General Comments

In their manuscript Hunt and Harrison provide a novel approach to reconstructing historical monsoon variability: using machine learning to assess the relationship between paleoclimate records as predictors and the IMD gridded rainfall dataset between 1901-present as the target, and then extending that record back through time using annually resolved paleoclimate records from tree-rings, speleothems, glaciers etc. The results are interesting and appear to provide a significant step forwards monsoon predictability, comparable or slightly better than previous techniques in developing time series at point locations or areal means, but a significant advance in reconstructing spatial heterogeneity of monsoonal variability. The section on assessing the importance of individual paleoclimate records in the model, and the 'sphere of influence' of individual sites is an exciting development and appears to be backed by our understanding of monsoon dynamics.

The manuscript is well written, and the language is largely appropriate but could be simplified in places to the intended audience of Climate of The Past who likely have a more limited understanding of machine learning techniques. This includes more high-level introduction to methods, and care explaining predictors, targets, training datasets, validation datasets, test datasets etc. There is also sometimes a disconnect between what is labelled on the figures, the figure captions and what the figures refer to. Consistency and clarity are needed in places.

I do not have any significant concerns over the science and resulting inferences presented in this manuscript, though I will happily defer to more qualified machine learning reviewers on the robustness of the methods. I believe that this paper is suitable for publication in Climate in the Past, after correction for some minor issues.

Thank-you

Nick Scroxton

Maynooth University

We would like to thank Nick for his positive assessment of our manuscript, and for his comments, which we respond to point-by-point in red below. Textual revisions to our manuscript will be highlighted in blue.

Specific Comments

- Could you explain LiPDs in either section 2.1.1 or 2.1.2 before you get to line 148.
  Thank you for the suggestion. We have added the following at the end of Sec 2.1.3: "All data from this database, as well as from PAGES2k and Iso2k, are available as Linked Paleo Data (LiPD) files. These provide a wealth of metadata and a standard format that makes them machine-readable."
- Section 3: A high-level couple of sentences at the start of the methods would help non-machine learning experts. For example, it doesn't actually say in the manuscript what is your predictor data-set is and what is the target data-set.
  We agree it would be useful to have a general introduction at the beginning of the methods. We have added the following at the beginning of Section 3: "In this study, we aim to reconstruct historical monsoon rainfall over India over the last 500 year using

deep learning models. Our predictor dataset – i.e., the model input – comprises a wide range of palaeoenvironmental records (Sec. 2.1) interpolated to annual resolution. Our target datasets – i.e., what we want the model to predict – are derived from observed monsoon rainfall (Sec. 2.2) The models were trained and tested on replicating the target datasets over their lengths, and once testing confirmed their predictions were robust, their predictions were extended backwards in time to 1500. To achieve this reconstruction, we tested two different architectures of model. The first was a regional model, built using a dense multilayer perceptron, and was trained to replicate longer instrumental timeseries of precipitation over the homogeneous regions (Sec. 2.2.3). The second was a spatial model, built using a decoding convolutional neural network, trained to replicate gridded precipitation data (Sec. 2.2.1). To avoid overfitting on small training datasets, we employed a range of common techniques including bagging, dropout, and regularisation. Finally, to understand how the models make their predictions given certain inputs, we employed an explainability method known as Shapley analysis. These models and techniques are described in greater detail in the sections below."

- What's the difference between validation and test datasets in Figure 2 and section 3.5 1a. What do they do and why?
  This is already explained briefly in Sec 3.4 (part 2.a.iv of the recipe). For clarity, we have also added this information to Sec 3.1, where the term "validation" is first introduced, in our revision: "The purpose of having distinct validation and training datasets also arises from the desire to avoid overfitting. As the models are trained, the loss function is computed on both the training dataset (which the training process is designed to minimise) and the validation dataset. If the validation loss starts to increase while the training loss continues to decrease, that is a sign that the model is starting to overfit. However, because the model has thus been tuned on the validation data, a true fair test requires that it is distinct from the test dataset, which must remain hidden from the model."

- Figure 3: This figure does not show clearly the information that is attributed to it. There are significant mismatches between the figure and figure caption. I don't see the ensemble median in black, the spread in red, or the Sontakke time series in green.
  Thank you for noticing this. This caption referred to an earlier version of the figure. We have now updated it: "Estimated seasonal precipitation anomalies from the regional model ensembles. For each of the seven homogeneous regions as well as all India (AI), the ensemble median is given by the coloured bars. Observed values, taken from the reconstructed timeseries in Sontakke et al. (2008), are given to the right of the green line. Where standardised anomalies lower than −0.5 occur in either the modelled or observed timeseries co-occur with known regional or national famines, these are marked with grey bands. Stated *r*-values measure the correlation between coincident actual and model test values."

  I think this figure should be expanded to take-up more space on the page to really highlight the key results attributed to it.
  We agree and have made this change.

- Line 380-383: The dismissal of the PCA technique is too strong at this stage of the manuscript. At this point the reader has only been introduced to figure 4. The PCA outperforms the CNN model 40% of the time (2 out of 5 years) in figure 4 so cannot be dismissed, although an argument can be made that it lacks spatial heterogeneity. Once we have seen figure 5 and the larger dataset, we see that outperformance of the PCA method in figure 4 is likely just an artifact of small sample size, and therefore the dismissal is more reasonable. This section therefore might need to be reworded
  This is true. We have added the following at the end of this section: "However, this is just one ensemble member compared across just five seasons. While the mean value of *r* is higher in the CNN (0.40) than the PCA (0.31) and linear (0.25) models, the latter two do beat the CNN model in two of the five years. Therefore, as a further test…"

- I wonder if dry years also correspond to major volcanic eruptions (as we might expect). This might provide an additional test of reconstruction performance that is less reliant on additional human societal complexities.
  Thank you for this very useful suggestion. We have created a new figure (Figure 8) which takes the area-averaged timeseries from what was Fig 7c and shows it alongside a popular ENSO reconstruction and historic large eruptions (VEI>4). We have added some new text describing the relationships that emerge: "The reconstructed ENSO anomalies (Fig. 8(b)) have a correlation with monsoon anomalies that varies centennially: the rolling 30-year correlation (not shown) is negative throughout almost all of the 18th and 20th centuries, but is positive in the 17th and early 19th centuries. This pattern is consistent with previous studies (Shi and Wang, 2019), in which it is speculated to be modulated by the Pacific Decadal Oscillation. Similarly, large volcanic eruptions (Fig. 8(c)) have a significant impact on the reconstructed monsoon anomalies. Lagged composites of monsoon rainfall for each VEI (not shown) suggest the impact on the monsoon grows with increasing VEI. In year 0 of VEI5 events, the mean reconstructed monsoon rainfall anomaly falls to −0.11, then to a minimum of −0.18 in year 2, before recovering by year 4. The pattern is different in VEI6 events, which initially cause an increase in monsoon strength, reaching a maximum mean anomaly of +0.25 in year 2. This then starts to fall and becomes negative by year 5, reaching a minimum of −0.56 in year 8, after which it recovers. These coherent responses of the reconstructed monsoon to volcanic eruptions gives us further confidence in our reconstruction."

- I disagree with some of your speleothem inferences. In section 4.3 If the EPI speleothem record is d18O then we might expect a better correlation with WPI rainfall than EPI, and thus this example belongs in the following paragraph instead. I don't see the relevance of the river basins argument here on speleothem proxy variability.
  We have not expressed this argument clearly. We agree that the river basin argument is not directly relevant. The EPI speleothem record is d18O from Jhumar Cave. Sinha et al. (2011) have shown that there is a significant inverse relationship between the changes in 18O and regional rainfall in the observational period and have argued that the variability in this record therefore reflects changes in both local and upstream rainfall. Thus, it is plausible that the record is influenced by the heavier monsoon rainfall of the WPI, as implied by the Shapley analysis. We have revised the relevant text so that it now reads: "An example of this is the single speleothem record from EPI (Jhumar Cave), which shows poor predictability for EPI though it is a useful predictor for the neighbouring WPI

region. The lack of predictive power may be because the record does not have a strong causal relationship with regional monsoon rainfall or because the record is not of high enough quality, e.g., because of low resolution. Shapley analysis implies the former case for the EPI speleothem. Sinha et al. (2011) have argued that the variability in this record reflects changes in both local and upstream rainfall and thus it is plausible that the heavier monsoon rainfall of the WPI may end up influencing speleothems in EPI."

Technical Corrections

- Line 196-198: Could you unreversed the first half of this sentence for clarity
  We have rewritten this sentence: "Bagging promotes diversity among models by training each one on a different dataset. This diversity causes the errors of each model to be at least partially orthogonal and as a result, averaging the errors across the ensemble reduces the overall variance."
- Figure 2 caption: 'one of the timeline model'
  Thank you, we have fixed this.
- Line 354: New paragraph?
  We have made this change.
- Line 364: 1988?
  This should read 1986. This has been corrected.
- Line 380: New paragraph?
  We have made this change.
- Line 528, 542: The cave name is spelt 'Mawmluh' or 'Krem Mawmluh" if cave needs to be included.
  Thank you for this suggestion. We originally followed the spelling given in Kathayat et al. (2022), but now understand this is nonstandard and have corrected our manuscript accordingly.
- Line 530: 'documented'?
  Thank you, we have made this change.
- Line 531: 'subcontinent' to be consistent throughout the manuscript.
  We have replaced "continent" here with "subcontinent". That was the only such instance we found in the manuscript.
- Line 561: clarify what you mean by wavelength or use less technical language.
  We have replaced "shorter wavelength signals" with "finer spatial detail".
- Line 578: delete 'have'
  Fixed.