I have finished my review of the paper "CREST-VEC: A framework towards more accurate and realistic flood simulation across scales", by Li et al., submitted to Geoscientific Model Development. This paper outlines the integration of two existing models – the gridded hydrological model CREST and the vector-based routing tool mizuRoute. The authors examine the (significant) improvement in computational cost associated with using a vector-based routing tool and explore the impact of including or not including lakes on simulations of 5350 streamflow hydrographs across the U.S., with impacts evaluated using standard model fit metrics (e.g., NSE/bias) and using flood detection measures (POD, FAR, and CSI). Lastly, the author examine the correlation of absolute model performance against environmental indices from the CAMELS database.

In general, I found this paper to be a bit disorganized and missing a number of important details. It uses less-than-ideal approaches for comparing populations of model runs and does not seem (to me) to represent a significant contribution either as a model development or in presenting new insights into models more generally. I document below a number of major issues with the paper which should be addressed prior to publication.

Response:

Thank you for your constructive comments and suggestions. Our responses to your comments are listed below.

- The title of this paper implies it may be a discussion of a fundamentally new model or modelling framework. However, it is the merger of an existing routing model which has already been coupled to multiple existing hydrological models (including gridded models) with yet another hydrological modelling code. The details of this integration are generally missing, so it is difficult to evaluate the challenge in performing this integration, whether it is file-based coupling or a single compiled code, whether anything other than the spatial aggregation of runoff is causing the computational improvements in the shift from CREST to CREST-VEC. It does not appear that this model development effort in and of itself is a contribution. If it is, then the authors need to demonstrate why.

Response:

Thank you for your comments. This framework is in fact not only a merger of existing codes, but we did improvements to the codes and added critical subsurface routing. Prior to this point, the evaluation of both subsurface routing and lake operation have not been comprehensively evaluated over the US. The scientific scope of this study is to draw attention to falsely alarmed floods if these two components in hydrologic framework are missing. We expanded the model integration details and described more about the newly added components.

L.208-214: "The framework has loosely coupled by two models written in different programming languages. A bash file calls three executables after model compilation subsequently (CREST-EASYMORE-mizuRoute). The input files for this model chain include forcing data (gridded precipitation, potential evaporation, and temperature), topography data (gridded digital elevation model, flow direction, flow accumulation, river network topology, and hydrologic response unit), and configuration files. The topography data can be accessed from the HydroSHEDS website which consists of grid-based and vector-based topography data."

More details shown in subsurface routing is described.

L. 191-199: “In this study, we enable an option to turn on or off subsurface routing as defined in the model configuration file. Similar to surface runoff routing, the subsurface flow is routed using the IRF scheme but with much slower velocity and reduced magnitude. We use a two-parameter Gamma distribution function to materialize the IRF method as shown in eq. 1.

$$y(t) = \frac{1}{\Gamma(a)\theta^{a}} t^{a-1} e^{-\frac{t}{\theta}}$$

Where t is the time variable, a is a shape parameter, and θ is a time-scale parameter. Both a and θ determine the flood peaking time and flashiness. After calculating instantaneous rates based on gamma function, we use a convolution to compute flow rates Q at time t. R(t-s) is the (sub)surface runoff at time (t-s), and s is an increment of time from 0 to tmax (also denoted as the time window). The default values of a and θ for hillslope surface routing are set to 2.5 and 8000. For subsurface flow routing, the a and θ are 10 and 86400, respectively.

$$Q(t) = \int_{0}^{tmax} y(t) \times R(t-s) ds$$”

- The authors use results from thousands of model runs in order to assess the benefit of including lakes in vector-based routing. The effort to do this is not trivial, and I applaud the authors for their ambition in the size of this computational experiment. However, there are some issues with both the magnitude of the contribution and the means of assessing the results. About the question of whether simulating lakes is useful, the inclusion of lakes has already been demonstrated to improve vector-based routing results (Han et al., 2020), though perhaps not in as rigorous of a manner as could potentially be done with this massive model set. However, there are several fundamental issues with the analysis herein. Firstly, the model in general (before and after lakes are included) performs very poorly, with median NSEs on the order of 0.3, and 45% of simulations having NSEs below zero. The authors make no attempt to reduce their analysis to focus on models with that satisfy some adequate base performance. They also spend most of their time discussing improvements to the median, which means that much of the performance can easily be due to changing very poorly performing models to slightly less terrible models. They miss the opportunity to statistically compare the probability distributions of metrics to see whether the distributions are functionally different. At the very least, I would ask that they compare median change in NSE (a more adequate metric for evaluating improvements) rather than the change in median NSE.

Response:

Thanks for your comments.

First, regarding the model performance, we compared our results with other studies since it is hard to perceive the model performance without relative comparison. One recent study by Tijerina et al., (2021) did continental simulation using two state-of-the-science hydrologic models - ParFlow and WRF-Hydro, and they used two indicators to describe (CC and Bias) as shown in Figure 1. Parflow has 56% gauges identified as “good” (low bias and good shape; green color), and the National Water Model (NWM) has 65%. For our results, 63.3% gauges have good agreements, in contrast to 61.2% of them without lake simulation. It is understandable since the NWM has been regularly calibrating their results as a part of their missions. We have lower than 10% gauges with low bias and poor shape (purple color). Overall, our model results have comparable results with modern continental-scale hydrologic models. As one objective of this study, we want to examine the potential improvement from with-lake configuration on streamflow simulation over a wide range of hydrometrical and geographical settings in the CONUS, rather than provide some optimal model setup and parameterization at the CONUS scale, which we believe is way beyond our scope and several steps forward from the current CREST-VEC or any existing CONUS models (as shown here). As far as what qualifies 'an adequate base simulation', there may be some room for debate but should be some bottom-line principles: first, one should be clearly aware of the sources of uncertainties including forcing, model structure, parameterization, streamflow observation as the reference, etc. Optimization, though effective in improving the model performance, compensates uncertainties from the other sources simply via adjusting model parameters. This has been acceptable for operational purposes but is not appropriate for this study where a modification of model structure is introduced. Instead, we use an a-priori parameter set that was developed based on remote sensing datasets and also evaluated at the CONUS scale (Vergara et al., 2016). The physical base of these a-priori parameters set a solid foundation for examining the new with-lake configuration, thus should not be compromised via parameter tunning,

Tijerina, D., Condon, L., FitzGerald, K., Dugger, A., O'Neill, M. M., Sampson, K., et al. (2021). Continental hydrologic intercomparison project, phase 1: A large-scale hydrologic model comparison over the continental United States. Water Resources Research, 57, e2020WR028931. https://doi.org/10.1029/2020WR028931

Vergara, H., Kirstetter, P., Gourley, J.J., Flamig, Z.L., Hong, Y., Arthur, A., and Kolar, R: Estimating a-priori kinematic wave model parameters based on regionalization for flash flood forecasting in the Conterminous United States, 541, 421-433. https://doi.org/10.1016/j.jhydrol.2016.06.011, 2016.

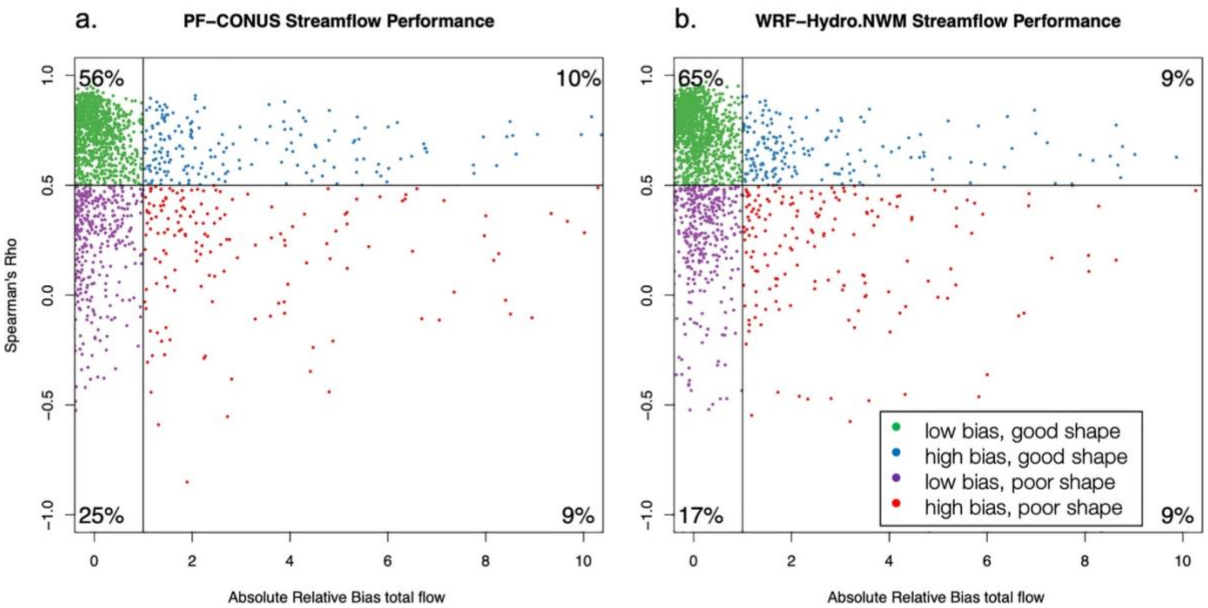


Figure 1. Simulation results by Tijerina et al. (2021). Copyright (2021) American Geophysical Union.

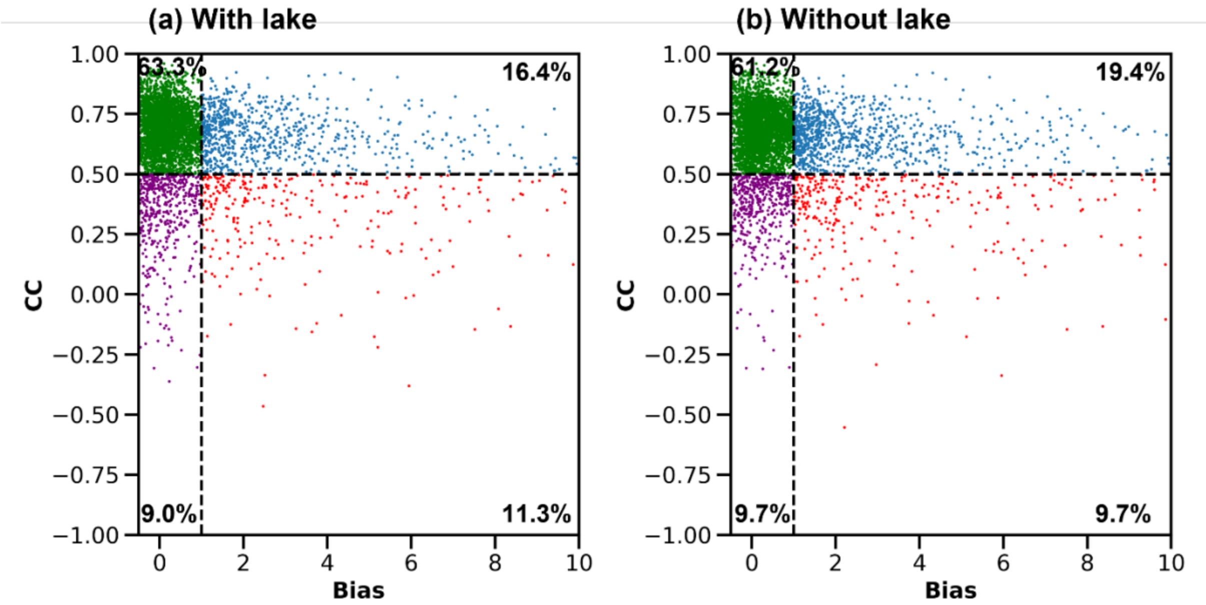


Figure 2. Similar to Figure 1, but produced with CREST-VEC.

Second, we discussed also the poor-performing regions, especially over the Great Plains where the model bias is large. We mentioned the challenges to hydrologic modelling community since it is ubiquitous to continental simulations.

L. 330-342: “CREST-VEC with lake module in regions like the West Coast and Upper Mississippi River Basin have relatively good performance (NSE>0.4), yet over the Great Plains and East Coast, the model bias is high (BIAS>1), yielding low NSE scores. Similar issues are found in the literature with other models (Clark et al., 2008; Konben et al., 2020; Lin et al., 2019; Mizukami et al., 2017; Newman et al., 2015; Salas et al., 2017; Knoben et al., 2020; Yang et al., 2021; Tijerina et al., 2021). Taking the Great Plains as an example (highlighted box in Fig. 6c), the model physics of CREST-VEC does not correctly represent the real hydrologic processes by two means. First, the surface runoff (before routing) simulated by CREST-VEC is biased. We compare the annual surface runoff by CREST-VEC to the public community dataset GRFR (Global Reach-level Flood Reanalysis) in Fig. S1. The runoff in GRFR is simulated by the VIC model and undergoes stringent bias correction against observations via the discrete quantile mapping technique (Yang et al., 2021; Lin et al., 2019). There is a 116.3% higher surface runoff by the CREST-VEC than theirs, partly explaining the high BIAS and low NSE scores in such region. We suspect the singular bulk soil layer represented in the CREST model yields such systematic differences. Second, the missing representation of playas, small and rain-fed lakes that are prominent in the Great Plains, leads to falsely produced runoff”

Besides, we attributed model performance to basin attributes in the Southeast and found the CREST-VEC model performs poorly for intermittent rivers and deep soil depth, both of which are related to hydrologic model physics.

“Figure 8b shows an example of poor-performing gauges in the Southeast. Analogous to the Great Plains, the soil depths in the Southeast are considerably high (1.5 meters), leaving CREST model simulations problematic. Evapotranspiration (ET) in the Southeast is also one of the highest among the US climate divisions due to abundant precipitation, permeable soils, dense vegetation, and substantial soil radiation. Because the CREST model does not account for transpiration from vegetation nor solve the energy balance explicitly, the simulated evaporation rates may be lower than actual evaporation rates, resulting in higher effective rainfall and thus positive bias of streamflow.

Therefore, the missing hydrologic processes such as transpiration and infiltration-excess process in the Southeast are likely the causations of lower NSE scores."

Third, we not only evaluated the relative change in NSE, but also the distribution as the boxplots shown in Figs. 5 and 9. There are distribution shifts in the lower tails. As suggested, we modified the change in median metrics to median change in metrics throughout the main text.

- Most of this paper is thematically consistent – examining the benefits of using a vector-based routing model with lakes across the continental U.S.. The second half of section 3.2, however, evaluates the raw performance of the model (with lakes) against CAMELS environmental variables. This would make sense if the authors evaluated the benefit of including lakes (i.e., shift in NSE) against these variables, but as is, this analysis is really an examination of the quality of CREST's runoff estimates. I suggest that if this is to be retained, this section be re-cast to answer questions such as "under what environmental conditions is the inclusion of lakes more likely to be beneficial", which can be done by (e.g.) performing the analysis of figure 7 with improvements to NSE rather than the raw score.

Response:

Thank you for your suggestions. As brought up by another reviewer, we think it is out of context for the scope of this study, so we deleted this section for brevity.

- The calibration in the Houston case study is missing important details. How many parameters were calibrated? Which ones? Using what objective function? It also uses an unconventionally small 1 year calibration period and 2.5 year validation/evaluation period, with no reference to a run-up period.

Response:

Thanks for your comments. We added more descriptions into our main text. For the split-sample selection, there is no standard strategy. The intention here is to include Hurricane Harvey (2017-08 to 2017-09) in the validation period. Obviously, different split-sample strategy can cause uncertainties in model performance. The model warm-up period is from 2015-06-01 to 2016-06-01.

L.235-236: "The NSCE is used as objective function for calibration, and model is warmed up for one year from 2016-06-01 to 2017-06-01"

- How were the model parameters determined for the CONUS application?

Response:

The CONUS parameters of CREST are based off previous publications in which parameter regionalization (random forest) is used after calibrating hundreds of basins in the US. We added more information in the main text.

L.216-217:"The parameters for IRF routing are based on default values provided by Mizukami et al. (2016), and the lake parameters, such as the outflow coefficient a and exponent b of eq.1, are based on suggested values in Döll et al. (2003) and Gharari et al. (2022)."

L.308-309: "For this case, the CREST-VEC model parameters are based on the pre-configured CONUS-wide parameters, the same as the ones used in Flamig et al. (2020)."

- The justification for selecting model populations is missing – why/how were the 283 models in section 3.3 selected? Are these all of the gauges downstream of natural lakes in the dataset? What distance threshold defines 'downstream'? How were the 5 local cases discussed in figure 10 chosen – given the minimum NSE difference is > 0.55, these are not random samples, but rather seem to be cherry-picked to illustrate the most successful inclusion of lakes. It would be useful to see at least one model where inclusion of lakes degraded model performance, with speculation as to what circumstances might cause this.

Response:

Thanks for your suggestions. Those filtered gauges are downstream of natural lakes, which are filtered by finding gauges located at reach that is downstream lakes. The distance is varying, depending on the length of reach. The five gauges are randomly picked as they are all downstream of large reservoirs. For your suggestions, we replaced Fig.3e with a case that inclusion of lakes downgrade the model performance (NSCE: 0.3 for results with lake and 0.4 for results without lake). We see that despite the baseflow simulated by model w/ lake is closer to recorded values, the flood peaks are underestimated. On the contrary, model w/o lake better captures the flood peaks. We described this in our main manuscript.

L.590-592: "Figure 9e shows that although the model with lake produces better baseflow, it underestimates flood peaks, resulting in lower NSE values (0.3) than results without lake (0.4). It implies that parameters governing the lake outflow need to be improved."

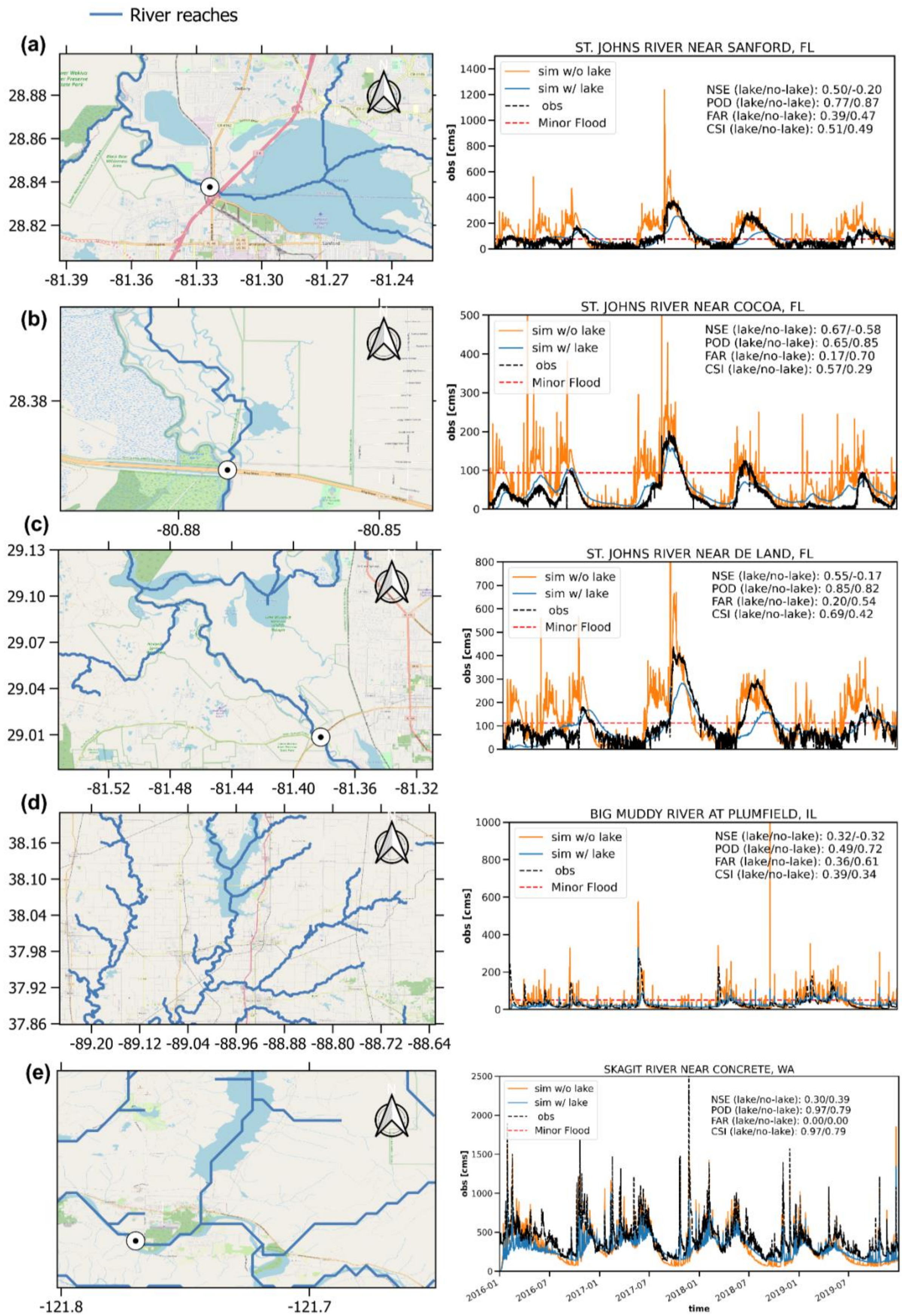


Figure 3. Five case examples of streamflow time series at gauges downstream of lakes: (a) St. Johns River near Sanford, FL; (b) St. Johns River near Cocoa, FL; (c) St. Johns River near De Land, FL; (d)

Big Muddy River at Plumfield, IL; (e) Skagit River Near Concrete, WA. Images courtesy of Google Map.

- The false detection analysis of section 3.3 also suffers from the inclusion of all models, regardless of quality. Why even analyse the false detection performance with an NSE of 0.1? These models are already not fit for the task, so including them in the analysis could very well skew the results such that they imply performance improvements with lake inclusion, even if this performance improvement is meaningless (a NSE of -0.35 is not functionally better than an NSE of -0.55).

Response:

Thanks for your comments. For Section 3.3, we selected gauges only downstream of lakes, where we expect to see improved performance by including lake routing. First, the median NSE (CC/Bias) for lake included results become 0.3 (0.7/0.1) and 0.1 (0.6/0.3) for results without lake. Second, the binary metrics POD, FAR, and CSI are closely related to CC and relatively insensitive to high flow magnitude, while NSE is highly favoured by high flow. So, the binary metrics offer an alternative yet important view to flood forecasting because the timing to detect a flood event is essential. Third, the intention of this framework is to develop an operational flood forecast system that weather forecasters use to issue warnings. A map of gauges with FAR could help us identify regions that model does not perform well, which is insightful for forecasters and decision makers.

We want to gently remind the reviewers/readers on the implication of a negative NSE, which suggests systematic bias but not necessarily poor CC. For a categorical detection of flood events (as indicated by the binary metrices), CC is as important as, if not more important than, bias. Therefore, NSE alone is not sufficient for making an informative judgement on the simulation performance, which is why we try to provide more metrics here.

- Supplementary materials should only be provided to corroborate existing evidence in the paper; the authors use Fig S1 to make a completely distinct point. This content should be either removed or incorporated into the main document.

Response:

Thanks for your suggestions. The supplementary Figure is now moved to the main text.

- The authors have multiple discussions of ideas that are very loosely related to the paper and/or not tied to any specific results herein, and should be removed:
    1. Ensemble forecasts at line 229
    2. Advocating for modular model structure at line 377
    3. Support for parameter regionalization line 388
    4. Two-way feedback between social systems and catchment signatures (ln 398)
    5. Future work on machine learning based reservoir operation simulation (line 402)

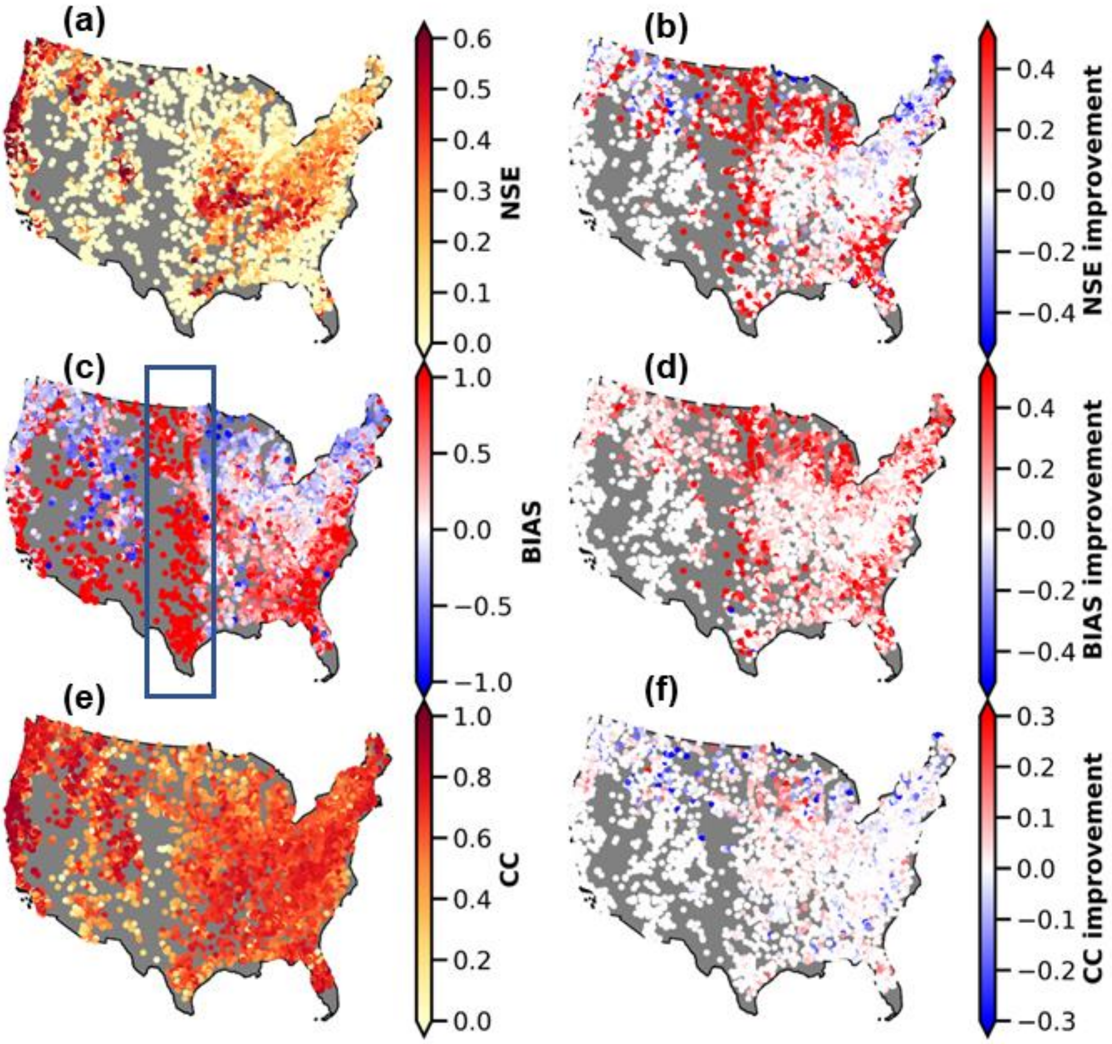
Response:

Thanks for your suggestions. We removed those discussion points mentioned above and expanded points (limitations and discoveries) that related to this study.

- The authors use percent change in NSE and BIAS to present results (e.g., in figure 6). However, this metric is very problematic for variables that can be positive or negative, because the denominator can go to zero. Another metric must be used.

Response:

Thanks for your comments. We changed the percent change to absolute change as attached below to replace the original Fig. 6.



- Multiple minor issues
    6. Bias is reported both as a percentage 0-100% and as a floating point 0-1 (fig 5)

Response: Thanks for your comments. We unified Bias as floating point throughout our main text.

    7. Speedup is usually used to evaluate improvements in computational costs; the authors here only report (less generalizable) differences in raw run times per time step.

Response:

The speedup is taken as the average run times per time step over the whole simulation period. We added one more row to compare the total computational costs.

Table 1. Statistical comparison of model performance over the continental U.S. Bolded numbers indicate the best metrics among the three model configurations. The computational speed is calculated as an average speed over a whole simulation period.

Metrics	Gridded CREST (Flamig et al., 2020)	CREST-VEC (w/o lake)	CREST-VEC (w/ lake)
Simulation resolution	1 km	**90 m**	**90 m**
Total computational cost (hours)	149.2	**29.9**	32.96
Computational Speed for routing (sec/step)	7.2	**0.35**	0.37
Max NSE	0.71	**0.87**	**0.87**
Median NSE	-0.06	0.12	**0.18**
% gauges NSE>0	41.8 %	50.6 %	**56.2 %**
Max CC	**1.0**	0.96	0.96
Median CC	0.40	**0.67**	**0.67**
Median bias	**9%**	27%	17%

8. The reasons provided for changing bias at line 247 are implausible – to influence bias you need to have water leave the domain by means other than streamflow. A more likely scenario is that this would be due to evaporation from the lake surface.

Response:

Thanks for your comments. Here the reduction of bias is measured at stream gauges, so if water is held in upstream lakes, streamflow measured at downstream gauges is much lower than simulation without lake. It is the primary (most obvious) reason for reducing the bias. Of course, there are other factors like lake evaporation or water abstraction, but those factors have so far not been considered in our continental simulation.

9. D-infinity (line 78) is also a grid-based (not vector-based) algorithm, and should be cited as Tarboton (1997)

Response:

Thanks for your suggestions. Yes, Dinf is a raster-based approach. Here we wanted to mention the vector-based routing can allow water infow or outflow from any direction. To avoid confusion, we deleted this sentence. For the reference you mentioned, we inserted in previous raster-based routing.

10. The CREST model original paper (Wang 2011) should be included upon its first mention. How is this paper not referenced?

Response:

Thanks for your comments. It was cited in original line L.129.

11. Many minor text errors not inventoried here given expectation of significant revisions

Response:

Thanks for your comments. After revisions, we carefully scrubbed the main text.

References

Han, M., J. Mai, B.A. Tolson, J.R. Craig, É. Gaborit, H. Liu, and K. Lee, Subwatershed-based lake and river routing products for hydrologic and land surface models applied over Canada, Canadian Water Resources Journal, doi:10.1080/07011784.2020.1772116, 2020

Tarboton, D. G., (1997), "A New Method for the Determination of Flow Directions and Contributing Areas in Grid Digital Elevation Models," Water Resources Research, 33(2): 309-319

Jiahu Wang, Yang Hong , Li Li , Jonathan J. Gourley , Sadiq I. Khan , Koray K. Yilmaz , Robert F. Adler , Frederick S. Policelli , Shahid Habib , Daniel Irwn , Ashutosh S. Limaye , Tesfaye Korme & Lawrence Okello (2011) The coupled routing and excess storage (CREST) distributed hydrological model, Hydrological Sciences Journal, 56:1, 84-98, DOI: 10.1080/02626667.2010.543087