the first validation day, I guess.

Response: added.

Figs. 4 and 5: the same for April 16

Response: this figure is deleted as it is for the 7-day calibration result

Table 3: the note is incorrect, it refers to information about the 3-day calibration

Response: fixed

Section 4.3: this is a purely "hydrological" analysis that could be skipped, given the numerous limitations of the calibration procedure and the focus on the implementation of the PEST software

Response:

We still keep this section and the figure in the revised manuscript. One of the reasons is that, it is more obvious using these stations to show the benefit (generate better hydrograph shape) of SVD-based regularization compared to non-regularization, which is the method we emphasize that should be applied for calibrating hydrological and environmental models. The other reason is that, these rivers are relatively small and are local rivers that included in the current study area, therefore, (1) the lag time between precipitation peak and discharge peak is much shorter than those for Station 3, 4, 5, and the hydrograph shape is well captured by the optimal parameter set; (2) the slow-discharge effect from baseflow is also relatively small so the discharge decrease faster after the peak than that for Station 3, 4, 5. This is also well captured by the optimal parameter set. Overall, the WRF-Hydro calibration with current configuration actually did a good job to capture the hydrograph features for these stations.

P16 LL9-10: please check the sentence

Response: this sentence is deleted, and changes are made accordingly through the entire manuscript.

P16 L18: to investigate

Response: corrected.

A parallel workflow implementation for PEST version 13.6 in high-performance computing for WRF-Hydro version 5.0: a case study over the ~~Midwestern~~midwestern United States

~~Jiali~~[1]Jiali Wang, ~~Cheng~~[1]Cheng Wang, ~~Andrew~~[2]Vishwas Rao, [1]Andrew Orr, ~~Rao~~[1]Rao Kotamarthi

~~Argonne~~[1]Argonne National Laboratory, Environmental Science Division, 9700 South Cass Avenue, ~~Argonne~~Lemont, IL 60439, USA

[2]Argonne National Laboratory, Mathematics and Computer Science Division, 9700 South Cass Avenue, Lemont, IL 60439, USA

Correspondence to: Jiali Wang (jialiwang@anl.gov)

Abstract. ~~Surface hydrological models must be calibrated for each application region.~~ The Weather Research and Forecasting Hydrological ~~system~~ (WRF-Hydro) system is a state-of-the-art numerical model that models the entire hydrological cycle based on physical principles. ~~However, as~~As with other hydrological models, WRF-Hydro parameterizes many physical processes. ~~As a result~~Hence, WRF-Hydro needs to be calibrated to optimize its output with respect to observations~~. However, when~~ for the application region. When applied to a relatively large domain, both WRF-Hydro simulations and calibrations require intensive computing resources and are best performed ~~in parallel.~~on multimode, multicore high-performance computing (HPC) systems. Typically, each physics ~~parameterization~~-based model requires a calibration process that works specifically with that model~~,~~ and is not transferrable to a different process or model. ~~Parameter Estimate Tool~~ The parameter estimation tool (PEST) is a flexible and generic calibration tool that can be used in principle to calibrate any ~~numerical code. However, PEST in~~of these models. In its ~~current~~existing configuration, however, PEST is not designed to work on the current generation of massively parallel ~~high-performance computing (~~HPC~~)~~ clusters. ~~This study~~To address this issue, we ported the parallel PEST to HPCs and adapted it to work with ~~the~~ WRF-Hydro. The porting involved writing scripts to modify the workflow for different workload managers and job schedulers, as well as developing code to connect ~~Parallel~~ parallel PEST to WRF-Hydro. ~~We~~To test the operational feasibility and the potential computational benefits of this first-of-its-kind HPC-enabled parallel PEST, we developed a case study using a flood in the ~~Midwestern~~midwestern United States in 2013~~ to test the operational feasibility of the HPC-enabled parallel PEST. We then~~

~~evaluate the WRF-Hydro performance in water volume and timing of the flood event. We also assess the spatial transferability of the calibrated parameters for the study area. We finally discuss the scale-up capability of~~. Results on a problem involving calibration of 22 parameters show that on the same computing resource used for parallel WRF-Hydro, the HPC-enabled parallel PEST ~~to provide insight for PEST's application to other hydrological models and earth system models on current and emerging HPC platforms. We find that, for this particular study, the HPC-enabled PEST calibration tool~~ can speed ~~up WRF-Hydro~~the calibration process by a factor of ~~30~~up to 15 compared ~~to~~with commonly- used PEST in sequential mode. The speedup factor is expected to be greater with a larger calibration ~~approaches.~~problem (e.g., more parameters to be calibrated or a larger size of study area).

1 Introduction

~~Hydrological models are important tools for research relevant but not limited to, water resource management, flood control, and hydrological response to climate change (Zanon et al., 2010; Papathanasiou et al., 2015). Conceptual hydrological models express hydrological processes in the form of abstract models that come from physical phenomenon and experience. Physically based hydrological models contain definite~~Physically based hydrological models contain detailed physical mechanisms to model the hydrological cycle, but many complex physical processes in these models are parameterized. For example, the state-of-the-art Weather Research and Forecasting Hydrological (WRF-Hydro) modeling system ~~(WRF-Hydro; (~~Gochis et al., 2015) has dozens of parameters that can be land- and river-type dependent and are typically specified in lookup tables. ~~Both conceptual hydrological models and physically based~~Therefore, these hydrological models need to be calibrated before they can be applied to research~~.~~ over different regions. In this context, calibration refers to ~~the hydrologists' need to adjust~~adjusting the values of the model parameters so that the model can closely match the behavior of the real system it represents. In some cases, the appropriate value for a model parameter can be determined through direct measurements conducted on the real system. ~~However, in~~In many situations, however, the model parameters are conceptual representations of abstract watershed characteristics and must be determined through calibration. In fact, model calibration is the most time-consuming step, not only for hydrological models, but also for ~~earth~~Earth system model development, because both parametric estimation and parametric uncertainty analysis require hundreds—if not thousands—

of model simulations to understand how perturbations in model parameters affect simulations of dominant physical processes and to find the optimum value of a single parameter.

WRF-Hydro is a ~~practical physics-based~~ numerical model that can simulate the entire hydrological cycle using advanced high-resolution data such as satellite and radar products. Compared ~~to~~with the traditional land surface model (LSM) used by WRF, WRF-Hydro provides a framework for multiscale representation of surface flow, subsurface flow, channel routing, and baseflow, as well as a simple lake/reservoir routing scheme. As a physics-based model, WRF-Hydro includes many complicated physical processes that are nonlinear and must be parameterized. ~~For example, the parameters for channel routing are prescribed as functions of stream order, not space; thus the~~The default parameters given by WRF-Hydro ~~are only~~may be valid ~~over a small~~for one region~~. Because channel routing can affect the accuracy of the model performance,~~ but not for another region. Hence calibration of related model parameters is often required in order to use the model in a new domain. In particular, for a large spatial domain such as the entire ~~Contiguous~~contiguous United States ~~(CONUS),~~ in order to develop the optimal parameter sets in a reasonable amount of time, the calibration must be conducted on ~~HPCs~~high-performance computing (HPC) systems in parallel instead of in the traditional sequential mode. To date, ~~there is~~ no such calibration tool ~~that~~ can ~~straightforwardly~~efficiently calibrate WRF-Hydro on ~~HPCs.~~HPC resources. Typically, each physics-based model needs a calibration code that is custom~~-~~ designed to work ~~with that particular numerical model. These custom-designed calibration codes/tools are highly challenging and do not offer flexibility; they are designed to operate~~ with that particular numerical model and its set of physics parameterizations, software architecture, and solvers. These custom-designed calibration codes are highly challenging and do not offer flexibility. Therefore, ~~there is a need for~~ a more flexible and generic calibration tool is needed that can calibrate any code that uses Message Passing Interface/Open Multi Processing (MPI/OpenMP) for parallelization on ~~HPCs~~HPC systems.

~~There are two general types of calibration methods for hydrological models: manual calibration and automatic calibration. Models for individual catchments have traditionally been calibrated by manually adjusting key model parameters within established ranges of parameters to obtain a best match between observed and simulated discharges. This procedure is time consuming, dependent~~

~~on the skill and experience of the modeler, and therefore prone to inconsistency between modelers. Automatic calibration is based on stochastic or mathematical methods and thus is more widely~~One widely used generic and independent calibration tool is the parameter estimation tool (PEST). PEST (Doherty, 2016) conducts calibration automatically based on mathematical methods and thus is applicable for optimizing nonlinear parameters. Compared with manual calibration, automatic calibration is more efficient and effective~~,~~ because it avoids interference from human factors (Madsen, 2000; Getirana, 2010). ~~One widely used automatic calibration tool~~The uniqueness of PEST is ~~Parameter Estimation Tool (PEST; Doherty 2016), which uniquely~~that it operates independent of models~~. There~~: there is no need to develop additional programs~~/~~ or codes for a particular model except preparing the files required by PEST (as described in ~~Sect~~Sec. 3.2~~, because~~). PEST ~~works with that model through~~has four modes of operation. One of the ~~model's own input~~modes is regularization mode, which supports the use of Tikhonov regularization and ~~output files. PEST implements a particularly robust variant of the Gauss-Marquardt-Levenberg method (Levenberg, 1944; Marquardt, 1963) to estimate parameters. This method requires a continuous relationship to exist between model parameters and model outputs, but it can normally find the~~ is found better for serving environmental models because, if implemented properly, it supports model predictions of minimum ~~in the objective function in a fairly shorter time period~~ error variance, is numerically stable, and embraces rather than ~~other parameter estimation methods. This is especially important when model runs are lengthy or when many parameters must~~ eschews the heterogeneity of natural systems. Singular value decomposition (SVD) can be ~~calibrated.~~used as a regularization device to guarantee numerical stability of the calibration problem. Parallel PEST is able to distribute many runs across many computing nodes using master-~~slave~~worker parallel programing. ~~However, to the~~ To our best ~~of our~~ knowledge, however, no approach is available that allows users to submit jobs using PEST parallelization to a typical supercomputing facility that uses job scheduling and workload management ~~using~~such as Simple Linux Utility for Resource Management (SLURM), Portable Batch System (PBS), and Cobalt. A previous study (Senatore et al., 2015) used PEST to calibrate WRF-Hydro over the Crati River Basin in ~~Southern~~southern Italy. ~~However, because~~Because the study area was relatively small, ~~they~~the authors were able to conduct the calibration using PEST in sequential mode~~.~~ (Alfonso Senatore, personal communication, 2018).

~~In this~~This study~~, we ported~~ aims to (1) port parallel PEST to HPC clusters operated by the U.S. Department of Energy (DOE) and ~~adapted~~adapt it to work with WRF-Hydro~~. Porting involved writing additional scripts to modify~~, (2) evaluate the ~~workflow for SLURM, Cobalt, and PBS and developing code to connect parallel PEST to WRF-Hydro. In particular, we aim to (1) calibrate the parameters of WRF-Hydro to improve model~~ performance ~~with realistic values maintaining their physical meanings; (2) speed up calibration for this particular study case and provide the capability to WRF-Hydro users; and (3) explore the scale-up capability~~ of HPC-enabled parallel PEST linked to WRF-Hydro~~.~~ by calibrating a flood event, and (3) explore the scale-up capability and computational benefits of HPC-enabled parallel PEST by assigning different computing resource to the entire calibration process.

2 Model description

2.1 Study area

The case presented here is one of the worst floods experienced by greater Chicago area in the ~~last~~past three decades~~, which~~; the storm occurred on April 18, 2013 (Campos and Wang, 2015). According to the National Weather Service (NWS), the heaviest 24-hour accumulated rainfall during this storm reached 201.4, 171.1, and 136.4 mm across Illinois, Iowa, and Missouri, respectively. The Mississippi River crested at 10.8 m (1.7 m above flood stage), and the Illinois River crested in Peoria, Illinois, at 8.95 m; ~~this~~these river cresting broke the previous record of 8.78 m, set in 1943, and was 4.55 m above the historical normal river stage (NWS, 2013). Campos and Wang (2015) conducted three-domain nested WRF simulations to understand the dynamical and microphysical mechanisms of the event. Our study builds on the smallest domain of that study, which covers the majority of Illinois, Iowa, and Missouri at a spatial resolution of 3 km ~~for the atmospheric and land surface model (Fig. 1).~~(Fig. 1). The domain size is 750 km from west to east and 660 km from south to north.

~~During the 10-day period of this studied case, light to moderate rain occurred on April 8 through 11, 2013, followed by a relatively dry period from April 12 to 14. Then a heavy rain event began on April 15 and peaked on April 18. The heaviest rain band moved east of the study area on April 19. The rainy event ended over the study area on April 20.~~

2.2 WRF-Hydro configuration

This study employs WRF-Hydro version 5 with a basic configuration. This configuration does not use nudging techniques or spatially distributed soil-related parameters as used in the National Water Model configuration. WRF-Hydro has been tested in several different cases that focused on different hydrometeorological forecasting and simulation problems (e.g., Gochis et al., 2018; Yucel et al., 2015; Senatore et al., 2015; Arnault et al., 2016), and it shows reasonable accuracy in simulated streamflow after being carefully calibrated. For details of the WRF-Hydro modeling system, see Gochis et al. (2018).~~WRF-Hydro employs a multiscale modeling approach to handle the local landscape gradient features. Specifically, WRF-Hydro uses a subgrid disaggregation-aggregation procedure. For each time-step at which forcing data are available, the column moisture stays within the LSM and is disaggregated from the LSM grid to a high-resolution routing grid (Gochis and Chen 2003). After disaggregation, the routing schemes are executed using the high-resolution grid values. After execution of the routing schemes, the high-resolution grid values are aggregated back to the native LSM grid. For details of each routing component, see Gochis et al. (2015), Yucel et al. (2015), and Senatore et al. (2015).~~

Currently, two LSMs are available in WRF-Hydro for representing land-surface column physics: Noah (Chen and Dudhia, 2001) and Noah Multi-parameterization (~~NoahMP~~Noah-MP; Niu et al. 2011). We utilize ~~NoahMP~~Noah-MP LSM because compared ~~to~~with Noah LSM it shows obvious improvements in reproducing surface fluxes, skin temperature over dry periods, snow water equivalent, snow depth, and runoff (Niu et al. 2011). ~~Compared to LSM, one major advantage of WRF-Hydro system~~The Noah-MP is ~~that, WRF-Hydro system can keep the infiltration capacity exceedance as ponded water within the model domain. This ponded water is subsequently available for lateral redistribution, which combine the ponded water with new precipitation for calculating the infiltration amount in the next time step~~configured~~. WRF-Hydro has been tested in several different cases that focused on different hydrometeorological forecasting and simulation problems (e.g., Gochis et al., 2015; Yucel et al., 2015; Senatore et al., 2015; Arnault et al., 2016), and it shows reasonable accuracy in simulated streamflow.~~

~~This study employs WRF-Hydro version 5 with a basic configuration. This configuration does not use nudging technique as used in the National Water Model configuration and spatially distributed~~

soil-related parameters. The LSM is at a grid spacing of 3 km, and the aggregation factor is 15,; that is, starting from a 3- km LSM resolution in the domain shown in Fig. 1, hydrological routing is performed at a spatialgrid resolution of 200 m., with 3285 south-north × 3735 west-east grid cells. We use a time step of 10 seconds for the routing grid in order to maintain model stability and prevent numerical dispersion of overland flood waves. The time step also meets the Courant condition criteria for diffusive wave routing on a 200- m resolution grid. The WRF-Hydro is configured to be in offline or uncoupled mode—there is no online interaction withbetween the WRF-Hydro hydrological model and the WRF atmospheric model. SurfaceOverland flow, saturated subsurface flow, gridded channel routing, and a conceptual baseflow are active in this study. The gridded channel network uses an explicit, one-dimensional, variable time-stepping diffusive wave. A direct output-equals-input "pass-through" relationship is adopted here to estimate the baseflow. Although the baseflow module is not physically explicit, it is very important because the water flow in the channel routing areis contributed by both the overland flow and baseflow. If the overland flow is active as it is in this study, it passes water directly to the channel model. In this case the soil drainage is the only water resource flowing into the baseflow buckets. IfHowever, if the overland flow is deactivated but channel routing is still active, then WRF-Hydro collects excess surface infiltration water from the land model, and passes this water into the baseflow bucket. This bucket then contributes the water from both overland and soil drainage to the channel flow. Therefore, the baseflow must be active if the overland flow is switched off. This study does not consider lakes and reservoirs.

We use the geographic information system (GIS) tool that are(Sampson and Gochis, 2018) developed by the WRF-Hydro team to delineate the stream channel network, open water (i.e., lake, reservoir, and ocean) grid cells, and groundwater/baseflow basins. Meteorological input for the WRF-Hydro model system includes hourly precipitation; near-surface air temperature, humidity, and wind speed; incoming shortwave and longwave radiation; and surface pressure. In this study, the hourly precipitation is from the National Centers for Environmental Prediction (NCEP) Stage IV analysis at a spatial resolution of 4 km. The Stage IV data is based on combined radar and gauge data (Lin and Mitchell, 2005; Prat and Nelson, 2015), and has been shown to be temporally well correlated with high-quality measurements from individual gauges (see, e.g., Sapiano and Arkin, 2009; Prat and Nelson, 2015). The other hourly meteorological inputinputs are from the

second phase of the multi-institution North American Land Data Assimilation System project, phase 2 (NLDAS-2) (Xia et al., 2012a,b), at a spatial resolution of 12 km. NLDAS-2 is an offline data assimilation system featuring uncoupled LSMs ~~that are~~ driven by observation-based atmospheric forcing.

During the 15-day period of this studied case, light to moderate rain occurred on April 8 through 11, 2013, followed by a relatively dry period from April 12 to 15. Then a heavy rain event began on April 16 and peaked on April 18. The heaviest rain band moved east of the study area on April 19. The rainy event ended over the study area on April 20 (see Fig. S1 in Supporting Information). We start the WRF-Hydro simulation on Jan. 1, 2013, and run the model for more than three months to reach equilibrium. This 3-month period is considered as spin-up time and is excluded from model calibration and evaluation. We calibrate the river discharge calculated by the WRF-Hydro model from 00UTC April 9 to 00UTC April 12, 2013, considering it long enough to achieve our objective. We then evaluate the model performance against U.S. Geological Survey (USGS) observed river discharge from 00UTC April 12 to 00UTC April 25, 2013.

3 Calibration

3.1 Platforms

We customized parallel PEST to work on ~~two~~three different workload managers and job schedulers: SLURM at the National Energy Research Scientific Computing Center (NERSC), PBS at the Argonne National Laboratory Computing Resource Center, and Cobalt at the Argonne Leadership Computing Facility. The tests presented here are conducted on Edison ~~of~~at NERSC, which uses the SLURM workload manager and job scheduler. Edison is a Cray XC30 with a peak performance of 2.57 petaflops per second, 133,824 compute cores, 357 terabytes of memory, and 7.56 petabytes of disk storage. It has 5,586 nodes and 24 cores per node.

~~3.2 PEST files~~

~~Parallel PEST requires four types of input file:~~

~~Template files, which~~

The interface we have built between parallel PEST and the management software (SLURM here) is, in general, used for (1) setting the number of workers and the nodes for each worker to conduct a model run (WRF-Hydro here); (2) finding the nodes that are available; (3) setting up the working directory for the workers; (4) identifying the nodes that work for each worker; (5) passing the global files (same for all the working directory) to all the workers (these files include the lookup table files that are not to be calibrated, the namelist files for both LSM and hydrological sector, and restart files that generated by the previous simulations, or spin-up period); and (6) submitting the job for the entire calibration process, including parallel PEST and parallel WRF-hydro. This job can be submitted as a cold-start run or as a restart. The main difference for this interface on different management software is that different management software has its own way to submit jobs and identify available nodes. This difference requires some changes in the script we developed.

3.2 PEST files and settings

1. PEST requires three file types in both sequential and parallel mode. They are template files to define the parameters to be calibrated. For example, we generated CHANNEL.TPL, HYDRO.TPL, and GENPARAM.TPL based on the format of their corresponding lookup tables, which are CHANNEL.TBL, HYDRO.TBL, and GENPARAM.TBL, respectively. CHANNEL.TBL describes the features of a channel, such as bottom width, channel side slope, and Manning's roughness coefficients. HYDRO.TBL contains Manning's roughness coefficients for land use types. GENPARAM.TPL describes the parameters used in the Noah-MP LSM.
2. An, an instruction file, which defines to define the format of model-generated output files. For example, WRF-Hydro can output time series of streamflow over the forecast points (frxst_pts_out.txt) specified during model configuration. The instruction file follows the format of frxst_pts_out.txt and specifies the line number of each calibrated forecast point in frxst_pts_out.txt.
3. A , and a control file, which supplies to supply PEST with the size of the problem (e.g., how many parameters to be calibrated; how many observational points); initial parameter values and their lower and upper bounds; the increment of each parameter for forward-calculation; the names of all template and instruction files; observational values, and

~~weight for each parameter to be calibrated. PEST requires all these three file types in both sequential and parallel mode.~~

~~4.~~ and the settings for the calibration method. Parallel PEST uses a "master-worker" paradigm that starts model runs simultaneously by different workers (or in different folders). The master of parallel PEST communicates with each of its workers many times during a calibration. To run PEST in parallel mode, one also needs a management file to inform PEST where ~~each slave's~~the working folder is~~, as well as~~ for each worker and what the names and paths ~~of~~are for each model input file that PEST must write (i.e., lookup tables that come from template files) and each model output file that PEST must read (such as frsxt_pts_out.txt).

~~Parallel PEST uses a "master-slave" paradigm that starts model runs simultaneously in different folders (or by different "slaves"). The master of parallel PEST communicates with each of its slaves many times during the course of a calibration. When PEST needs to run a~~ The management file also set the maximum running time for each worker. For those workers that take longer than the maximum running time, PEST will stop the model ~~in a~~run by that particular ~~folder, the master notifies the slave to start the~~worker and assign that model ~~in that folder. Each slave starts the model execution accordingly, and informs the master that the model starts running. Once the simulation is completed in a particular folder, the slave signals the master, so the mater can read the particular output However, to the best of our knowledge~~run to another worker if there is one with nothing else to do.

To the best of our knowledge, however, parallel PEST is not designed to run on HPCs directly. We developed scripts and an interface to enable parallel PEST to run on HPCs using SLURM, PBS, or Cobalt workload managers and job schedulers. ~~This enables~~The development involved writing scripts to modify the workflow for different workload managers and job schedulers, as well as developing code to connect parallel PEST to WRF-Hydro. These developments enable parallel PEST to run many ~~slaves on~~workers at the ~~HPC~~same time; each ~~slave~~worker runs a parallel code (~~such as~~here WRF-Hydro) that uses more than one node, which could significantly ~~increase~~reduce the ~~computational performance~~wall-clock time of model calibrations. Although this master-~~slave~~worker parallelism may not be as efficient as a fully MPI approach, it is sufficient

for model calibration and requires the least effort for the current parallel PEST to run on ~~HPCs~~HPC systems.

This study presents calibration results from PEST using the SVD-based regularization in regularization mode to ensure numerical stability (Tonkin and Doherty, 2005). We focus on calibrating 22 parameters (see Table 1 and detail description in Sec. 3.3) using 96 observation points and 22 items of prior information for the calibrated parameters. In each item of prior information, a value equal to its default value provided by the WRF-Hydro v5.0 (or the log of its default value) is assigned for each adjustable parameter, assuming that default values are the preferred values. All prior information equations are assigned a weight of 1.0. We assigned five different regularization groups to the prior information: Manning's roughness coefficients specified by Strahler stream order in CHANPARM.TBL to one group; the parameters in HYDRO.TBL (Manning's roughness coefficients for overland flow as a function of vegetation types) to another group; and three global parameters for the Noah-MP (xslop1, refdk, and refkdt) in GENPARM.TBL to the remaining three groups. The 96 observation points are given different weights based on the inversed mean of their observed discharge during the studied period (see the detailed description in Sec. 3.3 and Sec. 4.1). For a detailed description of these settings see the PEST User Manual (Doherty, 2016).

3.3 Calibrated experiments

The primary objective of this study is to ~~present~~build a bridge for linking the ~~operational~~ parallel PEST and WRF-hydro on the ~~scale-up capability~~basis of HPC clusters and to explore the ~~HPC-enabled parallel PEST for use with WRF-Hydro calibration over a relatively large domain.~~computational benefits of this bridge. We ~~focus less on~~ do not attempt to extensively ~~assessing the performance of~~ assess each individual tool or address questions in each individual domain, such as optimizing the objective functions in PEST or calibrating WRF-Hydro for a long time period considering all the ~~WRF-Hydro model.~~relevant parameters to achieve an optimal parameter set. The calibration ~~and validation~~period thus is limited to only ~~7~~three days, ~~considering it is~~which we believe long enough to achieve our objective and to understand WRF-Hydro's sensitivity to ~~multiple parameters. The calibration compares WRF-Hydro modeled river discharge~~

~~to U.S. Geological Survey (USGS) surface water observations.~~the calibrated parameters. We ~~originally choose 11 USGS sites across the study area. However, because of inaccuracies introduced when projecting geospatial data from one coordinate system to another by the ArcGIS tool, three of the observational sites were not properly assigned to the desired location on the channel network. This situation is common in hydrographic data processing and well known to hydrologists (Sampson and Gochis, 2018). Among the remaining eight sites, four have discontinuous or missing data over the calibration period. Therefore, we~~ calibrated WRF-Hydro using four USGS sites (referred to as Station 1, Station 2, Station 3, and Station 4 hereafter), as shown in Fig. 1~~ with their site number.~~. (More USGS sites could be included if one manually reallocated the stations that were not properly assigned to the desired location on the channel network by the GIS tool.) We then transfer the calibrated parameters to other ~~sub-basins~~subbasins in the study area to assess the transferability of the calibrated parameters. Although~~ there are~~ many parameters, including spatially distributed parameters and constant parameters in the lookup tables, ~~that~~ affect the model performance, we ~~only~~ calibrate only the parameters in lookup tables and do not consider the spatial variability of ~~each parameter~~other parameters or their scaling factors. We acknowledge that ~~there are~~some studies ~~that~~ calibrate a single scaling factor (without considering its spatial variability, however) of overland roughness coefficients (OVROUGHRTFAC) rather than the actual value of each land type in the lookup table (e.g., Kerandi et al., 2018). Although this approach reduces the number of calibrated parameters, however, it has less flexibility because changing one factor will change all the parameters that use the same proportion. ~~In addition, a single scaling factor holds the same for the entire domain, which may work well for a small domain, but could be problematic for a large domain. Thus, we suggest the calibration of spatially distributed parameters requires more knowledge and understanding of the study area and deserves future studies. In this study, we calibrate the roughness coefficients for each land type rather than calibrating a single scaling factor.~~

For ~~most~~the calibration exercises we ~~document~~conduct here, the retention depth factor (RETDEPRTFAC) is fixed at 0.001. This value is reasonable because the ~~modelled~~modeled discharge of our particular configuration (~~Sect~~Sec. 2.2) using default parameters is ~~much~~ lower than observed discharge. Reducing this factor from 1 to 0.001 keeps less water in water ponds and more water on the surface so it can contribute to river discharge. First, we calibrate 48 parameters

based on a 3-day simulation from April 9 to ~~12~~April 11, 2013 (Table S1 in Supporting Information). This calibration uses the estimation mode in the PEST tool and considers equal weight for all four USGS stations. We calibrate ~~the~~ Manning's roughness coefficients for both channels and land-use types, the deep drainage (SLOPE), infiltration-scaling parameter (REFKDT), and saturated soil lateral conductivity (REFDK). ~~The~~ Manning's roughness coefficients control the hydrograph shape and the timing of the peaks; the ~~infiltration factor, saturated hydraulic conductivity~~SLOPE, REFKDT, and ~~deep drainage~~REFDK control the total water volume. Second, based on the knowledge we learn from the ~~3-day~~48-parameter calibration (see details in ~~Sect~~Sec. 4.1), for the same 3-day period, we ~~redefine~~reduce the number of calibrated parameters from 48 to ~~calibrate~~22 according to the sensitiveness of the WRF-Hydro model to the adjustable parameters. For example, during the calibration we find that Manning's roughness coefficients for several land types barely change because these land types (e.g., tundra, snow/ice) are not present in the study area. We also learn that even though the calibrated WRF-Hydro parameters can generate discharge results that closely resemble observations, the physical meaning of several parameters are not appropriate because of the wide range of those parameters that we set in the PEST control file. For example, Manning's roughness coefficient for stream order 1 (0.199) is calibrated smaller than that for stream order 2 (0.218); the overland roughness coefficients for evergreen needleleaf forest (0.043) and mixed forest (0.023) are calibrated smaller than for cropland/woodland (0.046). Neither of these is true in the real world. We therefore adjust the range of many parameters according to the literature (Soong et al., 2012) to maintain their physical meanings (Table 1). ~~We also extend our calibration period to 7 days to include a heavy precipitation period. Although a period of 7 days is still very short compared to the traditional calibration period of at least 1 year, we find it provides more appropriate parameter estimation — as well as better results of simulated hydrograph shape and the total water volume — than does the 3-day calibration.~~We find that by using the same absolute weight for all four stations, the calibration helps three stations (Station 2, 3, and 4) with large water volumes to generate more reasonable results than do the default parameters; however, the results for Station 1, which has a relatively small volume of water, is not always better than the discharge that is modeled by using default parameters. Thus, we assign a weight of 15.0 for Station 1 versus a weight of 1.0 for the other three stations according to the inversed mean of observed discharge over these four stations

in April 2013. The ratio of the weights between Station 1 and the other three stations stays similar even if the means are calculated based on different time periods.

3.4 Statistics

This study employs three statistical criteria: Nash–Sutcliffe efficiency (NSE; Nash and Sutcliffe, 1970; Moriasi et al., 2007), root-mean-square error (RMSE), and Pearson correlation coefficient (PCC). RMSE and PCC evaluate model performance in terms of bias and temporal variation. NSE quantitatively describes the accuracy of ~~modelled~~modeled discharge compared ~~to~~with the mean of the observed data. Equation (1) calculates the NSE with defined variables:

$$NSE = 1 - \frac{\sum_{t=0}^{n}\left(Y_t^{obs} - Y_t^{sim}\right)^2}{\sum_{t=0}^{n}\left(Y_t^{obs} - Y_{mean}^{sim}\right)^2}, \quad (1)$$

where Y_t^{obs} is the *t*th observed value from USGS sites for river discharge , $Y_t^{sim}$is the *t*th simulated value from the WRF-Hydro output, Y_{mean}^{obs} is the temporal average of USGS observed discharge, and *n* is the total number of observation time points. An efficiency of 1 (NSE = 1) corresponds to a perfect match between modeled discharge and observed data. An efficiency of 0 (NSE = 0) indicates that the model predictions are as accurate as the mean of the observed data. An efficiency below zero (NSE < 0) occurs when the model is worse than the observed mean. Essentially, the closer the NSE is to 1, the more accurate the model is.

4 Results

4.1 ~~Three-day~~WRF-Hydro calibration and validation

Based on the knowledge we gained from the 48-parameter 3-day calibration, we adjust the range of critical parameters in the PEST control file to main their physical meanings. For example, we set Manning's roughness coefficient larger for stream order 1 than for stream order 2. We also adjust the parameter range of the overland roughness coefficient for multiple land covers, such as forests. We exclude the parameters that WRF-Hydro is not sensitive to for this study, in order to constrain the problem size considering the availability of computational resources. However, if the studied area is much larger with more land types than the study area here, then there would be more parameters to calibrate. Also, hundreds of constant parameters in the Noah-MP model could

affect the WRF-Hydro results (Cuntz et al. 2016) and can be calibrated. Both these situations would increase the burden of WRF-Hydro calibration. We perform the same 3-day calibration from April 9 to April 11, 2013. Figure 2 shows the results of the 3-day modeled discharge (in cubic meters) using default and calibrated parameters after five iterations, as well as observed discharge ~~from April 9 to 12.~~ . The four stations are calibrated by considering different weights. Compared with the results calibrated by using equal weights for all the stations, by giving a higher weight to Station 1 the model bias over Station 1 is significantly reduced, with a higher NSE (0.87 with higher weight versus 0.14 with equal weight) and lower RMSE (48.1 versus 123.6). Over Stations 2, 3, and 4, which sit on rivers with relatively large water volumes, the modeled discharge using the default parameter underestimates the streamflow by more than 65%. PEST detects this underestimation and immediately adjusts the parameters and increases the modeled discharge during the first iteration. After the ~~fifth~~third iteration, the difference in calibrated results between different iterations is relatively small~~, and~~ . We allow the PEST ~~performed 12~~to conduct five iterations ~~before finding the~~and use the parameters obtained from the fifth iteration as our optimum parameters. ~~Here we only show results generated by default parameters and by parameters calibrated from the 1st, 5th, and 12th iterations. Over Stations 2, 3, and 4, which sit on rivers with relatively large water volumes, the discharge modeled by the default parameters is much lower than discharge seen in observations. PEST detects this underestimation. It immediately adjusts the parameters and increases the modeled discharge during the first iteration. After adjusting the parameters for several iterations~~As shown in Table 2, when the optimum parameters are used, the modeled ~~discharge gets~~ discharges are much closer to the observations compared ~~to the modeled results that used the~~ with the modeled results when the default parameters were used. The NSEs for the four stations increased from 0.73 (Station 1), -54.4 (Station 2), 157.3 (Station 3) and -1316.9 (Station 4) to 0.87, 0.64, 0.05, and -58.78, respectively, being closer to 1. The RMSEs decreased from 69.3, 3925.2, 3981.3, and 4391.3 m^3/sec to 48.1, 318.2, 308.7, and 934.6 m^3/sec, respectively. Giving a lower weight for the three large river stations does not change the calibration results much.

During the validation period, compared with the modeled discharge using default parameters~~.~~, as shown in Table 2 ~~shows~~, the ~~statistics of model performance using default and calibrated parameters~~NSEs for all four stations ~~during~~ are increased to be closer to 1; RMSEs are decreased

by 50% or more; and the ~~calibration~~correlation coefficients between the observed and ~~validation period. Compared to the discharge that was modeled using the default WRF-Hydro parameters, overall, the calibrated~~ modeled discharge ~~matches observations better~~are increased from 0.8, 0.76, 0.21, and 0.72 to 0.98, 0.82, 0.80, and 0.75. Compared with the results of calibration using the estimation mode (no regularization) in PEST, the SVD-based regularization generates slightly better hydrograph shape with 24-hour later discharge peaks that are closer to the observations. However, a problem remains with the hydrograph shapes of the modeled discharge, especially ~~at the three stations with large volumes of~~ with the modeled peak of discharge. For Station 1, the WRF-Hydro almost captures the timing of the peak of discharge, although it still underestimates the water~~. Note~~ volume by ~25%. The reason is that this ~~3-day period only experienced a light rain over the study area. The streamflow in the rivers is, therefore, mostly from groundwater and overland flow from upstream or from previous precipitation events. The contribution of overland flow is small for this period because the amount of precipitation was also small, so the main contributor to river discharge in the real situation was from the groundwater. However, in this study WRF-Hydro~~ study uses a direct pass-through ~~groundwater model~~baseflow module, which does not account for slow discharge and long-term storage of the baseflow. ~~Therefore,~~ Therefore, the largest contribution to river discharge is from precipitation, and groundwater does not contribute much discharge to the channels~~. This situation causes the model to greatly underestimate discharge, so the calibration adjusts critical parameters aggressively to increase the streamflow to match the observations. When we apply these calibrated parameters to the following days, the calibrated discharge is much higher than the observed discharge during the heavy precipitation period, as shown by Fig. 3 and the RMSEs for the validation period in Table 2. However, after the~~ in a long-term view, as is also true for the other three large river stations. Different from Station 1, for the other three large river stations, the WRF-Hydro modeled discharge increases soon after the peak of precipitation and reaches a peak on April 21, 2013, which is much earlier than the observed peak of river discharge (near April 24). The reason is that the water contributions for these stations are from a larger river basin (Mississippi River) than we included in our current study area. Thus, when a heavy precipitation event~~, the modeled discharge~~ occurs over the entire river basin, there will be a significant lag time (especially at the lower part of the basin) between the peak of precipitation amount and the peak of river discharge. For example, the precipitation over the upper part of Mississippi River Basin (MRB) has a peak amount

on April 18–19, but the river discharge did not reach its peak until April 24. Because our studied area covers only half of the MRB, the modeled river discharge has a shorter delay period after the peak of precipitation than does the observed river discharge. Enlarging the study area to include the entire MRB may improve this situation. Alternatively, calibrating and validating local rivers that are included in the current study area may also reduce the bias in hydrograph shape compared to calibrating and validating large rivers. On the other hand, the WRF-Hydro simulated river discharge decreases soon after it reaches the peak and much ~~faster~~earlier than ~~in~~ the observed ~~situation (Fig. 3). This~~discharge. The reason is again ~~might be due to~~ that the direct pass-through ~~groundwater model we adopt in~~ baseflow employed by this study~~, which uses an output-equals-input relationship between soil drainage and the discharge into river channels. This model does not allow~~ does not account for slow discharge and long-term storage of ~~baseflow in each conceptual bucket, and thereby not be able to fully represent~~the baseflow. As a result, the contribution ~~of groundwater to streamflow~~from the baseflow to the river discharge in model simulations does not stay as long as in real situations. In the observations, the river discharge decreases from the peak at a speed of ~500 m^3/sec per day, while the modeled river discharge decreases from the peak at a speed of ~1667 m^3/sec per day. Using exponential storage-discharge function for the baseflow may improve this situation.

~~From this 3-day calibration experiment, we learn that the WRF-Hydro output is not sensitive to several parameters we calibrated in this particular study. For example, Manning's roughness coefficients for several land types barely change during the calibration because these land types (e.g., tundra, snow/ice) are not present in the study period and area. We also learn that even though the calibrated WRF-Hydro parameters can generate discharge results that closely resemble observations, the physical meaning of several parameters are not appropriate due to the wide range of those parameters that we set in the PEST control file. For example, as shown in Table S1, the Manning's roughness coefficient for stream order 1 (0.199) is calibrated smaller than that for stream order 2 (0.218); the overland roughness coefficients for evergreen needleleaf forest (0.043) and mixed forest (0.023) are calibrated smaller than cropland/woodland (0.046). Neither of these is true in the real world.~~

4.2 Seven-day calibration and validation

Based on the knowledge we gained from the 3-day calibration, we adjust the range of critical parameters in the PEST control file. For example, we set the Manning's roughness coefficient larger for stream order 1 than for stream order 2. We also adjust the parameter range of the overland roughness coefficient for multiple land covers, such as forests. With the adjusted range of parameters, we perform 7-day calibration from 00:00 UTC on April 9, 2013, to 00:00 UTC on April 16, 2013, when there is an increased streamflow that the simulation does not capture using 3-day calibrated parameters. The entire calibration takes 12 iterations. Figure 4 shows the results of modeled discharge (in cubic meters) using default and calibrated parameters (from the first, fifth, and 12th iterations), as well as observed discharge from April 9 to 16. Over Stations 2, 3, and 4, the modeled discharge using the default parameter underestimates the streamflow by more than 100%. PEST detects this underestimation and starts adjusting parameters to increase the discharge to match the observations. Compared to the modeled discharge using default parameters during the validation period, as shown in Table 3, the RMSE decreased from 624.9 (Station 1), 5162.9 (Station 2), 4990.0 (Station 3), and 5098.3 m^3/sec (Station 4) to 283.1, 637.9, 666.8, and 1202.8 m^3/sec, respectively. The correlation coefficient between observed and modeled discharge increased from 0.71, 0.90, 0.87, and 0.82 to 0.97, 0.99, 0.96, and 0.86. Note that, although the calibration helps three stations (Station 2, 3, and 4) with large water volumes to generate more reasonable results than the default parameters, the results for Station 1, which has a relatively small volume of water, is not always better than the discharge that is modeled using default parameters (Tables 2 and 3). This might be because we use the same absolute weight for all the stations when we perform the calibration. Using a higher weight for Station 1 may help improve this situation and generate better results for this station.

Although a period of 7 days is still very short for calibration compared to traditional calibration period of at least 1 year, we find that the 7-day period provides better and more appropriate parameter estimation than does the 3-day calibration, and it does a better job of capturing the hydrograph shape and the total water volume. Comparing the validation statistics between Tables 2 and 3 as well as Figs. 3 and 5, we find the 7-day calibrated parameters generate better results than do the 3-day calibrated parameters for both water volume and the hydrograph shape over the validation period. Compared to the discharge modeled using 3-day calibrated optimum parameters,

there is a 17–33% increase in the simulated streamflow (1,400–2,800 m^3/sec) calculated over Station 2, 3, and 4 using the 7-day calibrated parameters from April 12 to 16. The RMSE is 3,400–3,600 m^3/sec when calculated using the 3-day calibration, but this decreases to 600–1,200 m^3/sec when calculated using the 7-day calibration. The correlation coefficient between observed and modeled discharge using the 7-day calibration is 0.8–0.99, but that using the 3-day calibration is only 0.7–0.8. However, there is still a problem with the temporal variability of the modeled discharge, especially over the rivers with large discharge. When there is precipitation, the discharge immediately increases and is higher than the observed discharge. After the precipitation period, when the observed discharge still stays high, the modeled discharge decreases sooner and thus is smaller than the observed discharge. This might be the direct pass-through approach simplified groundwater flow, and does not represent the interaction between stream flow and groundwater properly in this case study.

4.3 Evaluation of spatial transferability of the modeling system

In this section, we apply the calibrated parameters for the four stations (black circles) in Fig. 1 to other 13 stations in the study area. As mentioned before, because of inaccuracies in the spatial location of station data and digital elevation models, and because of small errors introduced when projecting geospatial data from one coordinate system to another using the ArcGIS tool, only four stations are mapped on the river systems (crosses in Fig. 1); others are slightly shifted out of their closest grid cell. One of these four sites (Station 5) is located on a relatively small river, and others are located on larger rivers. The following analyses assess the transferability of the calibrated parameters from particular sites to other sites that are in the study area but not calibrated. The assessment compares the observed discharge with the closest grid cells from the discharge output of WRF-Hydro. Figure 6 shows the observed and modeled discharge using default and calibrated parameters. Overall, WRF-Hydro's default parameters underestimate the discharge. WRF-Hydro also generates an earlier discharge peak compared to observations over the four stations (Stations 5, 6, 7, and 8) in this particular study. The calibrated model results increase the discharge and generate a hydrograph shape that is closer to the observations than the default model results do. The absolute error of simulated discharge decreases by 12.8%, 22.8%, 46.8%, and 49.9%, respectively, over Stations 5 through 8, compared to the default simulated discharge. However,

~~because we did not specifically calibrate these stations based on observations, there are still differences between the calibrated results and observations.~~

~~5 Discussion and summary~~

~~5.1 Scale-up capabilities~~

4.2 Computational benefits of parallel PEST on HPCs

The ability to scale up the calibration of WRF-Hydro by using parallel PEST on ~~HPCs~~HPC systems is determined by two factors: the scale-up capability of ~~WRF-Hydro,~~parallel PEST and the scale-up capability of ~~PEST~~WRF-Hydro. In ~~the course of~~ calibrating WRF-Hydro, PEST ~~must run the WRF-Hydro model many times. PEST~~first makes ~~some~~as many model runs as there are adjustable parameters to calculate Jacobian matrix (Doherty, 2016). The Jacobian matrix has a column for each calibrated parameter and a row for each observation and each item of prior information that set in the PEST control file. These model runs are independent between ~~slaves.~~workers and can be easily parallelized. Each ~~slave run~~worker runs the model ~~using~~with temporarily incremented parameters that are defined in the template and control files. ~~These model runs can be easily parallelized. However~~Then, PEST ~~also need~~needs to make ~~some other~~additional model runs to test parameter ~~upgrades. These runs are calculated based on~~updates. Different from the Jacobian runs, these additional runs are performed by using different Marquardt lambdas~~. The~~, and the search for a Marquardt lambda that achieves the best set of parameters is a serial ~~procedure — what~~ iterative process. The lambda to use for the next run depends on the outcome of the model run conducted using the previously chosen lambda. ~~This in fact is the major bottleneck of parallelization of the PEST code.~~ Although serial testing of Marquardt lambdas may quickly find the optimal Marquardt lambda in the first or second series of model runs, it is an inefficient use of computing resources because other processors are idle while only one process is searching the lambdas. This is especially true when the model domain is large and requires extensive computing resources.

This study employs "partial parallelization" for the lambda-testing procedure (Doherty, 2016), so ~~all the processors~~multiple workers can be used to calculate parameter upgrades based on a series of lambda values that are related to each other by a factor of RLAMFAC set in the PEST control

file. ~~This partial parallelization makes the scale-up challenging when more processors are in use, because generating many Marquardt lambdas does not always guarantee that the best Marquardt lambdas were the ones generated. As a result, the calibration process may converge more slowly when using more slaves than it does when using less slaves. We tested different numbers of slaves (35, 50, 70, and 105) for the 32-parameter calibration experiment. In total, each of these tests uses 71, 101, 141, and 211 nodes; two nodes for each slave run WRF-Hydro, and one node runs PEST master to coordinate jobs and communicate with the slaves. The results shown in Figs. 2–6 and Tables 2–3 are from a calibration using 35 slaves; PEST conducted 12 iterations before finding the optimum parameters. We find that using different numbers of slaves generates slightly different parameter values and involves different numbers of iterations. For example, using 70 slaves only takes eight iterations and 41% of the wall-clock time 35 slaves used to find the optimum parameters. The calibrated parameters are slightly different from those generated by 35 slaves (Table 1, last column), and they generate slightly better results than does the 35-slave test compared to observed discharge. We finish the 12 iterations using 35 slaves (71 nodes) within 73 hours, and the eight iterations using 70 slaves (141 nodes) within 30 hours. More than 800 model runs were conducted for entire calibration process including calculating the Jacobian Matrix as well as testing the parameter upgrades. In fact if more nodes were used by each slave for the calculation, the wall-clock time can be further reduced. If these calibration were conducted sequentially on personal computers, the same calibration process would have taken 60–80 days for a 7-day calibration.~~ We also set the value of PARLAM to -9999 in the management file so only one cycle of parallel WRF-hydro runs is devoted to testing Marquardt lambdas. For additional details on these parameters and their settings see the PEST User Manual (Doherty, 2016).

~~Our~~

In this study ~~finds that, depending on~~ we test the computational performance of HPC-enabled parallel PEST using different number of ~~parameters being calibrated (e.g., 32 parameters in this particular study), using 32 to 64 slaves shows fairly good scale-up capability; most of the time consumed by PEST is for running WRF-Hydro~~workers (6, 12, and ~~the number of slaves can be used to carry out model runs to generate the Jacobian matrix. However, using 105 slaves (211 nodes) does not result in fewer iterations or a shorter wall-clock time than using 70 slaves. In fact, using more than 64 slaves may not be necessary because generating many~~ 23) for the 22-parameter

calibration. As shown in Table 3, we conducted five experiments: Test 1 uses 23 workers, Test 2 uses 12 workers, and Test 3 uses 6 workers. All three tests use two nodes for each worker to run WRF-Hydro in parallel. The maximum number of lambda-testing runs undertaken per iteration is set to 15, 10, and 5 for Test 1, 2, and 3, respectively, to make sure that only one cycle of WRF-hydro runs is devoted (using 15, 10 and 5 workers from Tests 1, 2, and 3, respectively) to testing Marquardt lambdas. Note that the maximum number of lambda-testing runs should be set equal to or less than the workers available. Otherwise, another cycle of WRF-hydro runs needs to be conducted. In fact, generating more Marquardt lambdas does not always guarantee ~~generating~~that the best Marquardt lambdas~~. In addition, at least for~~ are generated. In contrast, it may make the ~~calibrations conducted in this study, in each iteration~~model convergence slower (here, PEST ~~runs the~~) or even model ~~either 32 or 64~~failure.

In order to test the trade-offs between the computing nodes used for running parallel WRF-Hydro and the workers used for running parallel PEST, Tests 4 and 5 use different number of nodes for each worker to run WRF-Hydro in parallel. Explicitly, Test 4 uses four nodes per worker, and Test 5 uses six nodes per worker. Both tests use six workers for running the parallel PEST. The maximum number of lambda-testing runs undertaken per iteration is set to five for both Tests 4 and 5. Note that the time costs in Table 3 are limited to only one iteration. Conducting more iterations will increase the cost of wall-clock time and computing, but will not change the conclusion for the scale-up capability and computational benefits for HPC-enabled parallel PEST linked to WRF-hydro.

PEST needs to run the WRF-Hydro model at least as many times as the number of calibrated parameters (22 here). In fact, PEST runs the model 23 times in the first round (or the first iteration) with initial parameter values and for the first Jacobian matrix. From the second iteration, it runs the model 22 times to calculate ~~the Jacobin~~Jacobian matrix. Therefore, ~~having~~if there are fewer than 23 workers, the time cost for the first round of Jacobian matrix calculation will increase accordingly. For example, as shown in Fig. 4a, when we assign 12 (and 6) workers to parallel PEST, the time cost for calculating the Jacobian matrix is increased by a factor of 2 (and 4) compared with the time cost of using 23 workers. The time cost for the parameter upgrade stays similar for the three experiments because only one cycle of WRF-hydro simulation is conducted

to test the Marquardt lambdas. As a result, the total time cost for Test 2 is ~1.5 times more than ~~64 slaves~~that for Test 1, and the total time cost for Test 3 is ~1.5 times more than that for Test 2 (Fig. 4b). By extrapolating the speedup curve shown in Fig. 4a and Fig. 4b, we expect the total time cost to be ~1516 minutes when using only one worker (or sequential mode), which is about 15 times slower compared with running the PEST in parallel mode using 23 workers. For this particular study with 22 adjustable parameters, we expect the time cost most likely to stay the same even if one increases the number of workers to more than 23, because PEST runs WRF-Hydro only 23 or 22 times for each iteration. Assigning more workers for this particular study would most likely render some ~~slaves~~workers idle and is not an efficient use of computing resources. PEST may run WRF-Hydro more than 22 times (e.g., 44 times) if higher-order finite differences are employed. In this case, assigning more workers (e.g. 45 workers) may further speed up the calibration process. On the other hand, for the same case study and using the same number of nodes for running parallel WRF-Hydro, we can estimate the computing speedup by assuming an increase in the number of calibrated parameters to 50. This would be the case, for example, to evaluate model sensitiveness to the physics in Noah-MP or the spatial variabilities of certain parameters. We then expect to use 51 workers to achieve the best computing performance for parallel PEST. This would then be 28–30 times faster than running PEST using one worker (or in sequential mode). Similarly, if 100 parameters were used for the calibration for the same case study, a factor of up to 60 speedup in the calibration process would be achieved by running HPC-enabled parallel PEST.

In addition, by increasing the number of nodes for each worker to conduct WRF-Hydro (Tests 3, 4, and 5), the time cost for the entire calibration process is significantly reduced (Figs. 4c and 4d). Specifically, the WRF-hydro scales up well when using four and six nodes compared with using two nodes per worker for running the WRF-Hydro. Both the time spent on calculating the Jacobian matrix and the time spent on testing the parameter upgrades are decreased by 49% and 67%, respectively, when using four and six nodes. Therefore, the total time spent is also decreased when using more nodes for each worker (see Table 3). Increasing the number of nodes to eight for each worker will most likely further decrease the time cost by 70–75% compared with using only two nodes per worker. Moreover, if one has a larger study area such as the entire contiguous United States, we expect the WRF-Hydro to have an even better scale-up capability (e.g., on dozens of nodes) than this study. Overall, based on the experiments we conduct here, using 23 workers for

parallel PEST and six nodes for each worker to run parallel WRF-Hydro would cost the least wall-clock time—about 32 min for one iteration for this particular study.

4.3 Evaluation of spatial transferability of the calibrated parameters

To assess the transferability of the calibrated parameters, we apply the optimum parameters obtained from the calibration for the four stations (black circles) in Fig. 1 to another set of four stations (crosses in Fig. 1) in the study area. All four sites are located on relatively small rivers, so the lag time between precipitation peak and the discharge peak are much shorter than that for the stations on the lower part of MRB (e.g., Stations 2, 3, and 4). The assessment compares the observed discharge with the closest grid cells from the discharge output of WRF-Hydro. ~~5.2~~ Figure 5 shows the observed and modeled discharge using default and the optimum parameters. Overall, WRF-Hydro's default parameters underestimate the discharge and misrepresent the timing of discharge peaks compared with observations over the four assessed stations (Stations 5, 6, 7, and 8). By using the calibrated parameters from other sites over the area, the model results increase the discharge and shift the hydrograph shape so they are much closer to the observations than model results using default parameters. The absolute error of simulated discharge decreases by 13.1%, 38.3%, and 71.6%, respectively, over Stations 6 through 8 (Station 5 shows a 6% increase of absolute error), compared with the default simulated discharge. We also find that using the SVD-based regularization for the PEST calibration captures the timing of discharge peak better than using the estimation mode, which is one-day earlier than the observations reaching the discharge peak.

5 Summary and discussion

WRF-Hydro is a new, and perhaps the first practical, computer code that can run on ~~HPCs~~HPC systems and can model the entire hydrological cycle using physics-based ~~sub-models~~submodels and ~~very~~ high-resolution input datasets (e.g., radar). The hydrological community has desired this capability for decades, although it requires intensive computing resources. Thus, the calibration of this model would ideally be conducted on HPCs in parallel as well, especially when the model covers a large domain rather than the basin scale. This study ports an independent model calibration tool, parallel PEST, to HPC clusters and links it to WRF-Hydro to help WRF-Hydro

users calibrate the model within a much shorter wall-clock time period. ~~This tool's uniqueness lies in its flexibility~~The bridge we build here (between parallel PEST and ~~robustness to calibration any parameters in~~ WRF-Hydro~~. It is also unique in its use~~ on the basis of ~~two levels of parallelization across many slaves running PEST, with each slave running a simulation of WRF-Hydro. The calibration tool presented in this study also applies~~HPC systems) can be applied to any other hydrological models and ~~similar earth~~Earth system models that use ~~parameterization to~~parameterizations to represent model physics. We present the operational feasibility of the HPC-enabled parallel PEST by evaluating the performance of calibrated WRF-Hydro against observation in hydrograph features such as volume and timing of flood events. We examine the scale-up capability and computational benefits of the tool by assigning different computing resource for PEST and for WRF-Hydro. While this study presents the optimum parameters identified from the calibration of ~~this~~the particular ~~study case and area, but the calibrated~~flood event, the parameters can be significantly different if one uses different physics, such as exponential storage-discharge function for a groundwater model~~, or reach-based channel routing.~~ or reach-based channel routing. Our preliminary testing shows that using exponential storage-discharge function with the default parameters provided by WRF-Hydro, the modeled discharge was larger than that of observations. Thus, the calibration will need to adjust the parameters to reduce the discharge. Our study finds that for calibrating 22 parameters, using the same computing resource for running WRF-hydro, the HPC-enabled PEST calibration tool can speed up WRF-Hydro calibration by a factor of 15, compared with running PEST in sequential mode. The speedup factor can be larger when the number of parameters needing calibration is higher (e.g., 50 or 100).

~~We apply the HPC-enabled parallel PEST to WRF-Hydro to investigated a major flood event that occurred over the Midwestern United States in April 2013. Our precipitation inputs were derived from a radar, gauge, and satellite rainfall product named Stage IV. The calibrated parameters include Manning's roughness coefficients for both channels and land-use types, deep drainage, the infiltration scaling parameter, and saturated soil lateral conductivity. We evaluated the performance of WRF-Hydro in hydrograph features such as volume and timing of flood events. We also assessed the spatial transferability of the calibrated parameters in the study area.~~

The following are ~~the primary findings of~~ several key points that we would like to mention to inform future studies:

1. In this study~~:~~, we consider using the prior or regularization information only for the parameters that we calibrate. As is the case with solving inverse problems, prior information is added to improve the smoothness of the solutions. In order to build a more comprehensive calibration, an important aspect that can be considered is to enrich the prior with the available historical data. For example, in this particular case, one can use the historical observation data (e.g., April and May from the past few years) to enrich the prior information for the parameters. Hence, the regularization objective function in PEST will constitute not only the discrepancies between parameters and their "current estimates" but also the discrepancies between WRF-Hydro simulations and preferred values (which is the observed time series of historical discharge). Additionally, one can use the pilot points technique described by Doherty (2005) in conjunction with parameter estimation to add more flexibility to the calibration process. This will be potentially beneficial in improving the predictions.

~~1. For this particular study, the HPC-enabled PEST calibration tool can speed up WRF-Hydro calibration by a factor of 30, compared to a serial calibration procedure.~~

~~2. Calibrated WRF-Hydro improves the modeled hydrographs compared to the default model results. The RMSE of discharge over the three large stations are reduced by 76–86% with calibration for the validation period.~~

~~3. Although the calibration period in this study is relatively short, we found that the longer the calibration period is, the better the model results are when compared to observations. It is difficult to precisely define what length of data is sufficient to identify model parameters so that they can also be used for other periods, because different models have different levels of complexity and different catchments have different information content in each year of hydrological record. Because of the heavy computation load for the calibration of WRF-Hydro, it would be challenging to calibrate yearlong time series.~~

~~4. Although there are inaccuracies when mapping the USGS stations onto the river systems, we found the WRF-Hydro calibrated parameters are helpful for nearby locations. The absolute bias over the four assessed stations decreased by 12–50% as a result of using the~~

~~calibrated parameters, compared to using the default parameters for their simulations. The mapping issue are often random, or non-systematic; there is no generalizable way to automate the correction procedure with a high degree of fidelity. Manual manipulation or specification of the data is often required (Sampson and Gochis 2018).~~

~~5. Using different groundwater models can generate very different results and will require a completely different set of parameters for WRF-Hydro to model the observed discharge. Our preliminary testing shows that, using exponential storage-discharge function with the default parameters provided by WRF-Hydro, the modeled discharge was larger than observations. Thus, the calibration will need to adjust the parameters to reduce the discharge.~~

2. To focus on our main goal, we calibrate only the parameters in lookup tables. However, we acknowledge that using a single value to represent a physics for a large domain could be problematic, especially we expect the HPC-enabled parallel PEST to execute with WRF-Hydro for large domains. This situation often needs parameter regionalization. For example, WRF-Hydro v5 has many spatially distributed parameters available, such as the overland flow roughness scaling factor (OVROUGHRTFAC), the factor of maximum retention depth (RETDEPRTFAC), and the soil-related parameters (when compiled with SPATIAL_SOIL=1). Calibrating these spatial parameters based on grid scale (e.g., catchments) rather than a single value will give the model more flexibility and thus better fit the observations (Hundecha and Bardossy, 2004; Wagener and Wheater, 2006). In practice, for example, one can include regional OVROUGHRTFACs (e.g., their lower/upper bounds, and default values) in the PEST control file based on catchments. However, the selection of the locations and sizes of catchment may introduce significant uncertainties to the calibration results, which require systematic and comprehensive investigation and understanding of the study area.
3. This study is limited to calibrating the observed streamflow only based on the format of one of WRF-Hydro model outputs for individual station or point (frxst_pts_out.txt). It is feasible, however, to calibrate other variables as long as the observation data is available. For example, one can either find the closest point from the gridded dataset to the observation location and then compare that model grid to observations; or one can change

the WRF-Hydro input/output code to output other variables in the frxst_pts_out.txt file, so they can still use the same interface we developed here to calibrate other variables instead in addition to the discharge.

4. The optimal parameter set obtained from this study is from the 5th iteration of parallel PEST by testing five Marquardt lambdas. Testing different number of lambdas or calibrating different number of parameters may generate a different set of optimal parameters. These parameter sets can all make physical sense and be equally good for reproducing observed discharges. This problem is named equifinality (Beven and Freer, 2001; Savenije, 2001), which is an important source of model uncertainty. To reduce the model uncertainty through reducing the equifinality, hydrologists carry out additional modelling objective for model evaluation to find more useful parameter sets (Mo and Beven, 2004; Gallart et al., 2007). Alternatively, inspired by No. 3 discussed above, one can calibrate the WRF-hydro model based on more than one variables, such as discharge and soil moisture (or heat flux or water table depth) to reduce the number of optimal parameter sets, and thus reduce the model uncertainty of predictions for these variables.
5. While this study ported the parallel PEST to HPC system and linked it to WRF-Hydro, we note that BEOPEST is available in the PEST family. BEOPEST has the same functionality as parallel PEST but uses a different approach for communication between master and workers. Working with HPC-enabled BEOPEST may save total time cost since BEOPEST uses the Transmission Control Protocol (TCP) and the Internet Protocol (IP) instead of message files (reading input and writing output between master and works) for communication. We expect it to be relatively straightforward to use BEOPEST to calibrate WRF-hydro on HPCs since the interface remains similar, except one needs to copy the template and instruction files in addition to the global files (see Section 3.1) into each working folder.

Data and Code availability. The observed river discharge is downloaded from the USGS Surface-Water Data website, available at https://waterdata.usgs.gov/nwis/sw. The Stage IV precipitation data were downloaded from https://data.eol.ucar.edu/dataset/21.093. PEST was downloaded from http://www.pesthomepage.org/Downloads.php. We use the Unix PEST version 13.6. The scripts and files that are developed in this study and required by PEST for calibrating WRF-Hydro are

available at ~~https://www.zenodo.org/record/1490230#.W_XI6TFRdhE. DOI: 10.5281/zenodo.1490230.~~http://doi.org/10.5281/zenodo.2588506.

Author contributions. JW proposed the project and developed the study case in WRF and WRF-Hydro. CW developed the scripts/code to port the parallel PEST to DOE supercomputers and adapt it to work with WRF-Hydro. VR provided important input for the regularization calibration method. AO operated the ArcGIS tool to delineate the high-~~-~~resolution grid cells to include stream channel network, open water, and groundwater/baseflow basins. RK provide high-level guidance and insight for the entire project. All authors commented on this manuscript.

Competing interests. The authors declare that they have no conflict of interest

~~*Acknowledgements.*~~*Acknowledgments.* This work is supported under a Laboratory Directed Research and Development (LDRD) Program at Argonne National Laboratory, through U.S. Department of Energy (DOE) contract DE-AC02-06CH11357. Computational resources are provided by the DOE-supported National Energy Research Scientific Computing Center, Argonne National Laboratory Computing Resource Center, and Argonne Leadership Computing Facility. Our special thanks to the PEST ~~developer John Doherty~~developers and ~~the~~ entire WRF-Hydro team, especially Kevin Sampson~~,~~ for his guidance on the ArcGIS tool. We gratefully thank the two reviewers for their valuable comments and suggestions, which tremendously improved this manuscript.

References

Arnault, J., Wagner, S., Rummler, T., Fersch, B., Bliefernicht, J., Andresen, S., and Kunstmann, H.: Role of runoff–infiltration partitioning and resolved overland flow on land–atmosphere feedbacks: A case study with the WRF-Hydro coupled modeling system for West Africa, J. Hydrometeorol., 17, 1489–1516, 2016.

Campos, E~~.~~., and Wang, J.: Numerical simulation and analysis of the April 2013 Chicago Floods, J. Hydrol., 531, 454–474, 2015.

Chen, F. and Dudhia, J.: Coupling an advanced land surface-hydrology model with the Penn State-NCAR MM5 modeling system~~.~~, Part I: Model implementation and sensitivity, Mon. Weather Rev., 129, 569–585, 2001.

Cuntz, M., Mai, J., Samaniego, L., Clark, M., Wulfmeyer, V., Branch, O., Attinger, S., and Thober, S.: The impact of standard and hard-coded parameters on the hydrologic fluxes in the Noah-MP land surface model, J. Geophys. Res. Atmos., 121, 10,676–10,700, doi:10.1002/2016JD025097, 2016.

Doherty, J.: PEST: Model Independent Parameter Estimation, User Manual, 6th ed., Watermark Numerical Computing, Brisbane, Queensland, Australia, 2016.

Doherty, J.: Ground water model calibration using pilot points and regularization, Groundwater, 41(2), 170–177, 2005.

Gallart, F., Latron, J., Llorens, P., and Beven, K. J.: Using internal catchment information to reduce the uncertainty of discharge and baseflow predictions. Adv. Water Resour. 30(4), 808–823, 2007.

Getirana, A. C. V.: Integrating spatial altimetry data into the automatic calibration of hydrological models, J. Hydrol., 387 (3-4), 244–255, doi: 10.1016/j.jhydrol.2010.04.013, 2010.

Gochis, D. J~~. and Chen, F.: Hydrological enhancements to the community Noah land surface model, NCAR Technical Note NCAR/TN-454+STR, doi: 10.5065/D60P0X00, 2003.~~

~~Gochis, D.J., M.~~., Barlage, ~~A.~~M., Dugger, ~~K.~~A., FitzGerald, ~~L.~~K., Karsten, ~~M.~~L., McAllister, ~~J.~~M., McCreight, J~~.~~., Mills, ~~A.~~J., RafieeiNasab, ~~L.~~A., Read, ~~K.~~L., Sampson, ~~D.~~K., Yates, ~~W.~~D., and Yu, W.:~~(2018).~~ The WRF-Hydro modeling system technical description, (Version 5.0). NCAR Technical Note. 107 pages. Available online at: https://ral.ucar.edu/sites/default/files/public/WRFHydroV5TechnicalDescription.pdf~~.~~, 2018.

Hundecha, Y., and Bárdossy, A.: Modeling of the effect of land use changes on the runoff generation of a river basin through parameter regionalization of a watershed model, J. Hydrol., 292, 281–295, 2004.

Kerandi, N., Arnault, J., Laux, P., Wagner, S., Kitheka, J., and Kunstmann, H.: Joint atmospheric-terrestrial water balances for East Africa: A WRF-Hydro case study for the upper Tana River basin, Theor. Appl. Climatol., 131, 1337–1355, doi: 10.1007/s00704-017-2050-8, 2018.

~~Levenberg, K.: A method for the solution of certain non-linear problems in least squares, Q. Appl. Math., 2(2), 164–168, 1944.~~

Lin, Y~~.~~., and Mitchell, K. E.: The NCEP stage II/IV hourly precipitation analyses: Development and applications, Preprints, 19th Conf. on Hydrology, San Diego, CA, Amer. Meteor. Soc., 1.2., 2005.

Madsen, H.: Automatic calibration of a conceptual rainfall–runoff model using multiple objectives, J. Hydrol., 235, 276–288, 2000.

~~Marquardt, D. W.: An algorithm for least-squares estimation of non-linear parameters, J. Soc. Indust. Appl. Math., 11(2), 431–441, 1963.~~

Mo, X., and Beven, K.: Multi-objective parameter conditioning of a three-source wheat canopy model. Agricultural & Forest Meteorol. 122(1–2), 39–63, 2004.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., and Veith, T. L.: Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, Transactions of the ASABE, 50 (3), 885–900, 2007.

Nash, J. E~~.~~., and Sutcliffe, J. V.: River flow forecasting through conceptual models, part I ~~—~~ A discussion of principles, J. Hydrol., 10(3), 282–290, doi: 10.1016/0022-1694(70)90255-6, 1970.

Niu, G.-Y., Yang, Z.-L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., Kumar, A., Manning, K., Niyogi, D., Rosero, E., Tewari, M., and Xia, Y.: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements, J. Geophys. Res., 116, D12109, doi: 10.1029/2010JD015139, 2011.

NWS (National Weather Service): Record ~~River Flooding~~river flooding of April 2013, https://www.weather.gov/ilx/apr2013flooding, 2013.

~~Papathanasiou, C., Makropoulos, C., and Mimikou, M.: Hydrological modelling for flood forecasting: Calibrating the post-fire initial conditions, J. Hydrol., 529, 1838–1850, doi: 10.1016/j.jhydrol.2015.07.038, 2015.~~

Prat, O. P~~.~~., and Nelson, B. R.: Evaluation of precipitation estimates over CONUS derived from satellite, radar, and rain gauge data sets at daily to annual scales (2002-2012), Hydrol. Earth Syst. Sci., 19, 2037–2056, doi: 10.5194/hess-19-2037-2015, 2015.

Sampson, K~~.~~., and Gochis, D.: WRF Hydro GIS Pre-~~Processing Tools~~processing tools, Version 5.0 Documentation, 2018.

Sapiano, M. R. P~~.~~., and Arkin, P.A.: An intercomparison and validation of high-resolution satellite precipitation estimates with 3-hourly gauge data, J. Hydrometeor., 10, 149–166, doi: 10.1175/2008JHM1052.1, 2009.

Senatore, A., Mendicino, G., Gochis, D. J., Yu, W., Yates, D. N., and Kunstmann, H.: Fully coupled atmosphere-hydrology simulations for the central Mediterranean: Impact of enhanced hydrological parameterization for short and long time scales, J. Adv. Model. Earth Syst., 7(4), 1693–1715, doi: 10.1002/2015MS000510, 2015.

Soong, D. T., Prater, C. D., Halfar, T. M., and Wobig, L. A.: Manning's roughness coefficients for Illinois streams, U.S. Geological Survey Data Series 668, 2012.

Tonkin, M. J., and Doherty, J.: A hybrid regularized inversion methodology for highly parameterized environmental models, Water Resource Research, 41, W10412, doi:10.1029/2005WR003995, 2005.

Wagenera, T., and Wheater, H. S.: Parameter estimation and regionalization for continuous rainfall-runoff models including uncertainty, J. Hydrol., 320, 132–154, 2006.

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., Luo, L., Alonge, C., Wei, H., Meng, J., Livneh, B., Lettenmaier, D., Koren, V., Duan, Q., Mo, K., Fan, Y., and Mocko, D.: Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2~~).~~), 1: Intercomparison and application of model products, J. Geophys. Res., 117, D03109, doi: 10.1029/2011JD016048, 2012a.

Xia, Y., Mitchell, K., Ek, M., Cosgrove, B., Sheffield, J., Luo, L., Alonge, C., Wei, H., Meng, J., Livneh, B., Duan, Q., and Lohmann, D.: Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2). 2. Validation of model-simulated streamflow, J. Geophys. Res., 117, D03110, doi: 10.1029/2011JD016051, 2012b.

Yucel, I., Onen, A. Yilmaz, K. K., and Gochis, D. J.: Calibration and evaluation of a flood forecasting system: Utility of numerical weather prediction model, data assimilation and satellite-based rainfall, J. Hydrol., 523, 49–66, 2015.

~~Zanon, F., Borga, M., Zoccatelli, D., Marchi, L., Gaume, E., Bonnifait, L., Delrieu, G.: Hydrological analysis of a flash flood across a climatic and geologic gradient: The September 18, 2007 event in Western Slovenia, J. Hydrol., 394, 182–197, doi: 10.1016/j.jhydrol.2010.08.020, 2010.~~

Table 1: Calibrated 22 parameters and the optimum parameters found after five iterations.

Calibrated Parameter	Default	Lower Bound	Upper Bound	Optimum Parameter
mannn1	0.55	0.35	0.6	0.517599
mannn2	0.35	0.15	0.35	0.153894
mannn3	0.15	0.08	0.15	8.00E-02
mannn4	0.1	0.05	0.15	5.00E-02
mannn5	7.00E-02	0.02	0.1	6.677379E-02
mannn6	5.00E-02	0.015	0.1	1.628244E-02
mannn7	4.00E-02	0.01	0.08	1.298054E-02
mannn8	3.00E-02	0.005	0.06	5.00E-03
xslope1	0.1	1.00E-04	1	0.496680
refdk	2.00E-06	1.00E-08	1.00E-05	2.899043E-07
refkdt	1	0.01	5	1.66664
ovn1 (urban)	2.50E-02	0.005	0.06	6.00E-02
ovn2 (dry crop)	3.50E-02	0.015	0.06	1.50E-02
ovn3 (irrigated crop)	3.50E-02	0.015	0.06	1.50E-02
ovn5 (crop/grass)	3.50E-02	0.015	0.06	2.822497E-02
ovn6 (crop/wood)	6.80E-02	0.035	0.25	4.568903E-02
ovn7 (grass)	5.50E-02	0.015	0.25	1.50E-02
ovn10 (savanna)	5.50E-02	0.015	0.3	1.50E-02
ovn11 (deciduous forest)	0.2	0.1	0.3	0.30
ovn14 (evergreen forest)	0.2	0.1	0.3	0.164557
ovn15 (mixed forest)	0.2	0.1	0.3	0.112134
ovn16 (water)	5.00E-03	0.001	0.01	1.00E-02

Table 2: Statistics of model performance using optimum and default (in parentheses) parameters for Stations 1–4 during the calibration and validation period.[a]

Statistics	Station 1	Station 2	Station 3	Station 4
		Calibration		
NSE	**0.87 (0.73)**	**0.64 (-54.4)**	**0.05 (-157.3)**	**-58.78 (-1316.9)**
RMSE	**48.1 (69.3)**	**318.2 (3925.2)**	**308.7 (3981.3)**	**934.6 (4391.3)**
PCC	**0.95 (0.91)**	0.87 (0.92)	**0.91 (0.87)**	0.53 (0.66)
		Validation		
NSE	**0.83 (0.41)**	**-0.08 (-3.5)**	**-0.08 (-27.4)**	**-0.12 (-3.33)**
RMSE	**259.9 (487.3)**	**3264.3 (6670.1)**	**3170.1 (16305.7)**	**3283.9 (6854.3)**
PCC	**0.83 (0.8)**	**0.98 (0.69)**	**0.29 (0.19)**	**0.94 (0.64)**

[a] The calibration period is 3 days (April 9–11) and includes 22 parameters. The validation period is April 12–24. Bold typeface indicates the calibrated model results are closer to observations compared with the default model results. NSE and PCC are unitless; RMSE is in m^3/sec.

Table 3. Experiments designed to test the scale-up capability and computational benefits of HPC-enabled parallel PEST linked to WRF-Hydro.

Test	No. of Workers	No. of Lamdas	No. of Nodes for Each Worker	Total Time Cost (min)	Time Cost for Calculating Jacobian Matrix	Time Cost for Testing Parameter Upgrades
Test 1	23	15	2	103	52	51
Test 2	12	10	2	150	102	48
Test 3	6	5	2	264	211	53
Test 4	6	5	4	131	107	24
Test 5	6	5	6	86	70	16

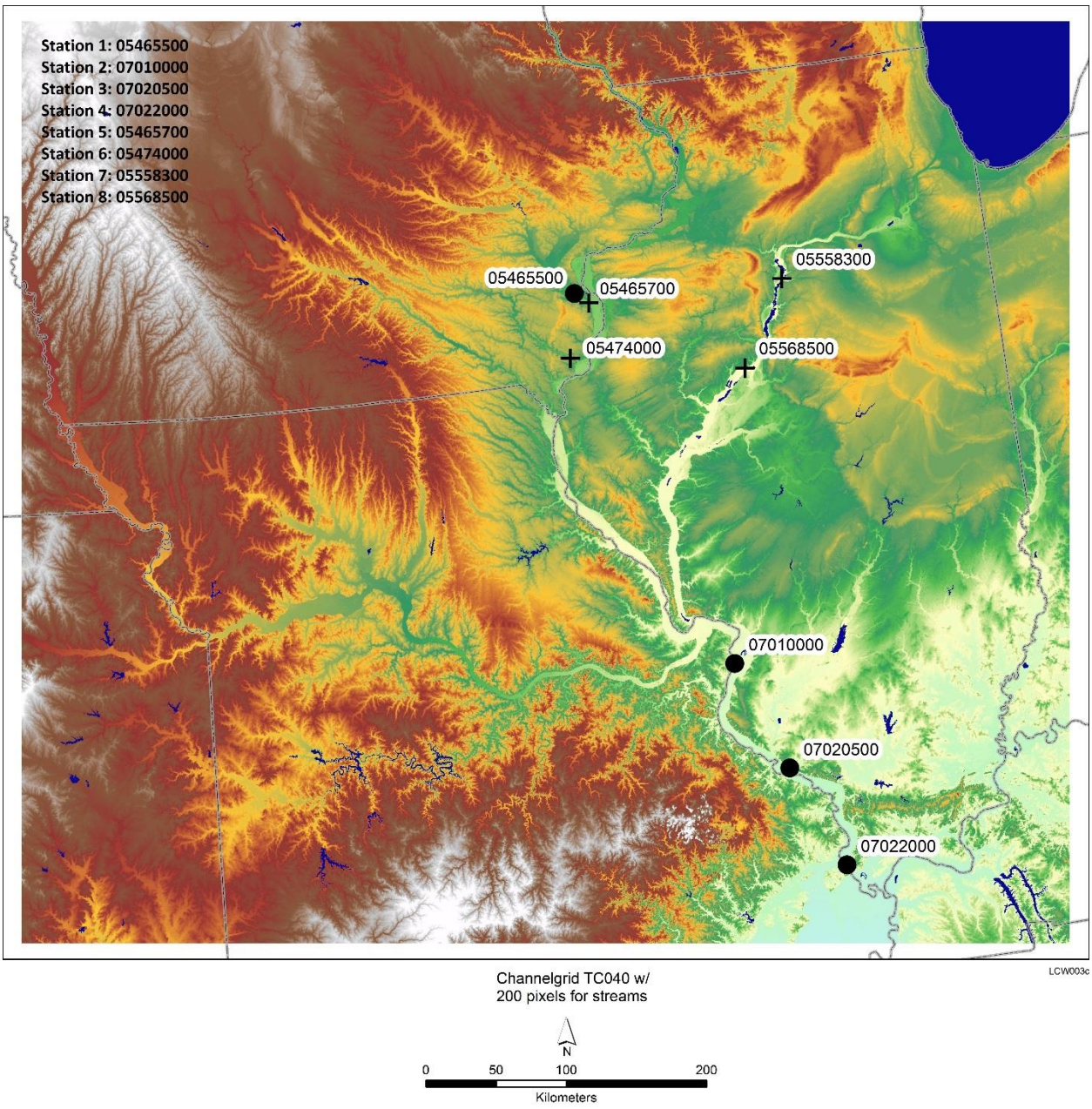


Figure 1: Eight USGS sites over the study area (750 km x 660 km). The four circles are sites that are used for calibrations; the four crosses are sites that are used for transferability assessment. USGS site numbers corresponding to the site index used in this study are listed on the top left corner of the map.

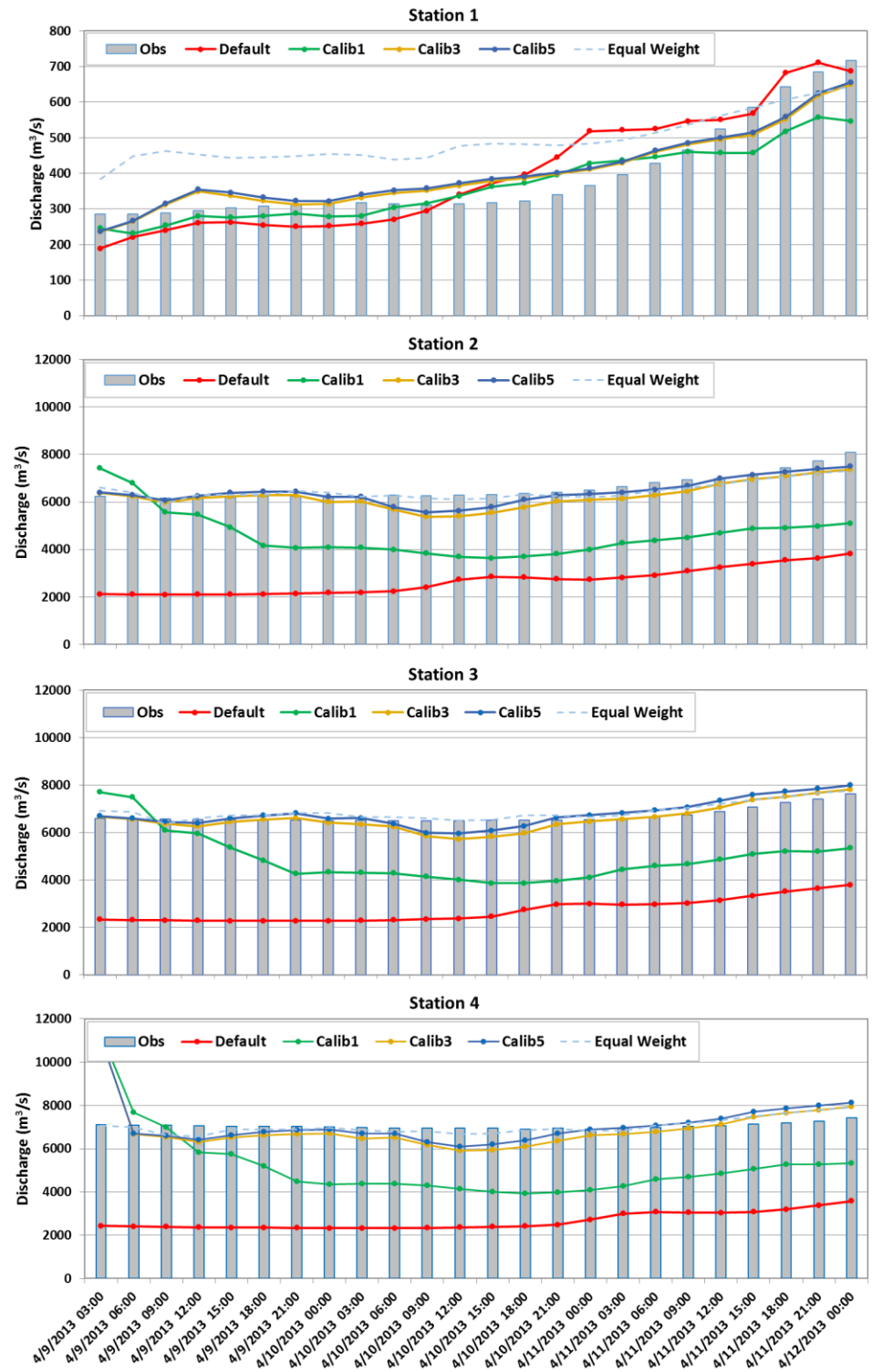
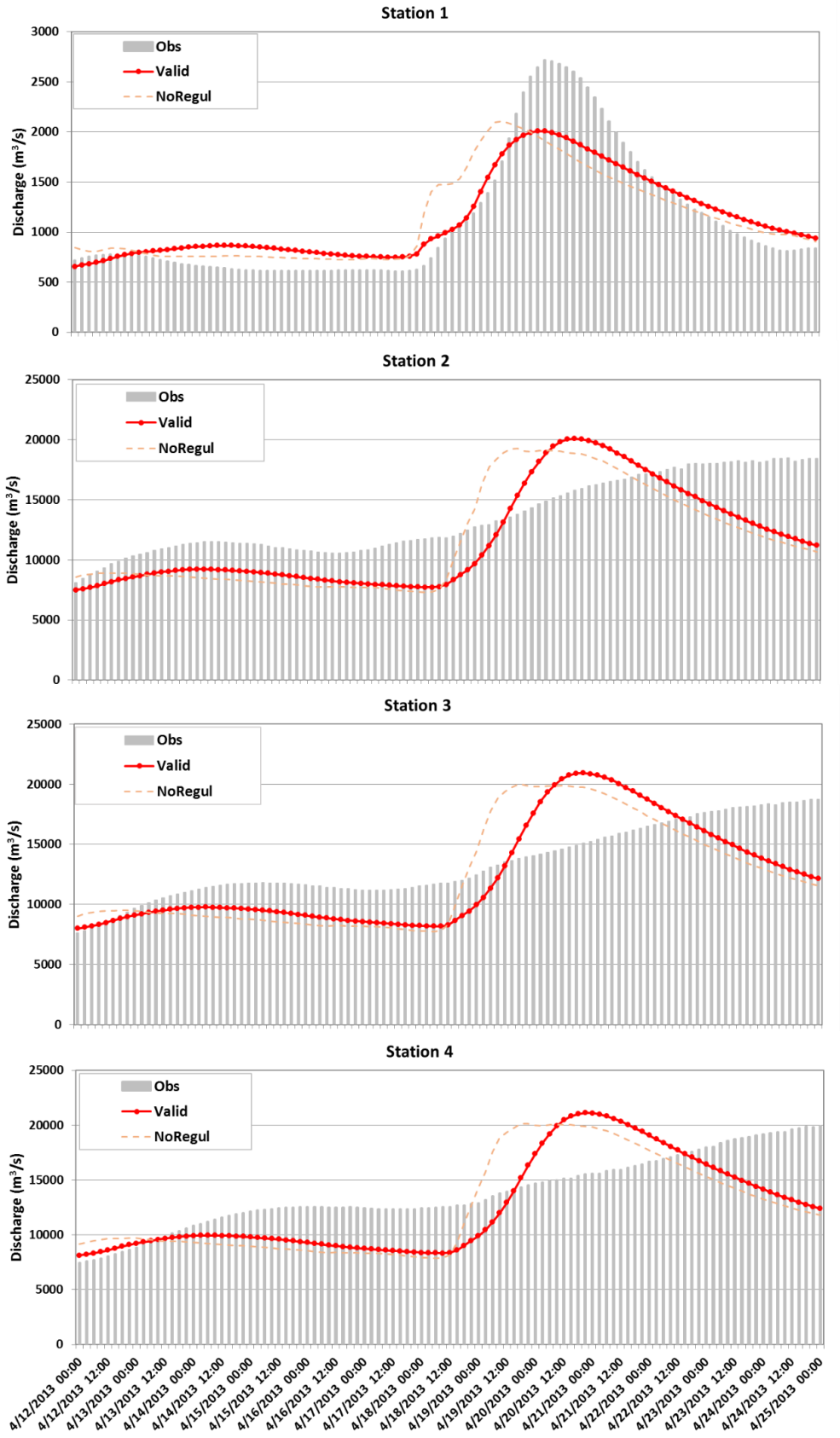


Figure 2: Observed and modeled discharge (m^3/sec) using default and calibrated parameters during a 3-day calibration period (April 9–11, 2013) over the four stations indicated by the black circles in Fig. 1. The calibrations shown in solid lines are conducted by using SVD-

based regularization and a higher weight for Station 1. The dashed line is the optimum result calibrated by using equal weight for all four sites.


Station 1
Obs
Valid
NoRegul
Discharge (m³/s)
3000
2500
2000
1500
1000
500
0
Station 2
Obs
Valid
NoRegul
Discharge (m³/s)
25000
20000
15000
10000
5000
0
Station 3
Obs
Valid
NoRegul
Discharge (m³/s)
25000
20000
15000
10000
5000
0
Station 4
Obs
Valid
NoRegul
Discharge (m³/s)
25000
20000
15000
10000
5000
0
4/12/2013 00:00
4/12/2013 12:00
4/13/2013 00:00
4/13/2013 12:00
4/14/2013 00:00
4/14/2013 12:00
4/15/2013 00:00
4/15/2013 12:00
4/16/2013 00:00
4/16/2013 12:00
4/17/2013 00:00
4/17/2013 12:00
4/18/2013 00:00
4/18/2013 12:00
4/19/2013 00:00
4/19/2013 12:00
4/20/2013 00:00
4/20/2013 12:00
4/21/2013 00:00
4/21/2013 12:00
4/22/2013 00:00
4/22/2013 12:00
4/23/2013 00:00
4/23/2013 12:00
4/24/2013 00:00
4/24/2013 12:00
4/25/2013 00:00

1

Figure 3: Observed and modeled discharge (m^3/sec) during a validation period (April 12–24, 2013) using optimum parameters identified from a 3-day calibration over the four stations indicated by black circles in Fig. 1. The solid line uses the optimum parameters that identified by PEST with SVD-based regularization and a higher weight for Station 1. The dashed line uses the optimum result calibrated by using estimation mode (no regularization).

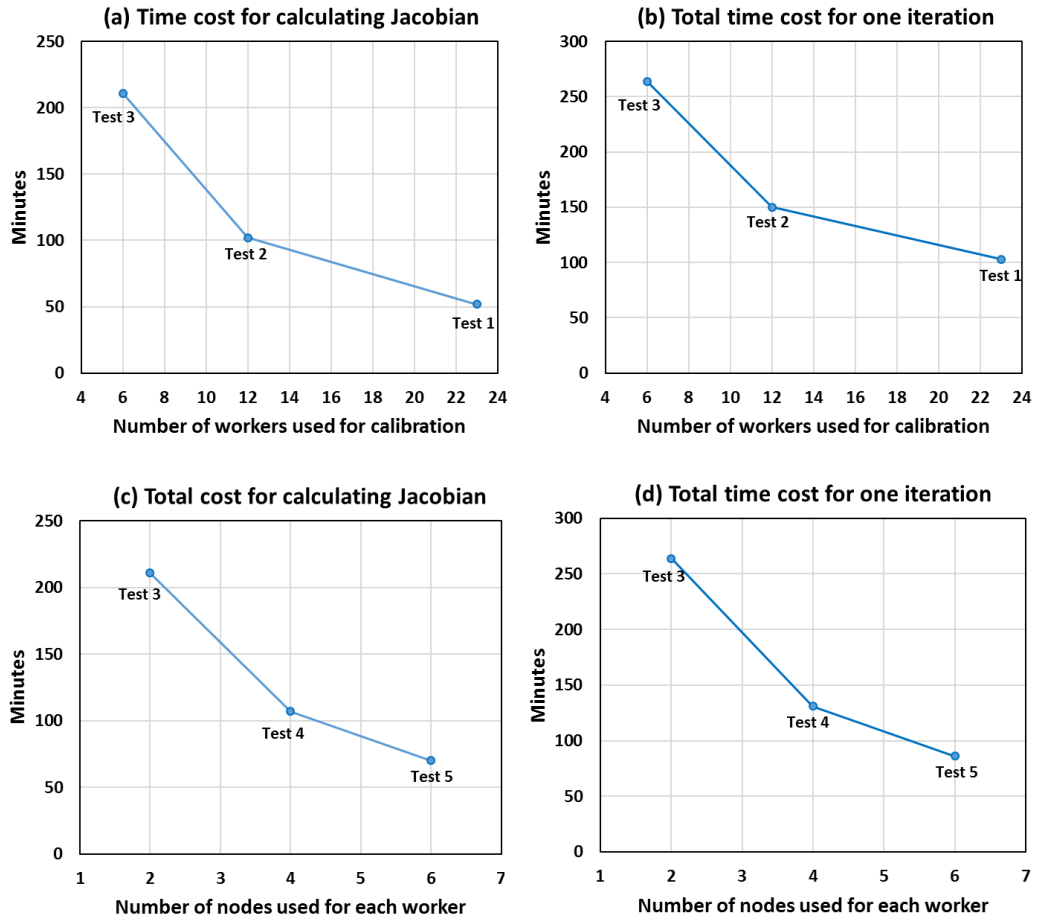


Figure 4. Time cost for calculating Jacobian matrix and total time cost for one iteration for the five experiments (Table 3) using different number of workers to conduct PEST (a, b) and different number of nodes for each worker (c, d) to conduct WRF-Hydro.

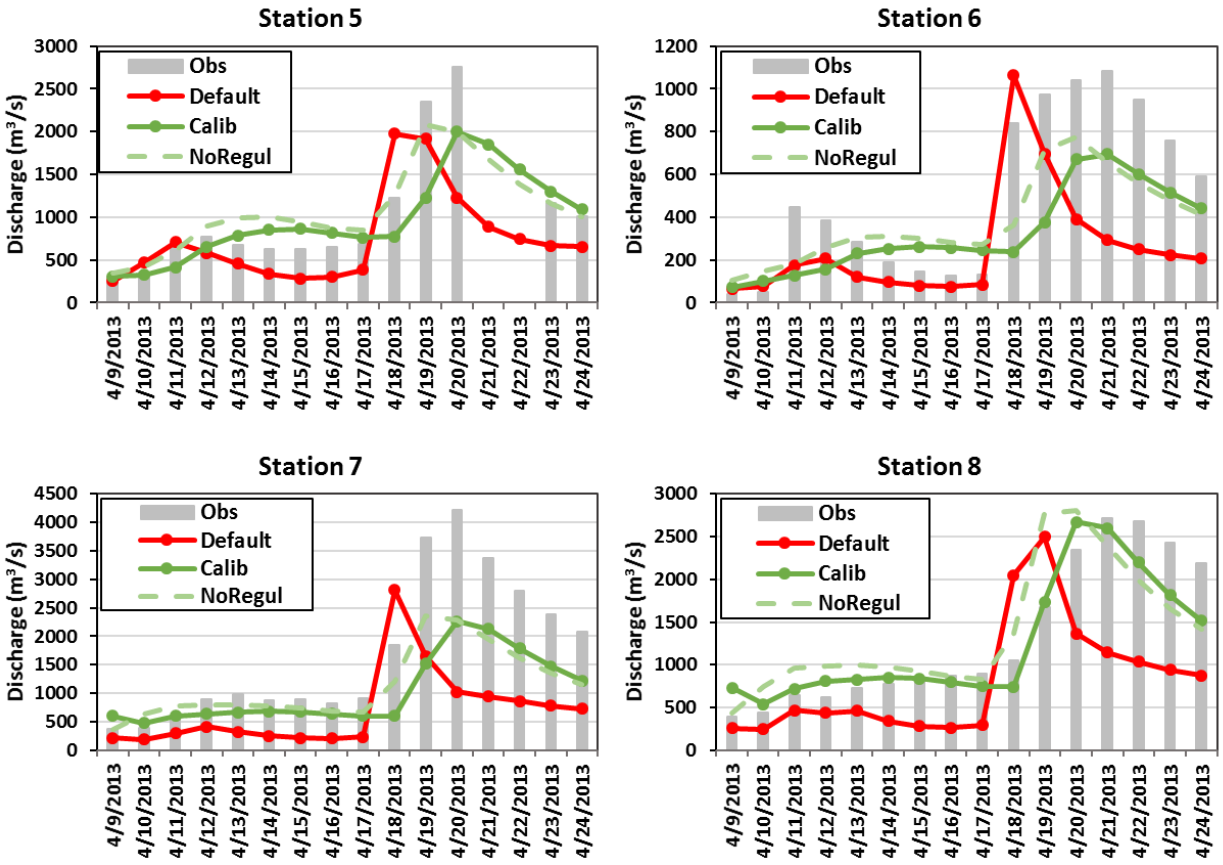


Figure 5: Observed and modeled daily averaged discharge (m^3/sec) from April 9–24 using default and the optimum parameters (shown in Table 1) identified by the 3-day calibration over four stations that are in the study area (indicated by crosses in Fig. 1). The calibrations shown in solid lines are conducted by using SVD-based regularization and a higher weight for Station 1. The dashed line is the optimum result calibrated by using estimation mode (no regularization).

A parallel workflow implementation for PEST version 13.6 in high-performance computing for WRF-Hydro version 5.0: a case study over the midwestern United States

[1]Jiali Wang, [1]Cheng Wang, [2]Vishwas Rao, [1]Andrew Orr, [1]Rao Kotamarthi

[1]Argonne National Laboratory, Environmental Science Division, 9700 South Cass Avenue, Lemont, IL 60439, USA

[2]Argonne National Laboratory, Mathematics and Computer Science Division, 9700 South Cass Avenue, Lemont, IL 60439, USA

Correspondence to: Jiali Wang (jialiwang@anl.gov)

Abstract. The Weather Research and Forecasting Hydrological (WRF-Hydro) system is a state-of-the-art numerical model that models the entire hydrological cycle based on physical principles. As with other hydrological models, WRF-Hydro parameterizes many physical processes. Hence, WRF-Hydro needs to be calibrated to optimize its output with respect to observations for the application region. When applied to a relatively large domain, both WRF-Hydro simulations and calibrations require intensive computing resources and are best performed on multimode, multicore high-performance computing (HPC) systems. Typically, each physics-based model requires a calibration process that works specifically with that model and is not transferrable to a different process or model. The parameter estimation tool (PEST) is a flexible and generic calibration tool that can be used in principle to calibrate any of these models. In its existing configuration, however, PEST is not designed to work on the current generation of massively parallel HPC clusters. To address this issue, we ported the parallel PEST to HPCs and adapted it to work with WRF-Hydro. The porting involved writing scripts to modify the workflow for different workload managers and job schedulers, as well as developing code to connect parallel PEST to WRF-Hydro. To test the operational feasibility and the potential computational benefits of this first-of-its-kind HPC-enabled parallel PEST, we developed a case study using a flood in the midwestern United States in 2013. Results on a problem involving calibration of 22 parameters show that on the same computing resource used for parallel WRF-Hydro, the HPC-enabled parallel PEST can speed the calibration process by a factor of up to 15 compared with commonly used

PEST in sequential mode. The speedup factor is expected to be greater with a larger calibration problem (e.g., more parameters to be calibrated or a larger size of study area).

1 Introduction

Physically based hydrological models contain detailed physical mechanisms to model the hydrological cycle, but many complex physical processes in these models are parameterized. For example, the state-of-the-art Weather Research and Forecasting Hydrological (WRF-Hydro) modeling system (Gochis et al., 2015) has dozens of parameters that can be land- and river-type dependent and are typically specified in lookup tables. Therefore, these hydrological models need to be calibrated before they can be applied to research over different regions. In this context, calibration refers to adjusting the values of the model parameters so that the model can closely match the behavior of the real system it represents. In some cases, the appropriate value for a model parameter can be determined through direct measurements conducted on the real system. In many situations, however, the model parameters are conceptual representations of abstract watershed characteristics and must be determined through calibration. In fact, model calibration is the most time-consuming step, not only for hydrological models, but also for Earth system model development, because both parametric estimation and parametric uncertainty analysis require hundreds—if not thousands—of model simulations to understand how perturbations in model parameters affect simulations of dominant physical processes and to find the optimum value of a single parameter.

WRF-Hydro is a numerical model that can simulate the entire hydrological cycle using advanced high-resolution data such as satellite and radar products. Compared with the traditional land surface model (LSM) used by WRF, WRF-Hydro provides a framework for multiscale representation of surface flow, subsurface flow, channel routing, and baseflow, as well as a simple lake/reservoir routing scheme. As a physics-based model, WRF-Hydro includes many complicated physical processes that are nonlinear and must be parameterized. The default parameters given by WRF-Hydro may be valid for one region but not for another region. Hence calibration of related model parameters is often required in order to use the model in a new domain. In particular, for a large spatial domain such as the entire contiguous United States, in order to develop the optimal parameter sets in a reasonable amount of time, the calibration must be conducted on high-

performance computing (HPC) systems in parallel instead of in the traditional sequential mode. To date, no such calibration tool can efficiently calibrate WRF-Hydro on HPC resources. Typically, each physics-based model needs a calibration code that is custom designed to work with that particular numerical model and its set of physics parameterizations, software architecture, and solvers. These custom-designed calibration codes are highly challenging and do not offer flexibility. Therefore, a more flexible and generic calibration tool is needed that can calibrate any code that uses Message Passing Interface/Open Multi Processing (MPI/OpenMP) for parallelization on HPC systems.

One widely used generic and independent calibration tool is the parameter estimation tool (PEST). PEST (Doherty, 2016) conducts calibration automatically based on mathematical methods and thus is applicable for optimizing nonlinear parameters. Compared with manual calibration, automatic calibration is more efficient and effective because it avoids interference from human factors (Madsen, 2000; Getirana, 2010). The uniqueness of PEST is that it operates independent of models: there is no need to develop additional programs or codes for a particular model except preparing the files required by PEST (as described in Sec. 3.2). PEST has four modes of operation. One of the modes is regularization mode, which supports the use of Tikhonov regularization and is found better for serving environmental models because, if implemented properly, it supports model predictions of minimum error variance, is numerically stable, and embraces rather than eschews the heterogeneity of natural systems. Singular value decomposition (SVD) can be used as a regularization device to guarantee numerical stability of the calibration problem. Parallel PEST is able to distribute many runs across many computing nodes using master-worker parallel programing. To our best knowledge, however, no approach is available that allows users to submit jobs using PEST parallelization to a typical supercomputing facility that uses job scheduling and workload management such as Simple Linux Utility for Resource Management (SLURM), Portable Batch System (PBS), and Cobalt. A previous study (Senatore et al., 2015) used PEST to calibrate WRF-Hydro over the Crati River Basin in southern Italy. Because the study area was relatively small, the authors were able to conduct the calibration using PEST in sequential mode (Alfonso Senatore, personal communication, 2018).

This study aims to (1) port parallel PEST to HPC clusters operated by the U.S. Department of Energy (DOE) and adapt it to work with WRF-Hydro, (2) evaluate the performance of HPC-enabled parallel PEST linked to WRF-Hydro by calibrating a flood event, and (3) explore the scale-up capability and computational benefits of HPC-enabled parallel PEST by assigning different computing resource to the entire calibration process.

2 Model description

2.1 Study area

The case presented here is one of the worst floods experienced by greater Chicago area in the past three decades; the storm occurred on April 18, 2013 (Campos and Wang, 2015). According to the National Weather Service (NWS), the heaviest 24-hour accumulated rainfall during this storm reached 201.4, 171.1, and 136.4 mm across Illinois, Iowa, and Missouri, respectively. The Mississippi River crested at 10.8 m (1.7 m above flood stage), and the Illinois River crested in Peoria, Illinois, at 8.95 m; these river cresting broke the previous record of 8.78 m, set in 1943, and was 4.55 m above the historical normal river stage (NWS, 2013). Campos and Wang (2015) conducted three-domain nested WRF simulations to understand the dynamical and microphysical mechanisms of the event. Our study builds on the smallest domain of that study, which covers the majority of Illinois, Iowa, and Missouri at a spatial resolution of 3 km (Fig. 1). The domain size is 750 km from west to east and 660 km from south to north.

2.2 WRF-Hydro configuration

This study employs WRF-Hydro version 5 with a basic configuration. This configuration does not use nudging techniques or spatially distributed soil-related parameters as used in the National Water Model configuration. WRF-Hydro has been tested in several different cases that focused on different hydrometeorological forecasting and simulation problems (e.g., Gochis et al., 2018; Yucel et al., 2015; Senatore et al., 2015; Arnault et al., 2016), and it shows reasonable accuracy in simulated streamflow after being carefully calibrated. For details of the WRF-Hydro modeling system, see Gochis et al. (2018). Currently, two LSMs are available in WRF-Hydro for representing land-surface column physics: Noah (Chen and Dudhia, 2001) and Noah Multi-

parameterization (Noah-MP; Niu et al. 2011). We utilize Noah-MP LSM because compared with Noah LSM it shows obvious improvements in reproducing surface fluxes, skin temperature over dry periods, snow water equivalent, snow depth, and runoff (Niu et al. 2011). The Noah-MP is configured at a grid spacing of 3 km, and the aggregation factor is 15; that is, starting from a 3 km LSM resolution in the domain shown in Fig. 1, hydrological routing is performed at a grid resolution of 200 m, with 3285 south-north × 3735 west-east grid cells. We use a time step of 10 seconds for the routing grid in order to maintain model stability and prevent numerical dispersion of overland flood waves. The time step also meets the Courant condition criteria for diffusive wave routing on a 200 m resolution grid. The WRF-Hydro is configured to be in offline or uncoupled mode—there is no online interaction between the WRF-Hydro hydrological model and the WRF atmospheric model. Overland flow, saturated subsurface flow, gridded channel routing, and a conceptual baseflow are active in this study. The gridded channel network uses an explicit, one-dimensional, variable time-stepping diffusive wave. A direct output-equals-input "pass-through" relationship is adopted to estimate the baseflow. Although the baseflow module is not physically explicit, it is important because the water flow in the channel routing is contributed by both the overland flow and baseflow. If the overland flow is active as it is in this study, it passes water directly to the channel model. In this case the soil drainage is the only water resource flowing into the baseflow buckets. However, if the overland flow is deactivated but channel routing is still active, then WRF-Hydro collects excess surface infiltration water from the land model and passes this water into the baseflow bucket. This bucket then contributes the water from both overland and soil drainage to the channel flow. Therefore, the baseflow must be active if the overland flow is switched off. This study does not consider lakes and reservoirs.

We use the geographic information system (GIS) tool (Sampson and Gochis, 2018) developed by the WRF-Hydro team to delineate the stream channel network, open water (i.e., lake, reservoir, and ocean) grid cells, and groundwater/baseflow basins. Meteorological input for the WRF-Hydro model system includes hourly precipitation; near-surface air temperature, humidity, and wind speed; incoming shortwave and longwave radiation; and surface pressure. In this study, the hourly precipitation is from the National Centers for Environmental Prediction (NCEP) Stage IV analysis at a spatial resolution of 4 km. The Stage IV data is based on combined radar and gauge data (Lin and Mitchell, 2005; Prat and Nelson, 2015), and has been shown to be temporally well correlated

with high-quality measurements from individual gauges (see, e.g., Sapiano and Arkin, 2009; Prat and Nelson, 2015). The other hourly meteorological inputs are from the second phase of the multi-institution North American Land Data Assimilation System project, phase 2 (NLDAS-2) (Xia et al., 2012a,b), at a spatial resolution of 12 km. NLDAS-2 is an offline data assimilation system featuring uncoupled LSMs driven by observation-based atmospheric forcing.

During the 15-day period of this studied case, light to moderate rain occurred on April 8 through 11, 2013, followed by a relatively dry period from April 12 to 15. Then a heavy rain event began on April 16 and peaked on April 18. The heaviest rain band moved east of the study area on April 19. The rainy event ended over the study area on April 20 (see Fig. S1 in Supporting Information). We start the WRF-Hydro simulation on Jan. 1, 2013, and run the model for more than three months to reach equilibrium. This 3-month period is considered as spin-up time and is excluded from model calibration and evaluation. We calibrate the river discharge calculated by the WRF-Hydro model from 00UTC April 9 to 00UTC April 12, 2013, considering it long enough to achieve our objective. We then evaluate the model performance against U.S. Geological Survey (USGS) observed river discharge from 00UTC April 12 to 00UTC April 25, 2013.

3 Calibration

3.1 Platforms

We customized parallel PEST to work on three different workload managers and job schedulers: SLURM at the National Energy Research Scientific Computing Center (NERSC), PBS at the Argonne National Laboratory Computing Resource Center, and Cobalt at the Argonne Leadership Computing Facility. The tests presented here are conducted on Edison at NERSC, which uses the SLURM workload manager and job scheduler. Edison is a Cray XC30 with a peak performance of 2.57 petaflops per second, 133,824 compute cores, 357 terabytes of memory, and 7.56 petabytes of disk storage. It has 5,586 nodes and 24 cores per node.

The interface we have built between parallel PEST and the management software (SLURM here) is, in general, used for (1) setting the number of workers and the nodes for each worker to conduct a model run (WRF-Hydro here); (2) finding the nodes that are available; (3) setting up the working

directory for the workers; (4) identifying the nodes that work for each worker; (5) passing the global files (same for all the working directory) to all the workers (these files include the lookup table files that are not to be calibrated, the namelist files for both LSM and hydrological sector, and restart files that generated by the previous simulations, or spin-up period); and (6) submitting the job for the entire calibration process, including parallel PEST and parallel WRF-hydro. This job can be submitted as a cold-start run or as a restart. The main difference for this interface on different management software is that different management software has its own way to submit jobs and identify available nodes. This difference requires some changes in the script we developed.

3.2 PEST files and settings

PEST requires three file types in both sequential and parallel mode. They are template files to define the parameters to be calibrated, an instruction file to define the format of model-generated output files, and a control file to supply PEST with the size of the problem and the settings for the calibration method. Parallel PEST uses a "master-worker" paradigm that starts model runs simultaneously by different workers (or in different folders). The master of parallel PEST communicates with each of its workers many times during a calibration. To run PEST in parallel mode, one also needs a management file to inform PEST where the working folder is for each worker and what the names and paths are for each model input file that PEST must write (i.e., lookup tables that come from template files) and each model output file that PEST must read (such as frsxt_pts_out.txt). The management file also set the maximum running time for each worker. For those workers that take longer than the maximum running time, PEST will stop the model run by that particular worker and assign that model run to another worker if there is one with nothing else to do.

To the best of our knowledge, however, parallel PEST is not designed to run on HPCs directly. We developed scripts and an interface to enable parallel PEST to run on HPCs using SLURM, PBS, or Cobalt workload managers and job schedulers. The development involved writing scripts to modify the workflow for different workload managers and job schedulers, as well as developing code to connect parallel PEST to WRF-Hydro. These developments enable parallel PEST to run many workers at the same time; each worker runs a parallel code (here WRF-Hydro) that uses

more than one node, which could significantly reduce the wall-clock time of model calibrations. Although this master-worker parallelism may not be as efficient as a fully MPI approach, it is sufficient for model calibration and requires the least effort for the current parallel PEST to run on HPC systems.

This study presents calibration results from PEST using the SVD-based regularization in regularization mode to ensure numerical stability (Tonkin and Doherty, 2005). We focus on calibrating 22 parameters (see Table 1 and detail description in Sec. 3.3) using 96 observation points and 22 items of prior information for the calibrated parameters. In each item of prior information, a value equal to its default value provided by the WRF-Hydro v5.0 (or the log of its default value) is assigned for each adjustable parameter, assuming that default values are the preferred values. All prior information equations are assigned a weight of 1.0. We assigned five different regularization groups to the prior information: Manning's roughness coefficients specified by Strahler stream order in CHANPARM.TBL to one group; the parameters in HYDRO.TBL (Manning's roughness coefficients for overland flow as a function of vegetation types) to another group; and three global parameters for the Noah-MP (xslop1, refdk, and refkdt) in GENPARM.TBL to the remaining three groups. The 96 observation points are given different weights based on the inversed mean of their observed discharge during the studied period (see the detailed description in Sec. 3.3 and Sec. 4.1). For a detailed description of these settings see the PEST User Manual (Doherty, 2016).

3.3 Calibrated experiments

The primary objective of this study is to build a bridge for linking the parallel PEST and WRF-hydro on the basis of HPC clusters and to explore the computational benefits of this bridge. We do not attempt to extensively assess each individual tool or address questions in each individual domain, such as optimizing the objective functions in PEST or calibrating WRF-Hydro for a long time period considering all the relevant parameters to achieve an optimal parameter set. The calibration period thus is limited to only three days, which we believe long enough to achieve our objective and to understand WRF-Hydro's sensitivity to the calibrated parameters. We calibrated WRF-Hydro using four USGS sites (referred to as Station 1, Station 2, Station 3, and Station 4

hereafter), as shown in Fig. 1. (More USGS sites could be included if one manually reallocated the stations that were not properly assigned to the desired location on the channel network by the GIS tool.) We then transfer the calibrated parameters to other subbasins in the study area to assess the transferability of the calibrated parameters. Although many parameters, including spatially distributed parameters and constant parameters in the lookup tables, affect the model performance, we calibrate only the parameters in lookup tables and do not consider the spatial variability of other parameters or their scaling factors. We acknowledge that some studies calibrate a single scaling factor (without considering its spatial variability, however) of overland roughness coefficients (OVROUGHRTFAC) rather than the actual value of each land type in the lookup table (e.g., Kerandi et al., 2018). Although this approach reduces the number of calibrated parameters, however, it has less flexibility because changing one factor will change all the parameters that use the same proportion.

For the calibration exercises we conduct here, the retention depth factor (RETDEPRTFAC) is fixed at 0.001. This value is reasonable because the modeled discharge of our particular configuration (Sec. 2.2) using default parameters is lower than observed discharge. Reducing this factor from 1 to 0.001 keeps less water in water ponds and more water on the surface so it can contribute to river discharge. First, we calibrate 48 parameters based on a 3-day simulation from April 9 to April 11, 2013 (Table S1 in Supporting Information). This calibration uses the estimation mode in the PEST tool and considers equal weight for all four USGS stations. We calibrate Manning's roughness coefficients for both channels and land-use types, the deep drainage (SLOPE), infiltration-scaling parameter (REFKDT), and saturated soil lateral conductivity (REFDK). Manning's roughness coefficients control the hydrograph shape and the timing of the peaks; the SLOPE, REFKDT, and REFDK control the total water volume. Second, based on the knowledge we learn from the 48-parameter calibration (see details in Sec. 4.1), for the same 3-day period, we reduce the number of calibrated parameters from 48 to 22 according to the sensitiveness of the WRF-Hydro model to the adjustable parameters. For example, during the calibration we find that Manning's roughness coefficients for several land types barely change because these land types (e.g., tundra, snow/ice) are not present in the study area. We also learn that even though the calibrated WRF-Hydro parameters can generate discharge results that closely resemble observations, the physical meaning of several parameters are not appropriate because of the wide

range of those parameters that we set in the PEST control file. For example, Manning's roughness coefficient for stream order 1 (0.199) is calibrated smaller than that for stream order 2 (0.218); the overland roughness coefficients for evergreen needleleaf forest (0.043) and mixed forest (0.023) are calibrated smaller than for cropland/woodland (0.046). Neither of these is true in the real world. We therefore adjust the range of many parameters according to the literature (Soong et al., 2012) to maintain their physical meanings (Table 1). We find that by using the same absolute weight for all four stations, the calibration helps three stations (Station 2, 3, and 4) with large water volumes to generate more reasonable results than do the default parameters; however, the results for Station 1, which has a relatively small volume of water, is not always better than the discharge that is modeled by using default parameters. Thus, we assign a weight of 15.0 for Station 1 versus a weight of 1.0 for the other three stations according to the inversed mean of observed discharge over these four stations in April 2013. The ratio of the weights between Station 1 and the other three stations stays similar even if the means are calculated based on different time periods.

3.4 Statistics

This study employs three statistical criteria: Nash–Sutcliffe efficiency (NSE; Nash and Sutcliffe, 1970; Moriasi et al., 2007), root-mean-square error (RMSE), and Pearson correlation coefficient (PCC). RMSE and PCC evaluate model performance in terms of bias and temporal variation. NSE quantitatively describes the accuracy of modeled discharge compared with the mean of the observed data. Equation (1) calculates the NSE with defined variables:

$$NSE = 1 - \frac{\sum_{t=0}^{n}\left(Y_t^{obs} - Y_t^{sim}\right)^2}{\sum_{t=0}^{n}\left(Y_t^{obs} - Y_{mean}^{sim}\right)^2}, \quad (1)$$

where Y_t^{obs} is the *t*th observed value from USGS sites for river discharge , $Y_t^{sim}$is the *t*th simulated value from the WRF-Hydro output, Y_{mean}^{obs} is the temporal average of USGS observed discharge, and *n* is the total number of observation time points. An efficiency of 1 (NSE = 1) corresponds to a perfect match between modeled discharge and observed data. An efficiency of 0 (NSE = 0) indicates that the model predictions are as accurate as the mean of the observed data. An efficiency below zero (NSE < 0) occurs when the model is worse than the observed mean. Essentially, the closer the NSE is to 1, the more accurate the model is.

4 Results

4.1 WRF-Hydro calibration and validation

Based on the knowledge we gained from the 48-parameter 3-day calibration, we adjust the range of critical parameters in the PEST control file to main their physical meanings. For example, we set Manning's roughness coefficient larger for stream order 1 than for stream order 2. We also adjust the parameter range of the overland roughness coefficient for multiple land covers, such as forests. We exclude the parameters that WRF-Hydro is not sensitive to for this study, in order to constrain the problem size considering the availability of computational resources. However, if the studied area is much larger with more land types than the study area here, then there would be more parameters to calibrate. Also, hundreds of constant parameters in the Noah-MP model could affect the WRF-Hydro results (Cuntz et al. 2016) and can be calibrated. Both these situations would increase the burden of WRF-Hydro calibration. We perform the same 3-day calibration from April 9 to April 11, 2013. Figure 2 shows the results of the 3-day modeled discharge (in cubic meters) using default and calibrated parameters after five iterations, as well as observed discharge. The four stations are calibrated by considering different weights. Compared with the results calibrated by using equal weights for all the stations, by giving a higher weight to Station 1 the model bias over Station 1 is significantly reduced, with a higher NSE (0.87 with higher weight versus 0.14 with equal weight) and lower RMSE (48.1 versus 123.6). Over Stations 2, 3, and 4, which sit on rivers with relatively large water volumes, the modeled discharge using the default parameter underestimates the streamflow by more than 65%. PEST detects this underestimation and immediately adjusts the parameters and increases the modeled discharge during the first iteration. After the third iteration, the difference in calibrated results between different iterations is relatively small. We allow the PEST to conduct five iterations and use the parameters obtained from the fifth iteration as our optimum parameters. As shown in Table 2, when the optimum parameters are used, the modeled discharges are much closer to the observations compared with the modeled results when the default parameters were used. The NSEs for the four stations increased from 0.73 (Station 1), -54.4 (Station 2), 157.3 (Station 3) and -1316.9 (Station 4) to 0.87, 0.64, 0.05, and -58.78, respectively, being closer to 1. The RMSEs decreased from 69.3, 3925.2, 3981.3, and 4391.3 m^3/sec to 48.1, 318.2, 308.7, and 934.6 m^3/sec, respectively. Giving a lower weight for the three large river stations does not change the calibration results much.

During the validation period, compared with the modeled discharge using default parameters, as shown in Table 2, the NSEs for all four stations are increased to be closer to 1; RMSEs are decreased by 50% or more; and the correlation coefficients between the observed and modeled discharge are increased from 0.8, 0.76, 0.21, and 0.72 to 0.98, 0.82, 0.80, and 0.75. Compared with the results of calibration using the estimation mode (no regularization) in PEST, the SVD-based regularization generates slightly better hydrograph shape with 24-hour later discharge peaks that are closer to the observations. However, a problem remains with the hydrograph shapes of the modeled discharge, especially with the modeled peak of discharge. For Station 1, the WRF-Hydro almost captures the timing of the peak of discharge, although it still underestimates the water volume by ~25%. The reason is that this study uses a direct pass-through baseflow module, which does not account for slow discharge and long-term storage of the baseflow. Therefore, the largest contribution to river discharge is from precipitation, and groundwater does not contribute much discharge to the channels in a long-term view, as is also true for the other three large river stations. Different from Station 1, for the other three large river stations, the WRF-Hydro modeled discharge increases soon after the peak of precipitation and reaches a peak on April 21, 2013, which is much earlier than the observed peak of river discharge (near April 24). The reason is that the water contributions for these stations are from a larger river basin (Mississippi River) than we included in our current study area. Thus, when a heavy precipitation event occurs over the entire river basin, there will be a significant lag time (especially at the lower part of the basin) between the peak of precipitation amount and the peak of river discharge. For example, the precipitation over the upper part of Mississippi River Basin (MRB) has a peak amount on April 18–19, but the river discharge did not reach its peak until April 24. Because our studied area covers only half of the MRB, the modeled river discharge has a shorter delay period after the peak of precipitation than does the observed river discharge. Enlarging the study area to include the entire MRB may improve this situation. Alternatively, calibrating and validating local rivers that are included in the current study area may also reduce the bias in hydrograph shape compared to calibrating and validating large rivers. On the other hand, the WRF-Hydro simulated river discharge decreases soon after it reaches the peak and much earlier than the observed discharge. The reason is again that the direct pass-through baseflow employed by this study does not account for slow discharge and long-term storage of the baseflow. As a result, the contribution from the baseflow to the river

discharge in model simulations does not stay as long as in real situations. In the observations, the river discharge decreases from the peak at a speed of ~500 m^3/sec per day, while the modeled river discharge decreases from the peak at a speed of ~1667 m^3/sec per day. Using exponential storage-discharge function for the baseflow may improve this situation.

4.2 Computational benefits of parallel PEST on HPCs

The ability to scale up the calibration of WRF-Hydro by using parallel PEST on HPC systems is determined by two factors: the scale-up capability of parallel PEST and the scale-up capability of WRF-Hydro. In calibrating WRF-Hydro, PEST first makes as many model runs as there are adjustable parameters to calculate Jacobian matrix (Doherty, 2016). The Jacobian matrix has a column for each calibrated parameter and a row for each observation and each item of prior information that set in the PEST control file. These model runs are independent between workers and can be easily parallelized. Each worker runs the model with temporarily incremented parameters that are defined in the template and control files. Then, PEST needs to make additional model runs to test parameter updates. Different from the Jacobian runs, these additional runs are performed by using different Marquardt lambdas, and the search for a Marquardt lambda that achieves the best set of parameters is a serial iterative process. The lambda to use for the next run depends on the outcome of the model run conducted using the previously chosen lambda. Although serial testing of Marquardt lambdas may quickly find the optimal Marquardt lambda in the first or second series of model runs, it is an inefficient use of computing resources because other processors are idle while only one process is searching the lambdas. This is especially true when the model domain is large and requires extensive computing resources. This study employs "partial parallelization" for the lambda-testing procedure (Doherty, 2016), so multiple workers can be used to calculate parameter upgrades based on a series of lambda values that are related to each other by a factor of RLAMFAC set in the PEST control file. We also set the value of PARLAM to -9999 in the management file so only one cycle of parallel WRF-hydro runs is devoted to testing Marquardt lambdas. For additional details on these parameters and their settings see the PEST User Manual (Doherty, 2016).

In this study we test the computational performance of HPC-enabled parallel PEST using different number of workers (6, 12, and 23) for the 22-parameter calibration. As shown in Table 3, we conducted five experiments: Test 1 uses 23 workers, Test 2 uses 12 workers, and Test 3 uses 6 workers. All three tests use two nodes for each worker to run WRF-Hydro in parallel. The maximum number of lambda-testing runs undertaken per iteration is set to 15, 10, and 5 for Test 1, 2, and 3, respectively, to make sure that only one cycle of WRF-hydro runs is devoted (using 15, 10 and 5 workers from Tests 1, 2, and 3, respectively) to testing Marquardt lambdas. Note that the maximum number of lambda-testing runs should be set equal to or less than the workers available. Otherwise, another cycle of WRF-hydro runs needs to be conducted. In fact, generating more Marquardt lambdas does not always guarantee that the best Marquardt lambdas are generated. In contrast, it may make the model convergence slower (here, PEST) or even model failure.

In order to test the trade-offs between the computing nodes used for running parallel WRF-Hydro and the workers used for running parallel PEST, Tests 4 and 5 use different number of nodes for each worker to run WRF-Hydro in parallel. Explicitly, Test 4 uses four nodes per worker, and Test 5 uses six nodes per worker. Both tests use six workers for running the parallel PEST. The maximum number of lambda-testing runs undertaken per iteration is set to five for both Tests 4 and 5. Note that the time costs in Table 3 are limited to only one iteration. Conducting more iterations will increase the cost of wall-clock time and computing, but will not change the conclusion for the scale-up capability and computational benefits for HPC-enabled parallel PEST linked to WRF-hydro.

PEST needs to run the WRF-Hydro model at least as many times as the number of calibrated parameters (22 here). In fact, PEST runs the model 23 times in the first round (or the first iteration) with initial parameter values and for the first Jacobian matrix. From the second iteration, it runs the model 22 times to calculate Jacobian matrix. Therefore, if there are fewer than 23 workers, the time cost for the first round of Jacobian matrix calculation will increase accordingly. For example, as shown in Fig. 4a, when we assign 12 (and 6) workers to parallel PEST, the time cost for calculating the Jacobian matrix is increased by a factor of 2 (and 4) compared with the time cost of using 23 workers. The time cost for the parameter upgrade stays similar for the three

experiments because only one cycle of WRF-hydro simulation is conducted to test the Marquardt lambdas. As a result, the total time cost for Test 2 is ~1.5 times more than that for Test 1, and the total time cost for Test 3 is ~1.5 times more than that for Test 2 (Fig. 4b). By extrapolating the speedup curve shown in Fig. 4a and Fig. 4b, we expect the total time cost to be ~1516 minutes when using only one worker (or sequential mode), which is about 15 times slower compared with running the PEST in parallel mode using 23 workers. For this particular study with 22 adjustable parameters, we expect the time cost most likely to stay the same even if one increases the number of workers to more than 23, because PEST runs WRF-Hydro only 23 or 22 times for each iteration. Assigning more workers for this particular study would most likely render some workers idle and is not an efficient use of computing resources. PEST may run WRF-Hydro more than 22 times (e.g., 44 times) if higher-order finite differences are employed. In this case, assigning more workers (e.g. 45 workers) may further speed up the calibration process. On the other hand, for the same case study and using the same number of nodes for running parallel WRF-Hydro, we can estimate the computing speedup by assuming an increase in the number of calibrated parameters to 50. This would be the case, for example, to evaluate model sensitiveness to the physics in Noah-MP or the spatial variabilities of certain parameters. We then expect to use 51 workers to achieve the best computing performance for parallel PEST. This would then be 28–30 times faster than running PEST using one worker (or in sequential mode). Similarly, if 100 parameters were used for the calibration for the same case study, a factor of up to 60 speedup in the calibration process would be achieved by running HPC-enabled parallel PEST.

In addition, by increasing the number of nodes for each worker to conduct WRF-Hydro (Tests 3, 4, and 5), the time cost for the entire calibration process is significantly reduced (Figs. 4c and 4d). Specifically, the WRF-hydro scales up well when using four and six nodes compared with using two nodes per worker for running the WRF-Hydro. Both the time spent on calculating the Jacobian matrix and the time spent on testing the parameter upgrades are decreased by 49% and 67%, respectively, when using four and six nodes. Therefore, the total time spent is also decreased when using more nodes for each worker (see Table 3). Increasing the number of nodes to eight for each worker will most likely further decrease the time cost by 70–75% compared with using only two nodes per worker. Moreover, if one has a larger study area such as the entire contiguous United States, we expect the WRF-Hydro to have an even better scale-up capability (e.g., on dozens of

nodes) than this study. Overall, based on the experiments we conduct here, using 23 workers for parallel PEST and six nodes for each worker to run parallel WRF-Hydro would cost the least wall-clock time—about 32 min for one iteration for this particular study.

4.3 Evaluation of spatial transferability of the calibrated parameters

To assess the transferability of the calibrated parameters, we apply the optimum parameters obtained from the calibration for the four stations (black circles) in Fig. 1 to another set of four stations (crosses in Fig. 1) in the study area. All four sites are located on relatively small rivers, so the lag time between precipitation peak and the discharge peak are much shorter than that for the stations on the lower part of MRB (e.g., Stations 2, 3, and 4). The assessment compares the observed discharge with the closest grid cells from the discharge output of WRF-Hydro. Figure 5 shows the observed and modeled discharge using default and the optimum parameters. Overall, WRF-Hydro's default parameters underestimate the discharge and misrepresent the timing of discharge peaks compared with observations over the four assessed stations (Stations 5, 6, 7, and 8). By using the calibrated parameters from other sites over the area, the model results increase the discharge and shift the hydrograph shape so they are much closer to the observations than model results using default parameters. The absolute error of simulated discharge decreases by 13.1%, 38.3%, and 71.6%, respectively, over Stations 6 through 8 (Station 5 shows a 6% increase of absolute error), compared with the default simulated discharge. We also find that using the SVD-based regularization for the PEST calibration captures the timing of discharge peak better than using the estimation mode, which is one-day earlier than the observations reaching the discharge peak.

5 Summary and discussion

WRF-Hydro is a new, and perhaps the first practical, computer code that can run on HPC systems and can model the entire hydrological cycle using physics-based submodels and high-resolution input datasets (e.g., radar). The hydrological community has desired this capability for decades, although it requires intensive computing resources. Thus, the calibration of this model would ideally be conducted on HPCs in parallel as well, especially when the model covers a large domain rather than the basin scale. This study ports an independent model calibration tool, parallel PEST,

to HPC clusters and links it to WRF-Hydro to help WRF-Hydro users calibrate the model within a much shorter wall-clock time period. The bridge we build here (between parallel PEST and WRF-Hydro on the basis of HPC systems) can be applied to any other hydrological models and Earth system models that use parameterizations to represent model physics. We present the operational feasibility of the HPC-enabled parallel PEST by evaluating the performance of calibrated WRF-Hydro against observation in hydrograph features such as volume and timing of flood events. We examine the scale-up capability and computational benefits of the tool by assigning different computing resource for PEST and for WRF-Hydro. While this study presents the optimum parameters identified from the calibration of the particular flood event, the parameters can be significantly different if one uses different physics, such as exponential storage-discharge function for a groundwater model or reach-based channel routing. Our preliminary testing shows that using exponential storage-discharge function with the default parameters provided by WRF-Hydro, the modeled discharge was larger than that of observations. Thus, the calibration will need to adjust the parameters to reduce the discharge. Our study finds that for calibrating 22 parameters, using the same computing resource for running WRF-hydro, the HPC-enabled PEST calibration tool can speed up WRF-Hydro calibration by a factor of 15, compared with running PEST in sequential mode. The speedup factor can be larger when the number of parameters needing calibration is higher (e.g., 50 or 100).

The following are several key points that we would like to mention to inform future studies:

1. In this study, we consider using the prior or regularization information only for the parameters that we calibrate. As is the case with solving inverse problems, prior information is added to improve the smoothness of the solutions. In order to build a more comprehensive calibration, an important aspect that can be considered is to enrich the prior with the available historical data. For example, in this particular case, one can use the historical observation data (e.g., April and May from the past few years) to enrich the prior information for the parameters. Hence, the regularization objective function in PEST will constitute not only the discrepancies between parameters and their "current estimates" but also the discrepancies between WRF-Hydro simulations and preferred values (which is the observed time series of historical discharge). Additionally, one can use the pilot points technique described by Doherty (2005) in conjunction with parameter estimation to add

more flexibility to the calibration process. This will be potentially beneficial in improving the predictions.

2. To focus on our main goal, we calibrate only the parameters in lookup tables. However, we acknowledge that using a single value to represent a physics for a large domain could be problematic, especially we expect the HPC-enabled parallel PEST to execute with WRF-Hydro for large domains. This situation often needs parameter regionalization. For example, WRF-Hydro v5 has many spatially distributed parameters available, such as the overland flow roughness scaling factor (OVROUGHRTFAC), the factor of maximum retention depth (RETDEPRTFAC), and the soil-related parameters (when compiled with SPATIAL_SOIL=1). Calibrating these spatial parameters based on grid scale (e.g., catchments) rather than a single value will give the model more flexibility and thus better fit the observations (Hundecha and Bardossy, 2004; Wagener and Wheater, 2006). In practice, for example, one can include regional OVROUGHRTFACs (e.g., their lower/upper bounds, and default values) in the PEST control file based on catchments. However, the selection of the locations and sizes of catchment may introduce significant uncertainties to the calibration results, which require systematic and comprehensive investigation and understanding of the study area.
3. This study is limited to calibrating the observed streamflow only based on the format of one of WRF-Hydro model outputs for individual station or point (frxst_pts_out.txt). It is feasible, however, to calibrate other variables as long as the observation data is available. For example, one can either find the closest point from the gridded dataset to the observation location and then compare that model grid to observations; or one can change the WRF-Hydro input/output code to output other variables in the frxst_pts_out.txt file, so they can still use the same interface we developed here to calibrate other variables instead in addition to the discharge.
4. The optimal parameter set obtained from this study is from the 5th iteration of parallel PEST by testing five Marquardt lambdas. Testing different number of lambdas or calibrating different number of parameters may generate a different set of optimal parameters. These parameter sets can all make physical sense and be equally good for reproducing observed discharges. This problem is named equifinality (Beven and Freer, 2001; Savenije, 2001), which is an important source of model uncertainty. To reduce the

model uncertainty through reducing the equifinality, hydrologists carry out additional modelling objective for model evaluation to find more useful parameter sets (Mo and Beven, 2004; Gallart et al., 2007). Alternatively, inspired by No. 3 discussed above, one can calibrate the WRF-hydro model based on more than one variables, such as discharge and soil moisture (or heat flux or water table depth) to reduce the number of optimal parameter sets, and thus reduce the model uncertainty of predictions for these variables.

5. While this study ported the parallel PEST to HPC system and linked it to WRF-Hydro, we note that BEOPEST is available in the PEST family. BEOPEST has the same functionality as parallel PEST but uses a different approach for communication between master and workers. Working with HPC-enabled BEOPEST may save total time cost since BEOPEST uses the Transmission Control Protocol (TCP) and the Internet Protocol (IP) instead of message files (reading input and writing output between master and works) for communication. We expect it to be relatively straightforward to use BEOPEST to calibrate WRF-hydro on HPCs since the interface remains similar, except one needs to copy the template and instruction files in addition to the global files (see Section 3.1) into each working folder.

Data and Code availability. The observed river discharge is downloaded from the USGS Surface-Water Data website, available at https://waterdata.usgs.gov/nwis/sw. The Stage IV precipitation data were downloaded from https://data.eol.ucar.edu/dataset/21.093. PEST was downloaded from http://www.pesthomepage.org/Downloads.php. We use the Unix PEST version 13.6. The scripts and files that are developed in this study and required by PEST for calibrating WRF-Hydro are available at http://doi.org/10.5281/zenodo.2588506.

Author contributions. JW proposed the project and developed the study case in WRF and WRF-Hydro. CW developed the scripts/code to port the parallel PEST to DOE supercomputers and adapt it to work with WRF-Hydro. VR provided important input for the regularization calibration method. AO operated the ArcGIS tool to delineate the high-resolution grid cells to include stream channel network, open water, and groundwater/baseflow basins. RK provide high-level guidance and insight for the entire project. All authors commented on this manuscript.

Competing interests. The authors declare that they have no conflict of interest

Acknowledgments. This work is supported under a Laboratory Directed Research and Development (LDRD) Program at Argonne National Laboratory, through U.S. Department of Energy (DOE) contract DE-AC02-06CH11357. Computational resources are provided by the DOE-supported National Energy Research Scientific Computing Center, Argonne National Laboratory Computing Resource Center, and Argonne Leadership Computing Facility. Our special thanks to the PEST developers and entire WRF-Hydro team, especially Kevin Sampson for his guidance on the ArcGIS tool. We gratefully thank the two reviewers for their valuable comments and suggestions, which tremendously improved this manuscript.

References

Arnault, J., Wagner, S., Rummler, T., Fersch, B., Bliefernicht, J., Andresen, S., and Kunstmann, H.: Role of runoff–infiltration partitioning and resolved overland flow on land–atmosphere feedbacks: A case study with the WRF-Hydro coupled modeling system for West Africa, J. Hydrometeorol., 17, 1489–1516, 2016.

Campos, E., and Wang, J.: Numerical simulation and analysis of the April 2013 Chicago Floods, J. Hydrol., 531, 454–474, 2015.

Chen, F. and Dudhia, J.: Coupling an advanced land surface-hydrology model with the Penn State-NCAR MM5 modeling system, Part I: Model implementation and sensitivity, Mon. Weather Rev., 129, 569–585, 2001.

Cuntz, M., Mai, J., Samaniego, L., Clark, M., Wulfmeyer, V., Branch, O., Attinger, S., and Thober, S.: The impact of standard and hard-coded parameters on the hydrologic fluxes in the Noah-MP land surface model, J. Geophys. Res. Atmos., 121, 10,676–10,700, doi:10.1002/2016JD025097, 2016.

Doherty, J.: PEST: Model Independent Parameter Estimation, User Manual, 6th ed., Watermark Numerical Computing, Brisbane, Queensland, Australia, 2016.

Doherty, J.: Ground water model calibration using pilot points and regularization, Groundwater, 41(2), 170–177, 2005.

Gallart, F., Latron, J., Llorens, P., and Beven, K. J.: Using internal catchment information to reduce the uncertainty of discharge and baseflow predictions. Adv. Water Resour. 30(4), 808–823, 2007.

Getirana, A. C. V.: Integrating spatial altimetry data into the automatic calibration of hydrological models, J. Hydrol., 387 (3-4), 244–255, doi: 10.1016/j.jhydrol.2010.04.013, 2010.

Gochis, D. J., Barlage, M., Dugger, A., FitzGerald, K., Karsten, L., McAllister, M., McCreight, J., Mills, J., RafieeiNasab, A., Read, L., Sampson, K., Yates, D., and Yu, W.: The WRF-Hydro modeling system technical description, (Version 5.0). NCAR Technical Note. 107 pages. Available online at: https://ral.ucar.edu/sites/default/files/public/WRFHydroV5TechnicalDescription.pdf, 2018.

Hundecha, Y., and Bárdossy, A.: Modeling of the effect of land use changes on the runoff generation of a river basin through parameter regionalization of a watershed model, J. Hydrol., 292, 281–295, 2004.

Kerandi, N., Arnault, J., Laux, P., Wagner, S., Kitheka, J., and Kunstmann, H.: Joint atmospheric-terrestrial water balances for East Africa: A WRF-Hydro case study for the upper Tana River basin, Theor. Appl. Climatol., 131, 1337–1355, doi: 10.1007/s00704-017-2050-8, 2018.

Lin, Y., and Mitchell, K. E.: The NCEP stage II/IV hourly precipitation analyses: Development and applications, Preprints, 19th Conf. on Hydrology, San Diego, CA, Amer. Meteor. Soc., 1.2., 2005.

Madsen, H.: Automatic calibration of a conceptual rainfall–runoff model using multiple objectives, J. Hydrol., 235, 276–288, 2000.

Mo, X., and Beven, K.: Multi-objective parameter conditioning of a three-source wheat canopy model. Agricultural & Forest Meteorol. 122(1–2), 39–63, 2004.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., and Veith, T. L.: Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, Transactions of the ASABE, 50 (3), 885–900, 2007.

Nash, J. E., and Sutcliffe, J. V.: River flow forecasting through conceptual models, part I – A discussion of principles, J. Hydrol., 10(3), 282–290, doi: 10.1016/0022-1694(70)90255-6, 1970.

Niu, G.-Y., Yang, Z.-L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., Kumar, A., Manning, K., Niyogi, D., Rosero, E., Tewari, M., and Xia, Y.: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements, J. Geophys. Res., 116, D12109, doi: 10.1029/2010JD015139, 2011.

NWS (National Weather Service): Record river flooding of April 2013, https://www.weather.gov/ilx/apr2013flooding, 2013.

Prat, O. P., and Nelson, B. R.: Evaluation of precipitation estimates over CONUS derived from satellite, radar, and rain gauge data sets at daily to annual scales (2002-2012), Hydrol. Earth Syst. Sci., 19, 2037–2056, doi: 10.5194/hess-19-2037-2015, 2015.

Sampson, K., and Gochis, D.: WRF Hydro GIS Pre-processing tools, Version 5.0 Documentation, 2018.

Sapiano, M. R. P., and Arkin, P.A.: An intercomparison and validation of high-resolution satellite precipitation estimates with 3-hourly gauge data, J. Hydrometeor., 10, 149–166, doi: 10.1175/2008JHM1052.1, 2009.

Senatore, A., Mendicino, G., Gochis, D. J., Yu, W., Yates, D. N., and Kunstmann, H.: Fully coupled atmosphere-hydrology simulations for the central Mediterranean: Impact of enhanced

hydrological parameterization for short and long time scales, J. Adv. Model. Earth Syst., 7(4), 1693–1715, doi: 10.1002/2015MS000510, 2015.

Soong, D. T., Prater, C. D., Halfar, T. M., and Wobig, L. A.: Manning's roughness coefficients for Illinois streams, U.S. Geological Survey Data Series 668, 2012.

Tonkin, M. J., and Doherty, J.: A hybrid regularized inversion methodology for highly parameterized environmental models, Water Resource Research, 41, W10412, doi:10.1029/2005WR003995, 2005.

Wagenera, T., and Wheater, H. S.: Parameter estimation and regionalization for continuous rainfall-runoff models including uncertainty, J. Hydrol., 320, 132–154, 2006.

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., Luo, L., Alonge, C., Wei, H., Meng, J., Livneh, B., Lettenmaier, D., Koren, V., Duan, Q., Mo, K., Fan, Y., and Mocko, D.: Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2), 1: Intercomparison and application of model products, J. Geophys. Res., 117, D03109, doi: 10.1029/2011JD016048, 2012a.

Xia, Y., Mitchell, K., Ek, M., Cosgrove, B., Sheffield, J., Luo, L., Alonge, C., Wei, H., Meng, J., Livneh, B., Duan, Q., and Lohmann, D.: Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2). 2. Validation of model-simulated streamflow, J. Geophys. Res., 117, D03110, doi: 10.1029/2011JD016051, 2012b.

Yucel, I., Onen, A. Yilmaz, K. K., and Gochis, D. J.: Calibration and evaluation of a flood forecasting system: Utility of numerical weather prediction model, data assimilation and satellite-based rainfall, J. Hydrol., 523, 49–66, 2015.