Response to Reviewer

November 22, 2024

Detailed Comments
Responses are marked in blue.

Anonymous Referee #1, 24 Oct 2024:
This is an interesting application for building sub-seasonal models but I have several concerns that would be good to address before it is published.

Dear Referee,

Thank you very much for your positive feedback and for the time and effort you dedicated to reviewing our manuscript and dataset.

We greatly appreciate your insightful comments, which have been invaluable in guiding improvements to our manuscript. Please find below a detailed, point-by-point response outlining our approach to addressing each of your suggestions. If you feel any of these adjustments might not fully meet the needs you highlighted, we would be grateful for further guidance.

Kind regards,
Víctor Galván (on behalf of the author team)

Major Corrections

1. It is not clear to me what the focus of this paper. Is it to present this new framework and the model you have trained is just an example of an application that could be done with the new framework? Or is the idea to present this new model? If the former, do you have plans to extend this into short lead time weather forecasting? It seems to me that most of what you have developed e.g. hyperparameter tuning and XAI could be useful here. If the latter then I think it would be good to have a better description of the model.

Thank you for your valuable comments on the focus of the paper. The primary objective of the paper is to present a framework that enables users to develop deep learning models with a specific focus on sub-seasonal and seasonal forecasting. The principal benefit of this framework is that it enables the user to undertake preprocessing, training and validation of the model in a straightforward manner. To exemplify this, we have selected a case study in which the Pacific sea surface temperature anomalies in September and October are employed as the predictor field, while the global sea level pressure anomalies in November and December constitute the predictand field. At the outset, the intention was not to extend this model to encompass short-lead-time weather forecasting. Nevertheless, if required by the user, an updated version of the model could be developed to enable this, thereby facilitating a comparison of its performance with that of other existing dynamic and machine learning models. However, this is out of the scope of the present study.

In this new version, we have highlighted the main goal of the paper and emphasised the flexibility of the tool presented to develop deep learning models (Lines 83-88):

“For these reasons, we developed the Neural Network foreCAST (NN4CAST) application, a Python library designed to facilitate the creation of deep learning models for non-linear modelling of climate teleconnections. One of the main objectives of NN4CAST is to avoid treating these deep learning models as “black boxes”, enabling users to analyse the origins of the predictability and assess the sensitivity of predictions to changes in the training period.”

2. As you mention in the introduction, sub-seasonal forecasting is very uncertain. I think for this framework to have a significant impact, it would need to be able to include a way to quantify uncertainty. For example, allowing multiple initial conditions, injections of Gaussian noise or the generation of ensembles.

We would like to thank you for this insightful comment. We are in full agreement with the proposal to provide a measure of the uncertainty of the model predictions. This might take into account both potential errors in the modelling of the teleconnections and the ability to analyse non-stationary behaviours of the teleconnections. In terms of the methodology for quantifying this, a new function (based on the existing “Model_build_and_test” one) has been created that allows the generation of ensembles by training the model in different periods of the training set. This enables the sensitivity of the model to these periods in the different regions to be evaluated. Specifically, what this new function does is to create a given number of models that differ only because they have been trained on different datasets, using the bagging (bootstrap aggregation) method. In this way, it quantifies the uncertainty associated with the period chosen to train the model and how this depends not only on the region we are analysing, but also on the teleconnection mechanisms involved. However, this function has not been implemented in the library, as it can be easily created and tailored by users to suit their specific needs, optimising it for the particular region and teleconnection mechanisms under analysis.

Furthermore, another method for introducing uncertainty into the model is to vary the initialisation seed of the trainable parameters of the model, which is already incorporated into the application. Nevertheless, our analysis of the case study revealed that variations in the seed do not result in significant alterations to the model predictions. Nevertheless, in other scenarios, this could prove to be a more pertinent consideration, as evidenced in the article of Scher, S., & Messori, G. (2021), where they tested different ways to generate ensembles, and how this could lead to better overall performances and uncertainty estimates. Furthermore, as you have stated, uncertainty can be estimated by utilising multiple initial conditions or by incorporating noise into them. This issue has been addressed by allowing the user to modify the predictor field files before introducing them as inputs to the model, despite the absence of a predefined function for this purpose within the application. Basically, the user just needs to add some noise to the data files (predictor and/or predictand) before doing the preprocessing phase.

We have highlighted in the discussion the possibilities of the model to perform sensitivity experiments by not only using different predictor and predictand fields, but changing the regions, the datasets and introducing noise to them to quantify the uncertainty (Lines 362-366):

"The NN4CAST framework also supports the development of sensitivity experiments, allowing users to explore not only different predictors and predictands, but also variations in the regions of them, different datasets and the introduction of noise. For instance, an attempt was made to introduce white noise into the predictor during the test period. However, this resulted in only minor alterations, indicating that the trained model is resilient to this type of noise in this specific case. These capabilities significantly enhance its versatility for exploring and understanding climate predictability."

3. It is unclear to me who provides the model? Are there example models provided in the repository? Or can the user prepare their own models and what format should they be in? Torch/tensorflow?

We are grateful for your observation regarding the construction of the model. The principal benefit of this framework, which to our knowledge is unique at least at seasonal timescales, is that the model code does not need to be programmed; it is fully implemented. The user is required to select the predictor and predictand, provide the initial hyperparameters (number of convolutional layers, activation functions, etc.) and then, the model will be created, trained and validated based on this information.

This application is founded upon the utilisation of pre-existing libraries, including TensorFlow and NumPy. Nevertheless, any neural network-based model constructed with alternative libraries, such as PyTorch, can be employed within the application. In this instance, the user would be required to program the model and utilise the library preprocessing and validation functions to assess the overall performance of their model.

4. For the model you present, I am not convinced that cross-validation is appropriate across an annual timescale. Is the idea to make the model robust against climate change? We know that ERA5 is also worse pre-1979 because of the lack of satellite observations.

We highly appreciate your contribution to the discussion. In this instance, detrending was conducted using the backward moving average method (using a sliding window of 50 years) in both fields with the objective of evaluating the predictive capacity of the model in terms of the internal variability of the climate system. The quality of data from the pre-satellite era, while subject to limitations, including reduced confidence in assimilated observations within ERA5, still allows for the detection of low-frequency signals. This is evident from the consistent oscillations observed when training with periods both before and after the 1970s.

In addition, and taking the above clarifications into account, the model allows the user to test the model for different periods changing the number of folds ,and not just a leave-one-out cross-validation. Figure 4 in the text (attached below) allows for a comparison of the model skill in terms of the anomaly correlation coefficient (ACC) between the predictions and the observations in the different folds of the cross-validation. To illustrate, in fold 4, where the model was trained on data from 1940 to 1999 and tested on data from 2000 to 2019, it exhibited a significant positive correlation not only in the tropics but also in the extratropics. This is despite the fact that the model was trained on data from a period with potentially inferior data quality. It may therefore be concluded that despite the potential errors in the data, the model is capable of learning the underlying mechanisms from these signals and extrapolating them to new unseen cases during training.

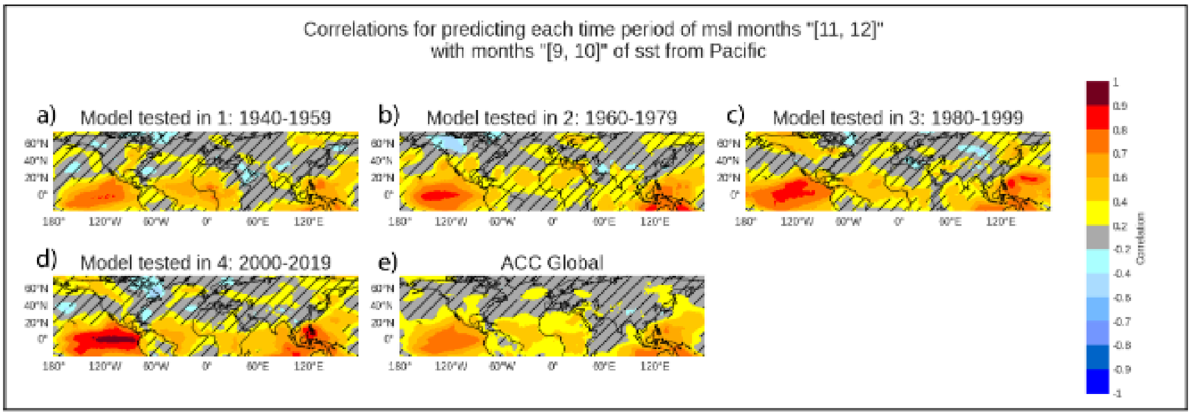
Correlations for predicting each time period of msl months "[11, 12]"
with months "[9, 10]" of sst from Pacific
a) Model tested in 1: 1940-1959
b) Model tested in 2: 1960-1979
c) Model tested in 3: 1980-1999
d) Model tested in 4: 2000-2019
e) ACC Global
60°N
40°N
20°N
0°
180°
120°W
60°W
0°
60°E
120°E
1
0.9
0.8
0.7
0.6
0.4
0.2
-0.2
-0.4
-0.6
-0.7
-0.8
-0.9
-1
Correlation

Figure 4. Panel of model performance in terms of spatial ACC over the 8-fold cross-validated periods. Concretely, each map depicts the spatial ACC between the predicted and observed ND SLP anomalies in the decades described, which is the one tested in each iteration during the cross-validation. The significant results of the simulation according to a T-test and the significance threshold defined in the hyperparamenters (in this case, 90% significance) is given by the non-dashed regions plotted in panel a) and the values above the dashed line of panel b

Minor

1. I think in the discussion around line 65, it would be good to mention Neural GCM as an example of an effective hybrid model.

We are pleased to receive this appreciation. We agree that this new model (Neural GCM) serves as an excellent illustration of how a hybrid deep learning model, trained to make short-term weather predictions, is capable of making forecasts on seasonal and decadal timescales for different atmospheric variables. Furthermore, it serves to elucidate some challenges of running this type of hybrid models at longer timescales (numerical instabilities and climate drifts). For this reason, in this new version, we have added this example in line 67:

“Some examples of these data-driven models are: PanguWeather, which is a 3D Earth-specific transformer module created by the Huawei Cloud group (Bi et al. (2022)); GraphCast and NeuralGCM, developed by researchers from DeepMind and Google (Lam et al. (2022); Kochkov et al. (2024)).”

2. I am not sure Lines 73-74 follow. You say the models are largely linear and then you say that this is important for non-linear relationships?

We will try to provide a more detailed clarification of this sentence in the text. However, the underlying concept is that classical statistical models have the limitation of assuming a linear relationship between predictor and predictand variables (Wilks, D. S. (2011)), which is a significant drawback when modelling complex systems such as the climate system, where the processes' interactions are highly non-linear. In this new version, we have clarified this topic in Lines 73-75:

“The underlying assumption of linear relationships between predictor and predictand fields is a common premise in these models. However, this assumption may not be entirely appropriate when modelling the Earth system, which is mainly composed of complex non-linear components.”

3. I think large parts of Section 2 could be removed. The basic theory of neural networks does not need to be included in a paper.

Thank you for your comments. The principal aim of this application is to provide a tool that enables climate experts to utilise non-linear modelling through the application of deep learning techniques, without the necessity to programme these models from their fundamental principles and without the requirement of a comprehensive understanding of this field. This section is designed to provide an overview of the fundamental principles involved, with the aim of ensuring that all users of this application are able to understand the various concepts associated with such techniques. However, in light of the comments received, some of the basic theory has been reduced in length and the reader is directed to literature on the subject for further information.