AO03: Predicting regional precipitation response to regional sea surface temperature anomalies

Candidate number: 1041620

Supervisors: Andrew Williams, Dr. Duncan Watson-Parris

Abstract

Sea surface temperatures (SSTs) are known to drive both local and non-local precipitation changes, and these relationships have important implications for weather forecasting. To this end, the skill of various linear and non-linear approaches in predicting the response of regional precipitation to regional SSTs is evaluated. A large ensemble of atmosphere-only general circulation model (GCM) simulations, forced by random SST perturbations, is used to train and test these methods. A simple convolutional neural network (CNN) is found to outperform traditional Green's function approaches in predicting precipitation anomalies generated by the GCM simulations. The computed relationship between SST and precipitation anomalies is also employed to reconstruct historical precipitation from a long observational SST dataset. Skilful reconstructions are made using linear and piecewise Green's function methods, but not using the CNN. It is suggested that, although the relationship between SSTs and precipitation anomalies may be non-linear, complex non-linear models trained on random SST fields are less robust for reconstruction and forecasting.

1 Introduction

Changes in sea surface temperature (SST) patterns can drive changes in many atmospheric variables, including regional precipitation. In particular, local SST anomalies can induce local precipitation anomalies by convection. The resulting vorticity anomalies can then excite Kelvin and Rossby waves, transporting climate signals and driving non-local precipitation changes [1]. The El Niño-Southern Oscillation (ENSO) phenomenon is a well-known example of these remote 'teleconnections', linking the atmospheric variability in much of the tropics and sub-tropics to anomalous SSTs in the Pacific. Understanding how to better model these teleconnections may be key to improving the accuracy of numerical weather forecasts.

Many studies in the field have investigated the response of atmospheric variables to SST forcings using a linear Green's function approach. Barsugli and Sardeshmukh [1] were among the pioneers of this technique, which consists of determining 'sensitivity maps' of atmospheric anomalies to SSTs. To do this, they performed a series of 'SST patch perturbation' experiments using an atmosphere-only general circulation model (GCM). By varying the location of the SST patches, they estimated the linear sensitivity of various target variables, such as precipitation, to regional SSTs. However, the patch method is limited by a low signal-to-noise ratio as the area-integrated SST anomaly in each patch is small. Later studies such as Li and Forest [2] and Baker et al. [3] used a random perturbation method to generate SST perturbation fields at a global scale, increasing the area-integrated SST anomaly and thereby the signal-to-noise ratio.

Once the linear sensitivity map of an atmospheric variable to regional SSTs is established, the historical response of this variable over a given time period can then be reconstructed purely from observed SST data. For example, Baker et al. [3] determined the sensitivities of the North Atlantic Oscillation (NAO) to SSTs and skilfully reconstructed historical time series of the NAO pressure difference using historical SSTs. Tsai et al. [4] applied the approach to precipitation in selected river basins and showed that the reconstructed responses correlated well with observational data between 1950 and 2000.

However, the linear approach does not always guarantee reconstruction skill. This is partly due to the limitations of using an atmosphere-only GCM to model a system with atmosphere-ocean coupling [3] and the existence of factors other than SSTs that drive precipitation changes. But a major reason appears to be that the linear approach fails to take into account the non-linear processes involved. In the case of precipitation, it is well known that there is a critical SST threshold above which deep convection may be triggered [5]. This strongly suggests that the response of precipitation to SST forcing is non-linear and motivates the use of a more complex model that captures these non-linearities. In particular, state-of-the-art deep learning models, such as convolutional neural networks (CNNs), have been shown to outperform traditional dynamical systems in forecasting ENSO [6] and also could prove useful in the context of predicting the response of regional precipitation to regional SST anomalies.

In this report, we benchmark the performance of two non-linear models, a piecewise linear model and a CNN, against that of the linear model in predicting regional precipitation response to SST anomalies. In § 2, we describe the datasets used in the study, define the models mathematically, and discuss the motivation for them in more detail. In § 3, we evaluate the models' performance on two key problems: validating the output of GCM perturbation simulations, and reconstructing regional precipitation between 1900 and 2010 from historical SST data. Here, we also discuss the limitations of the models. Finally, in § 4, we summarise key findings and suggest possibilities for future research.

2 Methods

2.1 Datasets

Following Baker et al. [3], the main dataset is generated from a 5544-member ensemble of perturbed SST runs on the HadAM3P GCM developed by the Met Office. Each ΔSST field is initialised as a 16-by-16 matrix with values chosen randomly from a uniform distribution satisfying $-2\,\mathrm{K} < \Delta\mathrm{SST} < 2\,\mathrm{K}$. Using bilinear interpolation, the matrix is mapped onto a latitude-longitude grid between 60°N and 60°S and then added to climatological monthly-varying SSTs calculated from 1980-1999 in the merged Hadley-NOAA optimum interpolation dataset [7]. For each ensemble member, a control simulation is run on the unperturbed SST field, followed by a forced simulation on the perturbed field. The resulting field of precipitation anomalies $\Delta P(\mathbf{x})$ is defined as the difference between the forced and control simulations. In this way, we obtain 5544 mappings between ΔSST fields and the corresponding ΔP fields. 4435 of these mappings (80%) are used to train the models, and the remaining 1109 (20%) are used for validation.

In addition to validating the mappings generated by the GCM, we also wish to reconstruct precipitation time series from historical SSTs. Here we use the Met Office's Hadley Centre SST dataset (HadSST3) from 1900 to 2010 [8], and compare the models' reconstructions against precipitation data from the ECMWF twentieth-century reanalysis (ERA-20C) [9]. A reanalysis dataset combines past climatological observations with modern forecasting models, thereby correcting for missing and erroneous data. We choose the ERA-20C in particular due to the long time period available and the atmosphere-only GCM on which it is based, which matches our random SST field experiments.

2.2 Linear model

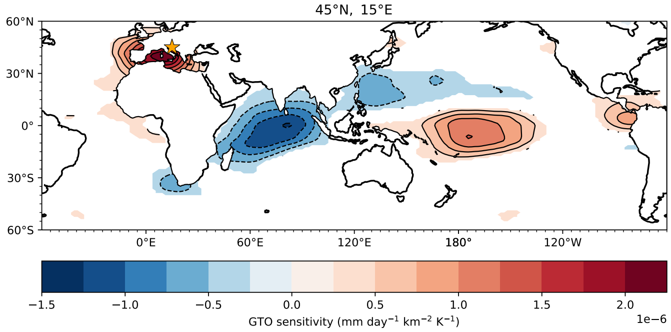


Figure 1. Example GTO for the precipitation grid point centred at $j = 45°\mathrm{N},\ 15°\mathrm{E}$ (near Croatia). SST anomalies in red (blue) regions are positively (negatively) correlated with precipitation anomalies at j. The grid point centre is indicated by the orange star symbol. The example illustrates how regional precipitation is influenced by both local SSTs and remote teleconnections.

The precipitation anomaly $\Delta P(\mathbf{x})$ at a particular location $\mathbf{x}$ in response to sea surface temperature anomalies $\Delta\mathrm{SST}(\mathbf{x}')$ at locations $\mathbf{x}'$ can be described by

$$\Delta P(\mathbf{x}) = \int G(\mathbf{x},\mathbf{x}')\Delta\mathrm{SST}(\mathbf{x}')\,dA' \; + \epsilon, \qquad (1)$$

where $G(\mathbf{x},\mathbf{x}')$ is a Green's function known as the global teleconnection operator (GTO), and ϵ is an error term accounting for deviations from linearity [1–3]. Intuitively, the GTO can be understood as a map of sensitivities to SST forcing at each precipitation grid point. In practise, we are using an SST grid with finite resolution, so Eq. (1) is discretised:

$$\Delta P_j = \sum_k G_{jk}\Delta\mathrm{SST}_k A_k, \qquad (2)$$

where G_{jk} is the relationship between the precipitation response at point j to the SST anomaly at point k, and A_k is the area of the grid point at k. Note that we have dropped the ϵ term as a first order approximation and that there is no requirement for the grids $\{j\}$ and $\{k\}$ to be identical. Here, the SST grid $\{k\}$ consists of 97 latitude points × 192 longitude points between 60°N and 60°S (excluding grid points over land). On the other hand, the precipitation grid $\{j\}$ spans all latitudes between 90° N and 90° S and is coarsened to 18 latitude points × 36 longitude points for ease of computation.

Following Baker et al. [3], a linear regression is performed between the precipitation anomalies at j and the SST anomalies at k to calculate G_{jk}:

$$G_{jk} = \frac{\mathrm{Cov}[\mathbf{\Delta SST_k}, \mathbf{\Delta P_j}]}{\frac{T_{\max}^2}{3} A_k}, \qquad (3)$$

where $\mathbf{\Delta SST_k}$ is a vector of the 4435 training SST anomalies at k, $\mathbf{\Delta P_j}$ is a vector of the 4435 corresponding precipitation anomalies at j, $T_{\max} = 2\,\mathrm{K}$ is the maximum SST perturbation, and A_k is the area of the grid point k. The derivation of the prefactors does not concern us here, but can be found in [2]. As a regularisation step, we set $G_{jk} = 0$ if the correlation is statistically insignificant by a two-tailed t test at the 5% significance level. A typical example of a calculated GTO for a point on the precipitation grid is shown in Fig. 1, illustrating how regional precipitation is sensitive to both local and non-local SST changes. In particular, we see that anomalous precipitation in Croatia increases in response to anomalous Mediterranean SSTs and decreases in response to anomalous Indian Ocean SSTs. The latter relationship is an example of a remote teleconnection.

Once we have obtained the GTO, Eq. (2) can be used to reconstruct the precipitation anomalies to a particular ΔSST field, up to a constant scaling factor. We first perform this reconstruction on the validation set of 1109 ΔSST fields to check how well the linear model reproduces the GCM-generated ΔP fields. We then use a time series of historical SST data from the HadSST3 dataset to obtain a reconstructed time series of ΔP_j from 1900 to 2010 for each grid point j. Note that, for each year n of the HadSST3 data, we define the SST anomaly to be $\Delta\mathrm{SST}_{k,n} = \mathrm{SST}_{k,n} - \overline{\mathrm{SST}}_{k,1980-99}$ where $\mathrm{SST}_{k,n}$ denotes the mean annual precipitation in year n, and $\overline{\mathrm{SST}}_{k,1980-99}$ denotes the mean SSTs between the years 1980 and 1999, inclusive. This is chosen to be consistent with the SST fields used in the control GCM simulations. Finally, the reconstructed ΔP_j time series is regressed onto the ERA-20C reanalysis time series using the equation

$$P_{j,\mathrm{reanalysis}} = \alpha_j \Delta P_{j,\mathrm{reconstructed}} + \beta_j, \qquad (4)$$

where α_j and β_j are coefficients to be determined. Note that the reconstructed precipitation anomalies can directly be regressed onto annual precipitation reanalysis data, since the baseline values can be absorbed into the regression coefficients. The Spearman's rank correlation coefficient r then provides a simple measure of the reconstruction skill at each grid point j.

2.3 Non-linear models

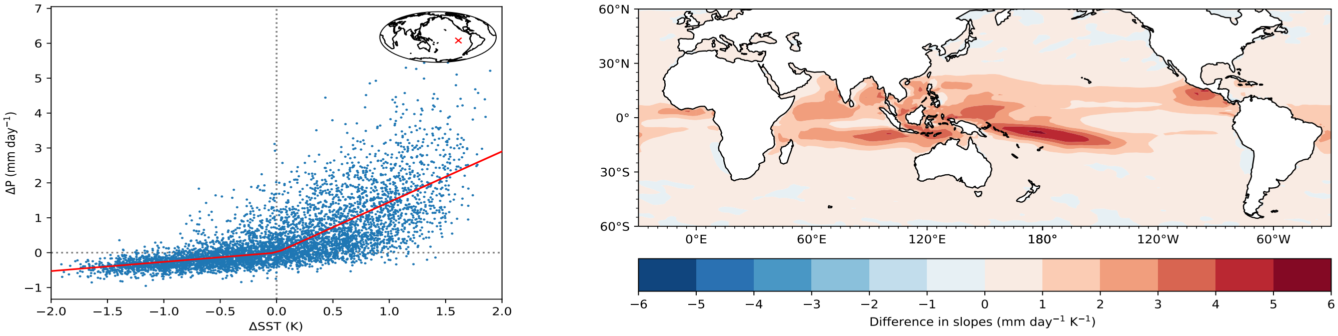
2.3.1 Motivation

Although previous studies have demonstrated good validation and reconstruction accuracy using the linear GTO, there is evidence to suggest that a non-linear model may perform better. For example, local correlations between ΔP and ΔSST appear to deviate significantly from linearity. For many grid points in the tropics, such as that illustrated in Fig. 2a, the precipitation response is notably more sensitive to a positive SST perturbation than to a negative one. We can obtain a simple measure of non-linearity by fitting a piecewise linear function through the origin and taking the difference in slopes for $\Delta\mathrm{SST} > 0\,\mathrm{K}$ and $\Delta\mathrm{SST} < 0\,\mathrm{K}$. A plot of these slope differences for local ΔP vs. ΔSST correlations is shown in Fig. 2b.

We see that the non-linearity is greatest in the tropics and not as significant in the temperate and polar regions. This appears to be consistent with a temperature threshold $T_c \approx 27.5$°C for large-scale deep convection as outlined in Graham and Barnett [5], since the mean SSTs in the tropics lie near this value. A simple argument for this threshold behaviour is as follows (cf. the C5 lectures). Consider the acceleration B due to buoyancy of a fluid parcel with density ρ_{parcel} at a level with ambient density ρ_{ambient}:

$$B = \frac{\rho_{\mathrm{ambient}} - \rho_{\mathrm{parcel}}}{\rho_{\mathrm{parcel}}} g = \frac{T_{\mathrm{parcel}} - T_{\mathrm{ambient}}}{T_{\mathrm{ambient}}} g, \qquad (5)$$

where we have used the ideal gas law $p = \rho RT$ to convert densities into temperatures. To first order, $T_{\mathrm{parcel}} \approx$ SST and $T_{\mathrm{ambient}} \approx T_c$, assuming that the boundary layer near the ocean surface is well mixed.



(a) Local ΔP vs. ΔSST correlation at 10°S, 243.75°E

(b) Local difference in slopes for $\Delta\text{SST} > 0\,\text{K}$ compared to $\Delta\text{SST} < 0\,\text{K}$

Figure 2. Plots illustrating the non-linear local relationship between ΔP and ΔSST.

The criterion for instability to convection is $B > 0$, which is therefore satisfied when the SST is above the threshold.

Fig. 2b also shows variation in the non-linearity within the tropics. More prominent threshold behaviour is observed in the western Pacific and Indian Ocean, compared to the eastern Pacific and Atlantic. This variation seems to agree with Graham and Barnett's [5] finding that surface wind divergence can suppress deep convection, even with SSTs above the critical threshold.

Having established that the underlying physical processes are non-linear, we now introduce two new approaches to try to capture these effects: a piecewise modification to the linear GTO and a simple CNN.

2.3.2 Piecewise linear model

A straightforward but physically motivated modification to the original model consists of defining separate GTOs for positive and negative SST perturbations. In particular, Eq. (2) is adjusted to read

$$\Delta P_j = \sum_{k|\Delta\text{SST}_k>0} G^{+}_{jk}\Delta\text{SST}_k A_k + \sum_{k|\Delta\text{SST}_k<0} G^{-}_{jk}\Delta\text{SST}_k A_k, \tag{6}$$

where G^{+}_{jk} and G^{-}_{jk} are the positive and negative GTOs, respectively. The first sum is taken over all the grid points for which $\Delta\text{SST}_k > 0$, and the second sum over all the grid points for which $\Delta\text{SST}_k < 0$. G^{+}_{jk} (and similarly G^{-}_{jk}) is calculated using Eq. (3), with one obvious modification: only the training examples for which $\Delta\text{SST}_k > 0$ are included in the vectors $\mathbf{\Delta SST_k}$ and $\mathbf{\Delta P_j}$. Validating the model results against the GCM, as well as performing the time series reconstruction, is done completely analogously to the linear case.

2.3.3 Convolutional neural network

We propose a CNN to model the relationship between ΔSST fields and ΔP_j at each grid point, using an architecture not dissimilar to the LeNet structure first proposed by LeCun et al. for the recognition of handwritten digits [10]. The CNN treats each ΔSST field as a two-dimensional image of size 97 × 192 and is able to build representations from geographically close regions of the input. This is an advantage compared to the linear model or a fully connected neural network, which have one-dimensional inputs that fail to encode geographical information and may be prone to overfitting. As the CNN architecture does not allow for a two-dimensional output, we train separate neural network parameters for each precipitation grid point j. In each case, the regression problem is defined as minimising a mean squared error loss function:

$$\arg\min_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i}^{n}(\Delta P_i - \widehat{\Delta P_i}(\boldsymbol{\theta}))^2, \tag{7}$$

where $n = 4435$ is the number of training examples, $\boldsymbol{\theta}$ are the neural network parameters, ΔP_i are the GCM precipitation anomalies (note that we have dropped the subscript j), and $\widehat{\Delta P_i}(\boldsymbol{\theta})$ are the CNN predictions.

A diagram showing the CNN architecture and the output dimensions of each layer is shown in Fig. 3. The network consists of three pairs of *convolutional layers* and *maximum pooling layers*, the output of which is flattened and fed into a fully connected neural network. In essence, each convolutional layer performs an image

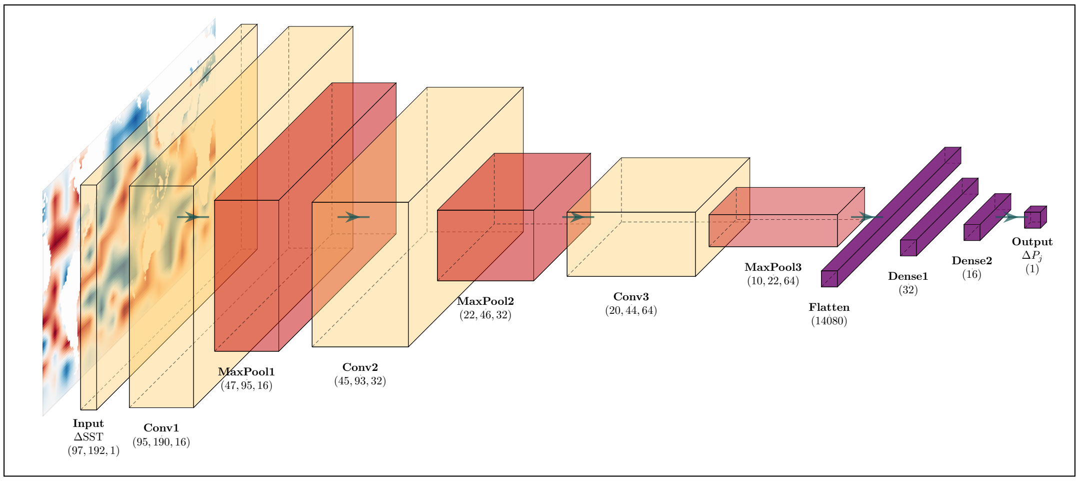


Figure 3. Diagram of the CNN architecture, showing the various layers and the output dimension of each layer. Note that while the input is two-dimensional, the third dimension of each convolutional layer is equal to the number of convolutional filters associated with that layer.

convolution of the previous layer. This is achieved using a number of convolutional filters or *kernels*, small square matrices with trainable weights that slide along the input matrix and perform inner products. An non-linear activation function is then applied to these inner products, forming the input to the next layer. As a concrete example, consider the action of the first convolutional layer, represented by the equation

$$y^1_{lm} = \sigma\left(\sum_{\alpha=0}^{s-1}\sum_{\beta=0}^{s-1}\theta_{\alpha\beta}\,y^0_{(l+\alpha)(m+\beta)}\right). \tag{8}$$

Here, y^1_{lm} is the layer output, y^0_{lm} is the layer input (the ΔSST grid), $\theta_{\alpha\beta}$ are trained parameters associated with the filter, s is the size of the filter, and σ is the non-linear activation function. Note that this equation only represents the action of a single filter in the convolutional layer. The number of filters is a hyperparameter, leading to a depth dimension equal to the number of filters (16 in this case). Next, each maximum pooling layer sub-samples its input by storing the maximum value of small square clusters. The main purpose of these subsampling layers is to reduce the dimensionality of the data and to prevent overfitting [10].

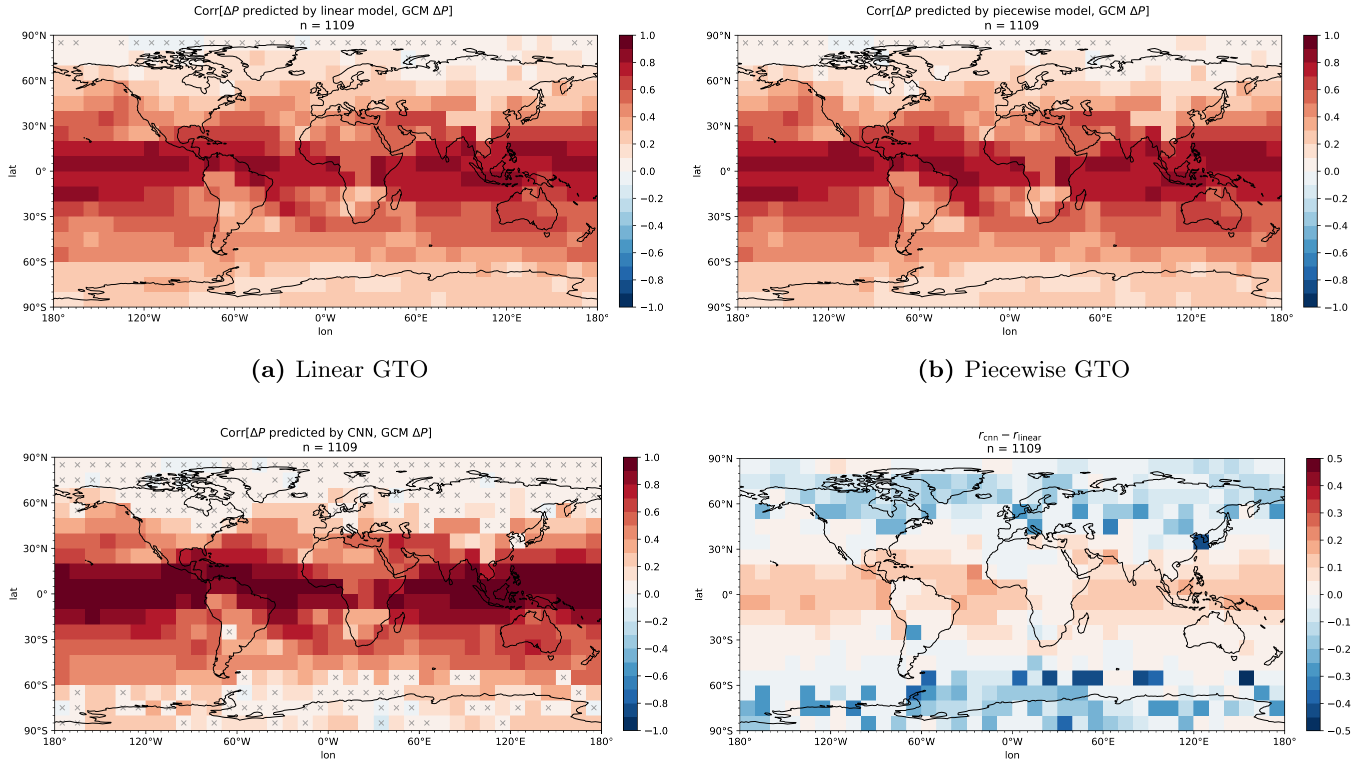
From a practical perspective, the implementation of the neural network code and its training algorithm is carried out in `Python` using the popular `TensorFlow` library developed by Google. No scaling is applied to the inputs, as the distribution of ΔSST has zero mean and close to unit variance. Training and cross-validation are performed for each precipitation grid point on the 4435 training examples, and the parameters $\boldsymbol{\theta}$ that give the lowest cross-validation loss are retained. There are naturally many hyperparameters to tune, such as the number of layers, the size and number of filters in each layer, the learning rate of each gradient descent step, the activation function, and so on. Given the large hyperparameter space and limited resources available, this process is largely carried out by trial and error. We find that sensible choices for the various hyperparameters are given by the values in Table 1 of Appendix A. Once the neural networks have been trained, validation against the GCM and time series reconstruction is done analogously to the previous models.

3 Results and discussion

3.1 Validation accuracy

We first consider the performance of the models on the validation set of 1109 mappings between ΔSST and ΔP. The Spearman's rank correlation coefficient r between ΔP_j generated by the GCM and ΔP_j predicted by the model can be used to quantify the validation accuracy at each grid point j. Figs. 4a to 4c show geographical plots of the correlation coefficient for the linear, piecewise, and CNN models respectively.

The general pattern of validation accuracy is similar for all three models and is statistically significant ($r > r_c = 0.059$) for almost all grid points except for selected regions near the poles. Meridionally, the highest validation accuracy is found in the equatorial regions and the lowest in the polar regions. This is perhaps



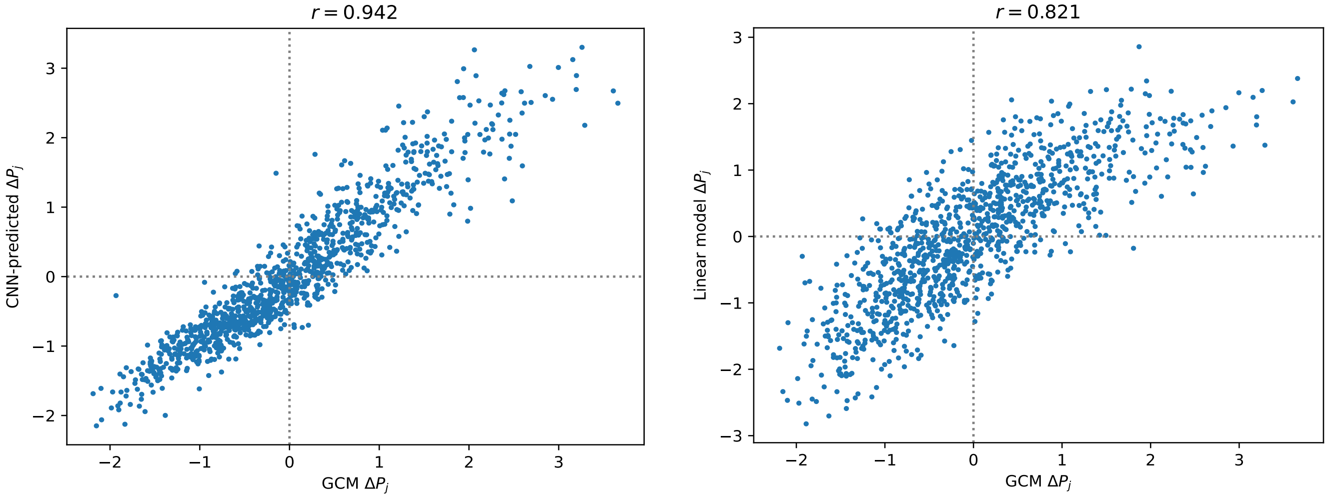
(a) Linear GTO

(b) Piecewise GTO

(c) CNN

(d) Difference between CNN (c) and linear GTO (a)

Figure 4. Plots of validation accuracy showing the correlation coefficient at each grid point between the GCM-generated ΔP_j and the model predicted ΔP_j. The validation is performed on the $n = 1109$ ΔSST to ΔP mappings not used for training. The critical correlation coefficient by a two-tailed t test is given by $r_c = 0.059$ assuming $n - 2 = 1107$ degrees of freedom and a significance level of $\alpha = 0.05$. In plots (a) to (c) showing the performance of the three models, a grey cross indicates that the correlation coefficient is not statistically significant at that grid point.



(a) CNN

(b) Linear GTO

Figure 5. Scatter plots showing the correlation between ΔP_j predicted by the CNN and linear model against the GCM-generated ΔP_j. This is for a representative tropical grid point centred at 5°N, 175°E. Note that while precipitation is in units mm day^{-1}, we have scaled the precipitation anomalies to zero mean and unit variance so the two plots can be easily compared.

unsurprising, as the extent of the ΔSST fields used in training the models is between 60°N and 60°S (refer to § 2.1 which describes the training data). In the polar regions the models are bound to perform poorly as they cannot rely on local SSTs, which are strongly correlated with local precipitation. In addition, given that the grid point size is proportional to cosine of the latitude, grid points in the polar regions are smaller in area and therefore are subject to more noise.

Zonally, the validation accuracy at the same latitude tends to be lower for grid points over land than for grid points over the ocean. For example, the validation accuracy appears unusually poor in southernmost regions of Africa and South America, as well as Central Asia, compared to other grid points at the same latitudes. This is again unsurprising as local SSTs are not defined over land, so the models are relying on remote teleconnections which are typically not as strongly correlated.

More interestingly, the CNN appears to outperform the linear and piecewise models between latitudes 30° N and 30° S. Fig. 4d shows the difference in correlation coefficients between the CNN and the linear model. In the tropics, the CNN performs better than the linear and piecewise models by about $\Delta r = 0.2$, with the CNN validation accuracy often exceeding $r = 0.9$ in these areas. We can examine the origin of this difference by considering a typical grid point in the tropics centred at 5°N, 175°E. The correlation between ΔP_j predicted by the two models compared to ΔP_j generated by the GCM is shown in Fig. 5. Not only is the correlation coefficient larger by about $\Delta r = 0.12$ for the CNN, the relationship between the CNN-predicted ΔP_j and the GCM-generated ΔP_j is considerably more linear. In the case of the linear GTO, we see that the relationship is somewhat curvilinear, suggesting a systematic error that we attribute to the non-linear ΔSST to ΔP relationship. On this evidence, it appears that the CNN is more effective at learning the threshold behaviour discussed in Graham and Barnett [5], which is particularly present in the tropics (recall Fig. 2b of the non-linearity distribution). These results also appear to support the idea that CNNs have the potential to outperform traditional approaches in the prediction of, say, ENSO events, echoing the findings of Ham et al. [6].

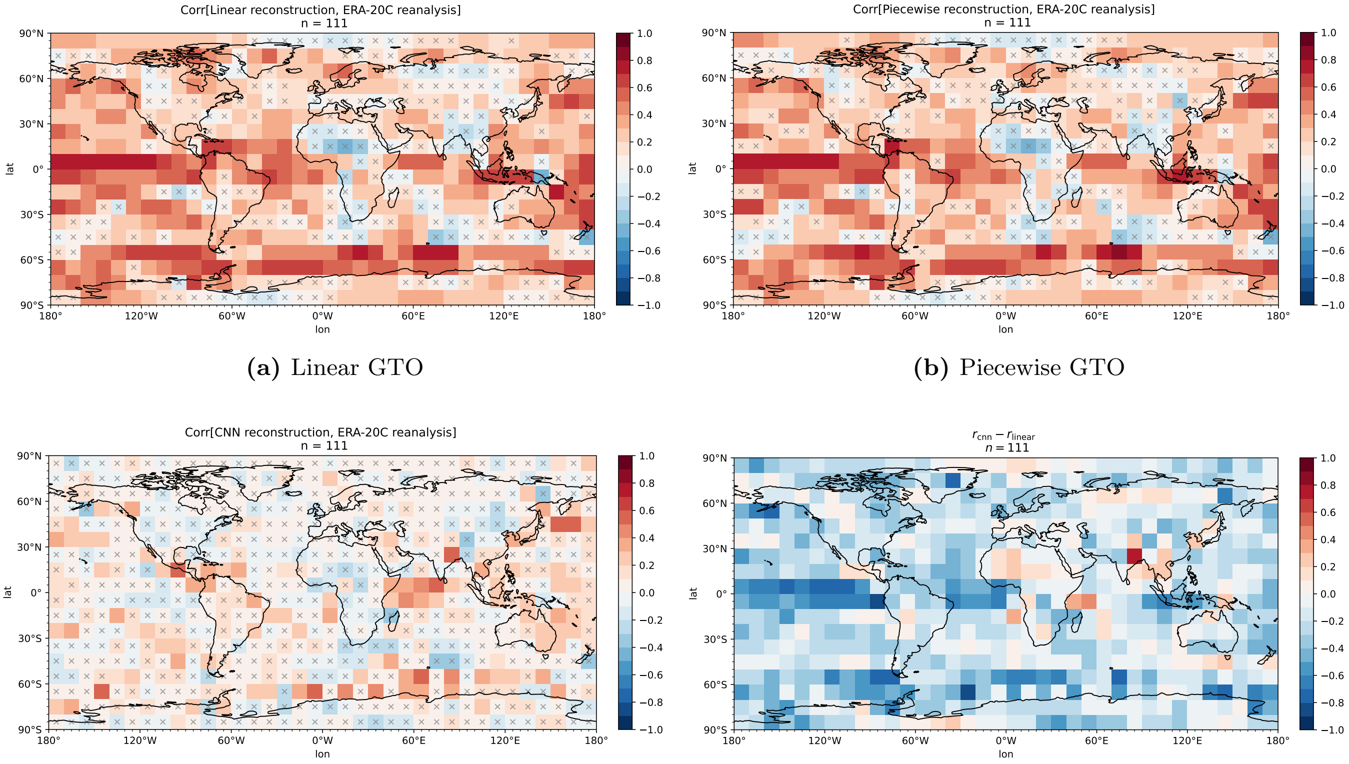
In contrast, in the polar regions and particularly for grid points north of 60°N and south of 60°S, the CNN does poorly compared to the simpler models by about $\Delta r = 0.3$. In these areas, the CNN parameters often fail to converge due to the amount of noise present and the lack of local SSTs, leading to poor predictions that are statistically insignificant. Another neural network architecture or different hyperparameters may be needed in these cases.

Quite surprisingly, Fig. 4 also reveals that the piecewise model not perform demonstrably better than the linear model in validation. There are statistically insignificant differences of order $\Delta r = 0.01$ for individual grid points, but the overall validation accuracy for the two models remains remarkably similar. We hypothesise that this could be due to the choice of $\Delta\mathrm{SST} = 0$ as the breakpoint for the piecewise function. As mean sea surface temperatures, and therefore T_{parcel} in Eq. (5), vary by grid point, it may be more accurate to define an arbitrary ΔSST breakpoint between the two GTO regimes, since $\Delta\mathrm{SST} = 0$ does not necessarily correspond to the SST threshold. More testing is required, but naively we would expect this approach to perform better than the linear GTO.

3.2 Time series reconstruction skill

We now consider the performance of the models in reconstructing mean annual precipitation between 1900 and 2010, inclusive. As described in § 2.2, reconstruction skill at each grid point j is measured using the correlation coefficient between the reconstructed precipitation anomaly time series $\Delta P_{j,\mathrm{reconstructed}}$ and the annual precipitation time series $P_{\mathrm{reanalysis}}$ from the ERA-20C reanalysis. Geographical plots of this correlation coefficient for the linear, piecewise, and CNN models respectively are shown in Figs. 6a to 6c. Here, statistical significance is defined as $r > r_c = 0.187$, again using a two-tailed t test at the 5% significance level. For the linear GTO, we include examples of reconstructions for selected grid points in Fig. 7 of Appendix B.

As in the validation case, both the pattern of reconstruction skill and the average performance of the linear and piecewise models are rather similar. Strong reconstruction skill characterised by $r > 0.5$ is observed in equatorial grid points in the eastern Pacific, much of the Southern Ocean, and selected areas in the mid-Atlantic and northern Indian Ocean. This agrees with the results of Tsai et al. [4] that skilful reconstructions of regional precipitation are possible with the purely linear GTO. Poor reconstruction skill characterised by $r < r_c$ is observed in Central Africa, parts of Central Asia and Russia, and some areas of the southern Indian Ocean. In some cases *negative* skill is observed.


Corr[Linear reconstruction, ERA-20C reanalysis]
n = 111
Corr[Piecewise reconstruction, ERA-20C reanalysis]
n = 111
Corr[CNN reconstruction, ERA-20C reanalysis]
n = 111
$r_{\rm cnn} - r_{\rm linear}$
$n = 111$
90°N
60°N
30°N
0°
30°S
60°S
90°S
180°
120°W
60°W
0°W
60°E
120°E
180°
lat
lon
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0

(a) Linear GTO

(b) Piecewise GTO

(c) CNN

(d) Difference between CNN (c) and linear GTO (a)

Figure 6. Plots of reconstruction skill showing the correlation coefficient at each grid point between the annual reconstructed precipitation anomalies between 1900 and 2010 and the ERA-20C reanalysis. There are $n = 111$ time steps. The critical correlation coefficient by a two-tailed t test is given by $r_c = 0.187$ assuming $n - 2 = 109$ degrees of freedom and a significance level of $\alpha = 0.05$. In plots (a) to (c) showing the performance of the three models, a grey cross indicates that the correlation coefficient is not statistically significant at that grid point.

This could be due to noise, as the identified regions with low skill roughly correlate well with regions that receive minimal precipitation in general based on the ERA-20C data [9], potentially lowering the signal-to-noise ratio in these areas. Negative skill could also indicate a systematic error arising from the lack of atmosphere-ocean coupling in the models (see § 3.3). As with validation accuracy, the reconstruction skill is usually higher for oceanic grid points compared to terrestrial grid points due to the influence of local SSTs. However, in contrast to validation accuracy, the distribution of reconstruction skill across latitudes is significantly more random.

Fig. 6c illustrates the rather disappointing result that the CNN does not appear to be skilled at performing reconstructions, with very few grid points satisfying $r > r_c$. This is especially apparent in Fig. 6d, which shows that the linear model outperforms CNN for most grid points.

We suspect that the CNN's poor reconstruction skill stems from the type of input data on which it is being trained. Recall that the input data consist of artificially generated ΔSST fields, with the values of the perturbation anomalies randomly drawn from a uniform distribution between $-2\,$K and $2\,$K. Unlike the linear and piecewise models, the CNN learns spatial relationships between input features. However, the SST fields on which the reconstructions are based take the form $\Delta\text{SST}_n = \text{SST}_n - \overline{\text{SST}}_{1980-99}$ where n runs from 1900 to 2010. Due to the increase in mean SSTs over the past century due to climate change, these SST anomaly fields tend to be mostly negative in the earlier years and mostly positive in the later years, and importantly, different from the input fields learned by the CNN. Thus, while the CNN is strong at emulating the GCM when making predictions on ΔSST fields similar to those used for training, it is much less robust than the Green's function-based methods for reconstructions, as the spatial correlations in real ΔSST fields are distinct from those found in artificially generated training examples. In contrast, Ham et al. [6] trained their CNN on historical simulations and reanalysis data from 1871 to 1973, in order to make ENSO-related forecasts during the validation period from 1984 to 2017. This proved to be a more fruitful approach and, combined with our findings, suggests that the random perturbation method may be less effective for training robust deep learning models.

3.3 Limitations

There are a number of other limitations to the study that could have affected both the validation accuracy and reconstruction skill of the models, which we now discuss. As identified by Baker et al. [3], we have ignored the effect of atmosphere-ocean coupling, as the ocean cannot respond instantaneously to atmospheric processes induced by a change in SSTs. Barsugli and Battisti [11] and Bretherton and Battisti [12] suggest that this effect is especially prominent in the mid-latitudes, limiting the predictability of using SST anomalies to simulate atmospheric variables over long time scales such as those used in our reconstructions. This effect could explain the regions of negative skill observed in Figs. 6a and 6b. Nevertheless, the high reconstruction skill in many regions can be explained by the fact that regional precipitation often contains a significant forced component due to remote teleconnections (as illustrated in Fig. 1), which are less affected by the lack of atmosphere-ocean coupling compared to local ΔSST to ΔP relationships.

There are also some limitations specific to the CNN. As discussed in § 2.3.3, we define separate regression problems for each precipitation grid point, using the ΔSST fields as the input and ΔP_j at each grid point as the scalar output. However, due to resource and time constraints, we train CNNs with identical hyperparameters for each regression problem. There are two main problems associated with this. First, the choice of architecture and hyperparameters may not be appropriate for all grid points, as our analysis of Fig. 4c suggested. Second, having a scalar output rather than an image output means that the CNN is unable to learn spatial correlations between *precipitation* grid points. Indeed, the GCM-generated ΔP fields indicate that anomalously high precipitation in a tropical area typically correlates with anomalously low precipitation in a neighbouring subtropical area, in broad agreement with the simple Hadley cell picture of atmospheric circulation. (Example plots illustrating the spatial correlation of the ΔP fields are shown in Fig. 8 of Appendix C.) Incorporating information about precipitation correlations could feasibly increase the validation accuracy of the model.

Another limitation is the use of the ERA-20C data as a 'ground truth' against which our models' reconstructions are compared. The reanalysis does not directly assimilate precipitation observations [9] and could therefore suffer from its own inherent biases. Finally, we note that the HadAM3P GCM itself has a relatively low vertical resolution in the stratosphere [3]. It is suggested that this could exclude some teleconnections that propagate in the stratosphere, leading to inaccurate estimates of the true atmospheric response to SST anomalies. Further work using a higher-resolution model could be profitable.

4 Conclusions

Building on previous work by Baker et al. [3], this study evaluated a linear Green's function model, a piecewise Green's function model, and a CNN in the context of predicting the response of regional precipitation to regional SST anomalies. We used two key performance indicators to evaluate the models: accuracy in validating the ΔSST to ΔP fields generated by GCM random perturbation simulations, and reconstruction skill in hindcasting yearly precipitation from historical SST data. In validation, all three models demonstrated skill in emulating the GCM, with higher accuracy in the tropics and lower accuracy in the polar regions. However, the CNN outperformed the traditional Green's function approaches in the tropics. We showed that this was due to its ability to learn the non-linear relationship between ΔSST and ΔP, which is likely driven by the temperature threshold for deep convection [5]. While less physically informative, these results suggest that machine learning-based models may be more effective emulators of GCMs than Green's function approaches.

In reconstruction, the two Green's function models performed similarly skilful precipitation hindcasts in the eastern Pacific, Southern Ocean, and mid-Atlantic, and less skilful hindcasts in regions such as Central Africa and Asia. This confirmed earlier results by Tsai et al. [4] showing some precipitation reconstruction skill using the linear GTO. Unfortunately, such skill was not observed for the CNN, which we attributed to the difference between the artificially generated ΔSST fields used for training and the historical

ΔSST fields. This seems to indicate that deep learning models trained on randomly generated SST fields may be less robust for reconstruction problems, provided that the models are trained on artificially generated data.

Several limitations to the study were identified, including the assumption of no atmosphere-ocean coupling, the fixed CNN architecture, the use of a reanalysis dataset as a benchmark, and the low resolution of the GCM itself. In view of our results and these limitations, we suggest a few possibilities for further research:

- The piecewise modification to the linear model could be adjusted to include a variable ΔSST breakpoint delineating the two piecewise regimes. This appears to be more consistent with the idea that the SST threshold does not necessarily occur at $\Delta\text{SST} = 0$ for all grid points, as we assumed in our study for simplicity.

- With a view to improving reconstruction skill using the CNN, the possibility of training a CNN using historical SST anomalies rather than random SST fields can be explored. This would be more consistent with existing approaches in the literature [6] and could lead to more accurate hindcasts (and hopefully forecasts).

- It is worth experimenting with alternate neural network architectures that capture other kinds of spatial correlation. One possibility is the U-Net architecture first proposed by Ronneberger et al. [13] for biomedical image segmentation. This is essentially a CNN consisting of both downsampling *and* upsampling layers, therefore allowing for image outputs. Such a model could prove effective in learning spatial relationships between precipitation anomalies. Another possibility is a graph neural network, which could more effectively capture distant spatial correlations in the input data and provide greater physical interpretability [14].

- Testing reconstructions against an purely observational precipitation dataset, such as the Global Precipitation Climatology Project (GPCP) [15], could help to check for any biases associated with the reanalysis data.

- Finally, more work is required to study the physical mechanisms (e.g. Rossby wave propagation) by which regional SSTs impact regional precipitation. This would further our physical understanding of calculated teleconnections such as those illustrated in Fig. 1.

References

[1] J. J. Barsugli and P. D. Sardeshmukh, “Global atmospheric sensitivity to tropical SST anomalies throughout the Indo-Pacific basin”, Journal of Climate **15**, 3427–3442 (2002).

[2] W. Li and C. E. Forest, “Estimating the sensitivity of the atmospheric teleconnection patterns to SST anomalies using a linear statistical method”, Journal of Climate **27**, 9065–9081 (2014).

[3] H. Baker et al., “The linear sensitivity of the North Atlantic Oscillation and eddy-driven jet to SSTs”, Journal of Climate **32**, 6491–6511 (2019).

[4] C.-Y. Tsai et al., “Estimating the regional climate responses over river basins to changes in tropical sea surface temperature patterns”, Climate Dynamics **45**, 1965–1982 (2015).

[5] N. E. Graham and T. P. Barnett, “Sea surface temperature, surface wind divergence, and convection over tropical oceans”, Science **238**, 657–659 (1987).

[6] Y.-G. Ham et al., “Deep learning for multi-year ENSO forecasts”, Nature **573**, 568–572 (2019).

[7] J. W. Hurrell et al., “A new sea surface temperature and sea ice boundary dataset for the community atmosphere model”, Journal of Climate **21**, 5145–5153 (2008).

[8] N. A. Rayner et al., “Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century”, Journal of Geophysical Research: Atmospheres **108** (2003).

[9] P. Poli et al., “ERA-20C: An atmospheric reanalysis of the twentieth century”, Journal of Climate **29**, 4083–4097 (2016).

[10] Y. LeCun et al., “Gradient-based learning applied to document recognition”, Proceedings of the IEEE **86**, 2278–2324 (1998).

[11] J. J. Barsugli and D. S. Battisti, “The basic effects of atmosphere–ocean thermal coupling on midlatitude variability”, Journal of the Atmospheric Sciences **55**, 477–493 (1998).

[12] C. S. Bretherton and D. S. Battisti, “An interpretation of the results from atmospheric general circulation models forced by the time history of the observed sea surface temperature distribution”, Geophysical Research Letters **27**, 767–770 (2000).

[13] O. Ronneberger et al., “U-Net: convolutional networks for biomedical image segmentation”, `10.48550/ARXIV.1505.04597` (2015).

[14] S. R. Cachay et al., “The world as a graph: improving El Niño forecasts with graph neural networks”, `10.48550/ARXIV.2104.05089` (2021).

[15] R. F. Adler et al., “The Global Precipitation Climatology Project (GPCP) monthly analysis (new version 2.3) and a review of 2017 global precipitation”, Atmosphere **9**, `10.3390/atmos9040138` (2018).

[16] D. P. Kingma and J. Ba, “Adam: a method for stochastic optimization”, `10.48550/ARXIV.1412.6980` (2014).

[17] N. Srivastava et al., “Dropout: a simple way to prevent neural networks from overfitting”, Journal of Machine Learning Research **15**, 1929–1958 (2014).

Appendices

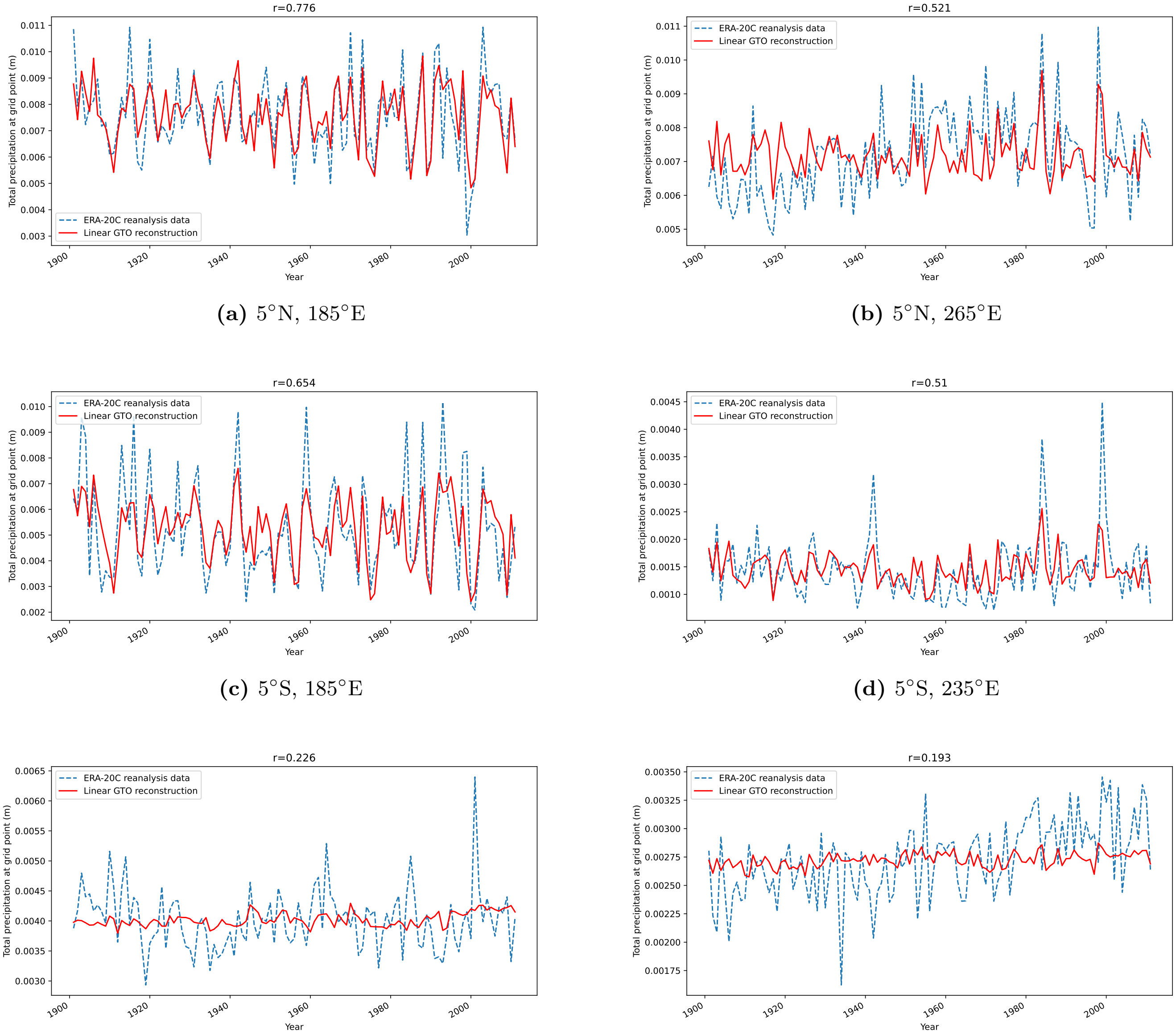
A CNN hyperparameters

Hyperparameter	**Value/Description**
Activation function	ReLU*
Learning rate	10^{-3}
Batch size	4
Optimiser	Adam*
No. of convolutional layers	3
No. of pooling layers	3
Convolutional filter size	3×3
Pooling filter size	2×2
No. of convolutional filters in each layer	16, 32, 64
Other regularisation	10% dropout* after Dense1 layer

Table 1. Chosen CNN hyperparameters. The ReLU (rectified linear unit) is commonly used activation function defined as $f(x) = \max(0, x)$. The Adam optimiser is an efficient algorithm for stochastic gradient descent [16]. Dropout regularisation is a technique shown to prevent overfitting in supervised learning tasks, which consists of ignoring a random percentage of layer outputs during training [17].

B Example reconstructions using the linear GTO

The plots in Fig. 7 show example reconstructions of annual precipitation from 1900 to 2010 for selected grid points. Recall from § 2.2 that the reconstructed time series $\Delta P_{j,\text{reconstructed}}$ at each grid point j is regressed onto the ERA-20C reanalysis time series using the equation $P_{j,\text{reanalysis}} = \alpha_j \Delta P_{j,\text{reconstructed}} + \beta_j$, where we determine the regression coefficients α_j and β_j. A time series of the *total* reconstructed precipitation can then be computed using these regression coefficients, using $P_{j,\text{reconstructed}} = \alpha_j \Delta P_{j,\text{reconstructed}} + \beta_j$. This conveniently allows us to plot $P_{j,\text{reconstructed}}$ and $P_{j,\text{reanalysis}}$ on the same set of axes.



(a) 5°N, 185°E

(b) 5°N, 265°E

(c) 5°S, 185°E

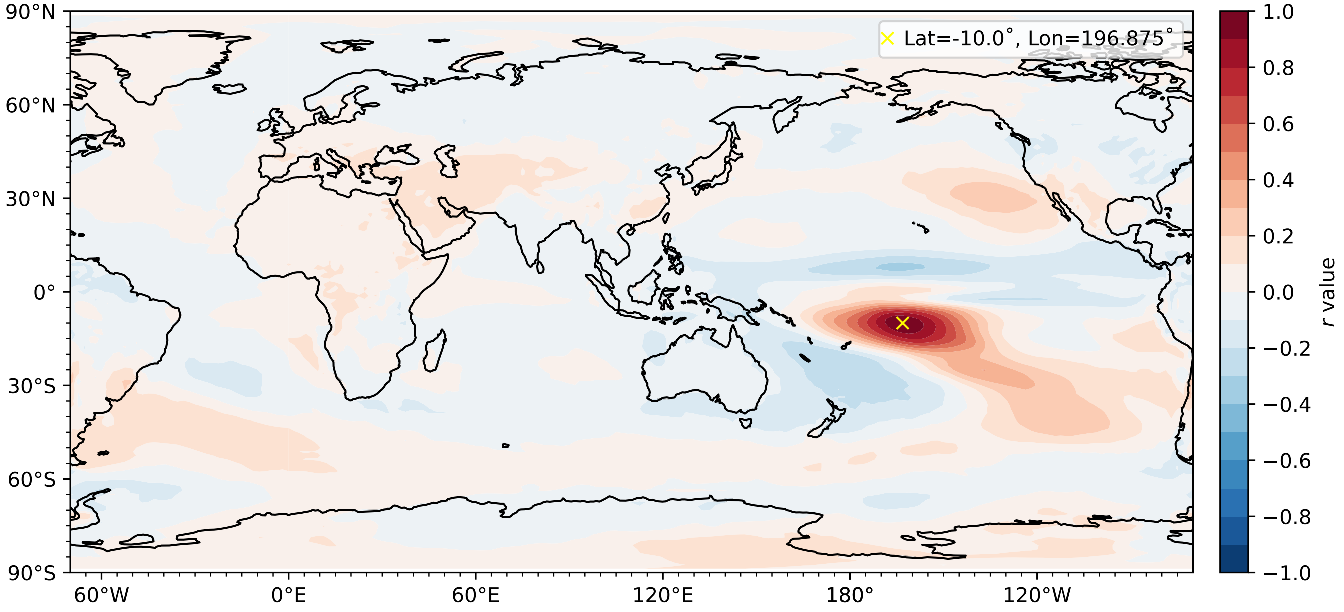
(d) 5°S, 235°E

(e) 35°N, 185°E

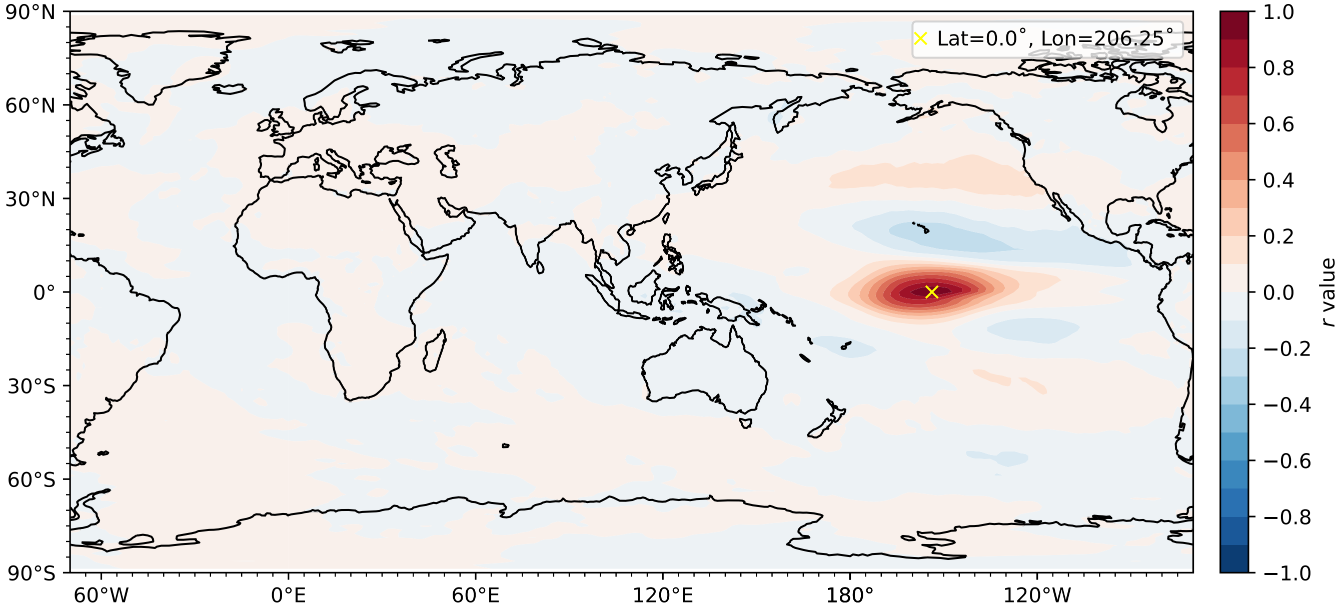
(f) 55°N, 355°E

Figure 7. Examples of reconstructions at various grid points, showing the the reconstructed time series $P_{j,\text{reconstructed}}$ and reanalysis time series $P_{j,\text{reanalysis}}$ on the same axes. The grid points were chosen to illustrate varying levels of reconstruction skill.

C Example plots of spatial correlation between precipitation anomalies



(a) 10°N, 196.875°E



(b) 0°N, 206.25°E

Figure 8. Plots showing the Spearman's rank correlation coefficient between ΔP_j at a selected grid point (indicated by the yellow cross) and ΔP_j at all the other grid points. We observe that anomalously high precipitation in a tropical area is often correlated with anomalously low precipitation in a neighbouring subtropical area, as the alternating red and blue regions indicate. This observation is broadly consistent with the Hadley cell model of atmospheric circulation, characterised by rising air near the equator and descending air in the subtropics. (Note that the precipitation grid used in these plots has *not* been coarsened.)