This manuscript introduces a vector-based routing scheme (mizuroute) into the CREST hydrologic model. Additionally, the authors augment the routing scheme by adding a new subsurface routing scheme and a lake module. The new model is then tested for multiple scenarios. The manuscript is well-written and the conclusions drawn are consistent with the presented results. I have the following suggestions:

Response:
Thanks for your suggestions and comments. We have modified our manuscript as detailed below.

As it stands, the manuscript does not do a good job of separating the two distinct contributions - subsurface routing and lake module - in the text. The figures show clear delineation. I would suggest adding separate sections with detailed descriptions for both the subsurface routing scheme and the lake module.

Response:
Thanks for your suggestions. We have extended paragraphs to introduce the newly added lake module and subsurface routing.

In Section 2.4, we added following description of how to model subsurface flow.

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.

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

In Section 2.5, we have enriched descriptions of lake module:

L. 215-225: “We use the IRF scheme in this study for both terrain routing and channel routing in this study and activate the lake model with the Döll et al. (2003) lake model. The parameters for lake parameters such as the outflow coefficient $a$ and exponent $b$ of eq.3, are based on suggested values in Döll et al. (2003) and Gharari et al. (2022). For lakes that have monitored storage provided by the US Geological Survey (USGS), we directly insert storage time series into the model. As reservoir operation is not considered in this study, we exclude observed streamflow that is regulated by reservoirs and regulated lakes, as shown in Fig. 1c. So, only results from natural lakes, which account for 98% of US lakes or reservoirs, are considered valid for statistical comparison. To initialize model states, especially for initial lake volumes, we warm up the CREST-VEC model from 1948 to 2014 using the GLDAS forcing (Global Land Data Assimilation System) at a daily time step.

$Q_{out} = a \times S_f \times (S_f / S_{f,max})^b$,

where $a$ and $b$ are the outflow coefficient (1/day) and exponent, respectively; $S_f$ is the actual lake storage ($m^3$); $S_{f,max}$ is the maximum lake storage ($m^3$)."

The results section is well structured and clearly flows from regional application to a continental use case, and finally a flood forecasting example.
However, the discussion section is insufficient. Section 4.2 is largely unnecessary as the paper does not deal with hydrologic simulation at all, unless the authors want to test ensemble simulation with varying catchment processes.

Response:
Thanks for your suggestions. In a nutshell, we still consider this framework CREST-VEC as hydrologic simulation, since it has both water balance module and routing module. This section intends to lay out future strategies to improve large scale hydrologic simulation after we see commonly poor scores in hydrologic models in regions like the Great Plains. The ensemble simulation could make a difference but has been so far being ignored in large-scale hydrologic simulation. However, we agree that previous version overstated some sentences that are out of context, so we have shortened this section and only kept parts that are relevant to our study.

Additionally, more discussion of the results from the flood forecasting example is needed. The authors need to contextualize the results within the large body of flood forecasting literature. In addition, the reasons for the improved FAR and/or reduced POD is not adequately addressed. I would suggest providing concrete mechanistic reasons for both improved FAR and reduced POD.

Response:
Thanks for your suggestions. After integrating other three reviewers' comments, we realize that suggesting flood forecasting is off our intention and distract the main focus of this work. We chose not to expand on this topic. However, we agree on your suggestion that more mechanistic reasons for improved FAR and reduced POD are needed. We have made such revisions to our main text.

L.566-573: "The decrease in FAR values implies five instances: (1) decrease in false alarms while hits remain the same; (2) increase in hits while false alarms remain the same; (3) decrease in false alarms while increase in hits; (4) decrease in both false alarms and hits; (5) increase in both false alarms and hits. We find that however POD values decrease from 0.87 (without lake) to 0.85 (with lake), which indicates that both hits and false alarms are decreasing, but false alarms decrease at a higher rate. This is due to the reduced, simulated flood peak, and consequently less hits in flood forecasts but meanwhile less falsely alarmed floods. As most studies focus on flood detection, they inevitably arrive at more falsely detected floods. Too many false alarms could make people disregard the warnings, despite a real threat, causing the "cry wolf" effect."