The manuscript presents a machine learning framework that has already been used in practice to predict flood extent in the Sudd wetlands. The study addresses an important humanitarian problem, and the operational deployment is impressive. The two-stage design, with a basin-scale temporal model followed by a ConvLSTM spatial model, is a reasonable approach, and the seasonal differencing idea is also interesting. However, I have several important concerns about the methodology, especially about how the main claim, predicting "out-of-sample extreme values" (OSEVs), is presented and tested. In my opinion, some of these issues are serious and should be addressed before the paper can be published. I would recommend major revisions.

Below are my major comments:

1) This is my main concern: OSEV is defined in Eqs. 7–8 using the raw target, where any test value above 5σ from the training mean is considered extreme. Based on that definition, there are 35 such cases after 2019. However, the model is not trained on the raw target. It is trained and tested on the transformed seasonal anomaly, $\Delta\tilde{y}_t$. According to Figure 8c, there are no OSEVs in the test set under this transformed variable, and the training set even contains more extreme values than the test set in that space. So, the model is not predicting test cases that lie outside the training distribution. Instead, it is predicting values in a transformed space that appear to be within the training range, even though after reconstruction with Eq. 22 they become extreme in the original variable. That is a weaker claim than saying the model predicts truly out-of-sample extremes, but the paper often treats these two ideas as the same. Can the authors clarify the main claim here? Is it that (a) the transformation makes the test extremes look in-distribution to the model, which seems to be what is actually happening, or (b) the model truly extrapolates beyond the training extremes, which is what the title and abstract seem to suggest? Have the authors tested cases where $\Delta\tilde{y}_t$ itself is extreme in the test set? Those would be the real out-of-sample extremes for the model's actual prediction task.

2) The temporal model uses the previous 36 dekads of inundation as input. Since flooding in the Sudd is highly persistent and can last for years, the model already receives very high inundation values during the post-2019 extreme period. In other words, even if the model was not trained on these extreme periods, its inputs already show that the system is in a high-flood state. This is not exactly data leakage, but it does affect how the results should be interpreted. The model is not predicting an extreme event from normal-looking conditions. Instead, it is predicting the continuation of an already extreme situation. This may also explain why the persistence baseline performs almost as well as the transformer in Table 2. Have the authors tested the transformer without inundation history as an input? That would help show how much of the OSEV performance comes from other predictors, such as rainfall, lake levels, and climate indices, and how much simply comes from persistence in the inundation record itself.

3) Section 4.4 says that for deployment, the temporal and spatial models were retrained using all historical data, including the OSEV cases that were previously held out. This means the reported test results do not represent the same model that was later used to make the

2024/2025 forecasts in Figure 12 and share them with humanitarian partners. Has the deployed model been tested on any truly held-out data? If so, what was its performance?

4) Eq. 17: $L_t = MSE \cdot SignLoss + 0.1 \cdot SumLoss$. The SignLoss multiplier of 20 (Eq. 18) is unusually heavy and the 0.1 weight on SumLoss appears arbitrary. How were the constants 20 and 0.1 chosen? Was a sensitivity analysis performed?

5) Only one train/test split was used, with training before July 21, 2018 and testing after that date. Given the small effective sample size, this is a concern. Did the authors use any time-series cross-validation, such as a rolling-origin test, to see how sensitive the results are to the choice of cutoff date? Furthermore, the paper says that 10% of the training data were used for validation with a random split. In a time-series setting, this can cause leakage between training and validation. Was the validation set separated by time instead?

6) The paper compares the model with persistence, linear interpolation, linear regression, lasso, random forest, and FFNN, but it does not compare it with some other important baselines. These include a standard LSTM or GRU, which are common in flood prediction studies and are also used in works cited by the paper, such as Google's AI Flood model and Frame et al. (2022). It also does not compare the results with GloFAS or another physically based model for the same region. In addition, Section 2.2 discusses the FEWS NET LASSO plus constant-fill method, but the paper does not provide a direct comparison with that approach on the same test set. It would also be useful to compare against a simpler transformer model, without the custom loss or Monte Carlo dropout, to better understand which design choices really matter. Why was an LSTM not included, given how common it is in flood prediction research? Can you run a head-to-head comparison with FEWS NET's actual outputs for the operational period, since both are operational systems addressing the same problem?

Minor comments:

Section 7.1 says that this approach could also be useful for forecasting other spatio-temporal hydrometeorological events, such as rainfall, cyclones, and heatwaves. This feels like too strong a claim based on the evidence presented in the paper. Those processes behave very differently from flood extent in the Sudd, which changes slowly over time and has strong persistence. That is also one reason why the persistence baseline works so well here. Can the authors either soften this statement or provide stronger evidence that the seasonal differencing approach also works for faster-changing processes, where the target has much weaker autocorrelation?

The 70 HydroATLAS features are PCA'd to 16 components, but only the first PC is used in the spatial model. Why retain 16 in the pipeline if only 1 is used? Was the number of PCs itself tuned?

Page 4 line 1: "Slater et al., 202" truncated citation.

Page 12 line 9: "are primarily used or monitoring" has to be "for monitoring"

Section 5.2 is missing.

"CHRIPS" appears multiple times as a typo for "CHIRPS"

Many figures lack units, axis labels, or sufficient caption detail to be self-contained. Figure 13 in particular needs y-axis units.