Interpretable Deep Learning for Probabilistic MJO Prediction

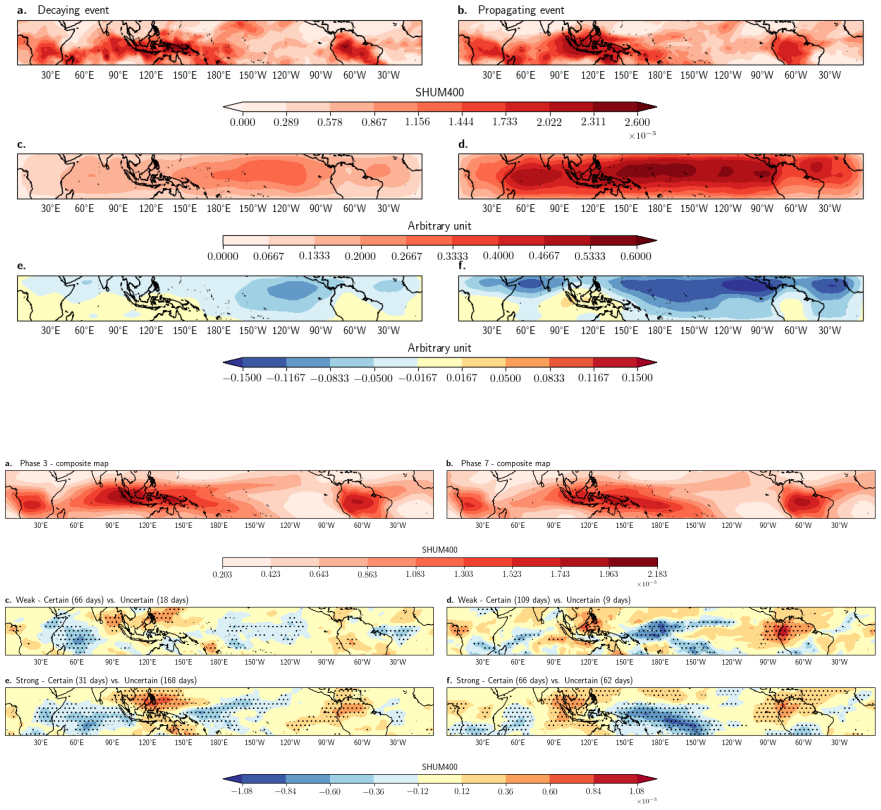
Antoine Delaunay[1,1] and Hannah Christensen[2,2]

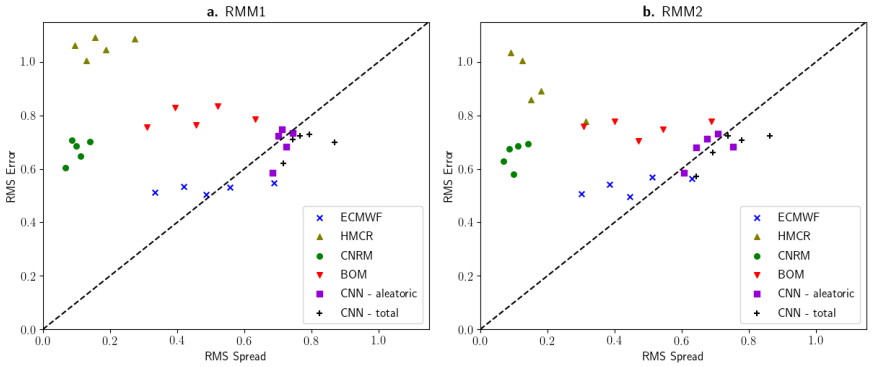
[1]Ecole Polytechnique
[2]University of Oxford

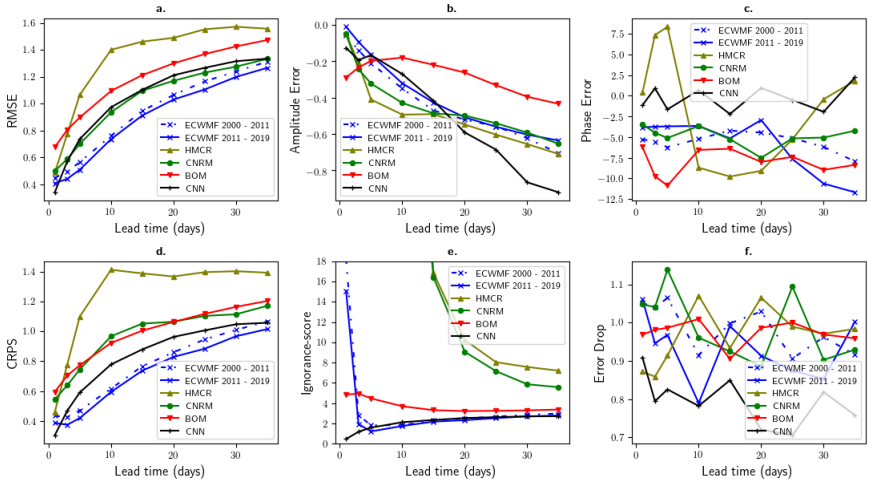
November 30, 2022

Abstract

The Madden–Julian Oscillation (MJO) is the dominant source of sub-seasonal variability in the tropics. It consists of an Eastward moving region of enhanced convection coupled to changes in zonal winds. It is not possible to predict the precise evolution of the MJO, so sub-seasonal forecasts are generally probabilistic. We present a deep convolutional neural network (CNN) that produces skilful state-dependent probabilistic MJO forecasts. Importantly, the CNN's forecast uncertainty varies depending on the instantaneous predictability of the MJO. The CNN accounts for intrinsic chaotic uncertainty by predicting the standard deviation about the mean, and model uncertainty using Monte-Carlo dropout. Interpretation of the CNN mean forecasts highlights known MJO mechanisms, providing confidence in the model. Interpretation of forecast uncertainty indicates mechanisms governing MJO predictability. In particular, we find an initially stronger MJO signal is associated with more uncertainty, and that MJO predictability is affected by the state of the Walker Circulation.







Interpretable Deep Learning for Probabilistic MJO Prediction

Antoine Delaunay[1] **and Hannah M. Christensen**[2]

[1]Department of Applied Mathematics, Ecole Polytechnique, Palaiseau, France

[2]Department of Physics, University of Oxford, Oxford, UK

Key Points:

- A deep convolutional neural network (CNN) is used to produce probabilistic forecasts of the MJO
- The forecasts provide well-calibrated state-dependent estimates of forecast uncertainty
- The CNN forecasts are used to probe sources of predictability for the MJO

Corresponding author: Hannah M. Christensen, hannah.christensen@physics.ox.ac.uk

Abstract

The Madden–Julian Oscillation (MJO) is the dominant source of sub-seasonal variability in the tropics. It consists of an Eastward moving region of enhanced convection coupled to changes in zonal winds. It is not possible to predict the precise evolution of the MJO, so sub-seasonal forecasts are generally probabilistic. We present a deep convolutional neural network (CNN) that produces skilful state-dependent probabilistic MJO forecasts. Importantly, the CNN's forecast uncertainty varies depending on the instantaneous predictability of the MJO. The CNN accounts for intrinsic chaotic uncertainty by predicting the standard deviation about the mean, and model uncertainty using Monte-Carlo dropout. Interpretation of the CNN mean forecasts highlights known MJO mechanisms, providing confidence in the model. Interpretation of forecast uncertainty indicates mechanisms governing MJO predictability. In particular, we find an initially stronger MJO signal is associated with more uncertainty, and that MJO predictability is affected by the state of the Walker Circulation.

Plain Language Summary

The Madden-Julian Oscillation (MJO) is an important tropical climate phenomenon. It consists of enhanced convective thunderstorms and anomalous winds that propagate eastward along the Equator for a few weeks. The MJO is difficult to predict and exhibits great variability. This means that forecasts are often probabilistic. However, current models have difficulty in correctly predicting the uncertainty in the forecast based on the current conditions. In this paper, we propose a model using neural networks capable of making reliable probabilistic forecasts. We interpret the behaviour of the algorithm to verify its consistency with the known physical mechanisms of the MJO and to highlight new physical conditions that affect MJO prediction uncertainty.

1 Introduction

The Madden-Julian Oscillation (MJO: Madden & Julian, 1971) is an envelope of enhanced tropical convection with associated changes to the atmospheric circulation. It is characterised by its period of 40-50 days, its planetary scale, and its Eastward propagation at speeds of 4–8 ms^{-1}. It is the major source of predictability on sub-seasonal timescales in the Tropics (Zhang, 2013) and influences phenomena such as the North Atlantic Oscillation and Arctic sea ice cover through global teleconnections (Ferranti et al.,

1990; Cassou, 2008; Yoo et al., 2012; Henderson et al., 2014). Subseasonal forecasts are of great socio-economic value through their potential to predict extreme weather events several weeks ahead (Vitart & Robertson, 2018). There is therefore great interest in improving predictions of the MJO, and in understanding sources of MJO predictability (Kim et al., 2018).

The chaotic nature of the Earth System means that it is not possible to predict the precise evolution of the MJO beyond a few days, so subseasonal forecasts are generally probabilistic (J. Slingo & Palmer, 2011; Bauer et al., 2015). If the probabilistic forecast *mean* is assessed, averaging out the unpredictable 'noise', current dynamical models have a prediction skill up to three weeks (Lim et al., 2018; Vitart, 2017). However, systematic biases remain, especially in the propagation of the MJO convective anomaly over the Maritime Continent (Kim et al., 2016; Barrett et al., 2021; Li et al., 2020). In contrast to the mean skill, the *probabilistic* skill of MJO forecasts is low (Lim et al., 2018; Vitart, 2017). Improving probabilistic forecasts is essential to quantify our confidence in the predictions, and to advance understanding of the predictability of this phenomenon.

While prediction skill is a property of the forecast model, predictability is a property of the Earth-system. MJO predictability studies have focused on the theoretically achievable prediction limit that one could achieve with a perfect model, quantified as 6–7 weeks (e.g. Neena et al., 2014; Wu et al., 2016; Kim et al., 2018). This is complementary to an approach taken in the medium-range forecasting community, where 'predictable' forecasts are those for which the forecast uncertainty is small (e.g. Palmer, 2000). This identification is possible because medium-range forecasts exhibit state-dependent reliability (Leutbecher & Palmer, 2008). If reliable, state-dependent, MJO forecasts could be produced, forecast uncertainty could be used as an indicator of instantaneous MJO predictability.

Increasing volumes of data, advances in computational power, and developments in statistical modelling have led to substantial interest in the use of machine learning in Earth-system science (Reichstein et al., 2019; Huntingford et al., 2019). Deep learning has been applied to the MJO for phase classification (Toms et al., 2020; Martin et al., 2021), post processing (Kim et al., 2021), and deterministic prediction (Martin et al., 2021). Here, we develop a neural network that produces well calibrated probabilistic forecasts of the MJO. We use a convolutional neural network (CNN), which has proved ef-

fective at identifying hidden patterns and processes in climate (Ham et al., 2019; Arcomano et al., 2020; Schultz et al., 2021) and other areas such as image recognition (Russakovsky et al., 2015).

The paper is structured as follows: in Section 2, we describe the CNN, including the data used to train the model. In Section 3 we present our results. We evaluate the CNN compared to dynamical models from the Subseasonal-to-Seasonal (S2S) prediction project. We validate the CNN by seeking to understand its mean forecasts, before using the CNN to uncover potential sources of predictability for the MJO. Finally we discuss the significance of our results and draw conclusions in Section 4.

2 Methods

2.1 Data

Observational data used to train and test the CNN are taken from the ECMWF Reanalysis version 5 (ERA5) dataset between 1979–2019 (Hersbach, H., et al., 2020). We compare the CNN to models from the S2S database (F. Vitart et al., 2017). We select reforecast data from four representative models, chosen to span the range of performances of models in the S2S database. In particular, we include the European Centre for Medium-Range Weather Forecasts (ECMWF) model, which is known to produce the most skilful MJO forecasts (Lim et al., 2018). The remaining models chosen had the largest reforecast ensemble size, enabling probabilistic forecast skill to be assessed. Details are presented in Supporting Table S1 and Text S1.

2.2 Overview of Predictive Model

The MJO is a coupled convective-dynamic anomaly that can be summarised by the bivariate Real-time Multivariate MJO (RMM) index (Wheeler & Hendon, 2004). The RMM index classifies active MJO events (amplitude greater than one) into one of eight phases depending on geographical location (e.g. Supporting Figure S1). Using observed daily-mean maps for a single date t as inputs, we train a deep CNN to predict the mean and uncertainty in RMM1 and RMM2 computed from daily means at a later date $t+\tau$, training a separate CNN for each lead time. The chosen lead times are one, three and five days, then every fifth day up to 35 days. The architecture of the CNN is shown in Supporting Figure S2.

We compute the observed values of the RMM following Wheeler and Hendon (2004) (Supporting Text S2). Subseasonal anomalies of daily-mean Outgoing Longwave Radiation (OLR) and daily-mean zonal winds at 200 hPa (UA200) and 850 hPa (UA850) between 20°S–20°N are latitudinally averaged and divided by their global variance. The first two Empirical Orthogonal Functions (EOFs) of the combined fields are computed. RMM1 and RMM2 are the projection of the daily fields onto EOFs 1 and 2.

Even though the MJO shows seasonal behaviour, we train a single model for all seasons to maximise the available training data. As inputs we use subseasonal anomalies of OLR, UA200, and UA850, consistent with fields used to compute the RMM indices. We supplement these with four further fields which provide complementary information: daily mean Specific Humidity at 400 hPa (SHUM400) was included because Barrett et al. (2021) reported large differences in SHUM400 between MJO events which propagate and weaken over the Maritime Continent; daily mean geopotential at 850 hPa (Z850) provided skill in previous work (Toms et al., 2020); daily mean Downwelling Longwave Radiation at the surface (DLR) has a marked annual cycle, which we found a more effective means of accounting for the seasonality of the MJO than including a dummy variable. Finally, daily anomalies of sea surface temperature (SST) are included, since the MJO is known to be linked to El Nino-Southern Oscillation (ENSO: e.g. Kessler, 2001). Sensitivity of CNN performance to the choice of input feature is shown in Supporting Figure S3, providing insights into sources of predictability for the MJO. Inputs are provided as maps spanning 0–360^{o}E, 20^{o}S–20^{o}N on a 2.5^{o}x2.5^o grid. The different variables are input to the CNN as separate channels. This allows the CNN to learn to identify co-located phenomena. To ensure independence between the training and testing data sets, we use the first 80% of the dates for training, and the remaining 20% for testing.

We model the two forecast RMM indices as following a Gaussian Bivariate distribution with null correlation (Wheeler & Hendon, 2004). The network outputs the predicted means and variances of RMM1 and RMM2, and is trained by minimising the negative log-likelihood. The output variance represents the intrinsic chaotic (aleatoric) uncertainty in the prediction. In addition, we represent the epistemic uncertainty in the CNN model weights using a Monte-Carlo Dropout method to produce an ensemble of forecasts (Gal & Ghahramani, 2016; Gal, 2016; Scalia et al., 2019). The total forecast uncertainty is the sum of the aleatoric and epistemic variances. More details are provided in Supporting Text S3.

2.3 Interpretation using PatternNet

We use the PatternNet algorithm (Kindermans et al., 2017) to interpret forecasts made by the CNN, as it outperforms other approaches including Guided BackProp and Layerwise Relevance Propagation in both idealised test cases and for image classification problems (Kindermans et al., 2017). Inputs to the CNN include a *signal*, that contains information about the future state of the MJO, and a *distractor*, that is a residual containing information irrelevant to the prediction task (Kindermans et al., 2017). PatternNet is a distinct network to the CNN, but whose structure reflects that of the CNN in reverse, propagating the estimated signal from the output to the input space, thereby disentangling the signal from the distractor: for more details, see Supporting Text S4.

3 Results

3.1 Network performance

Figure 1 compares the network's performance to models from the S2S database (see Supporting Text S5 for definitions of all metrics). Figures 1(a–c) show the deterministic skill of the CNN mean forecasts in terms of the Root Mean Square Error (RMSE), Amplitude Error, and Phase Error respectively. In terms of RMSE, the CNN is competitive with models from the S2S database, though has larger errors than ECMWF. Similarly to the dynamical models, the CNN forecasts suffer from an increasing amplitude error with time, indicating a decay in MJO strength over the duration of the forecast. It is known that dynamical models simulate slower MJO propagation speeds than observed, resulting in a negative phase error (Lim et al., 2018). Here the CNN outperforms the dynamical models, accurately capturing the MJO propagation speed. A fourth metric, the bivariate correlation, is shown in Supporting Figure S4: the CNN performance is poorer than ECMWF, but similar to CNRM and BOM.

Figures 1(d–f) assess the probabilistic skill of the CNN. The Continuous Ranked Probability Score (CRPS: Marshall et al. (2016)) compares forecast and observed cumulative distribution functions. The CNN is competitive with forecasts from the S2S database, outperforming three of the four dynamical models considered. Despite being widely used, the CRPS can give unintuitive rankings (e.g Bolin & Wallin, 2019), as it penalises errors in the forecast mean more than poor calibration of spread (Christensen et al., 2015).

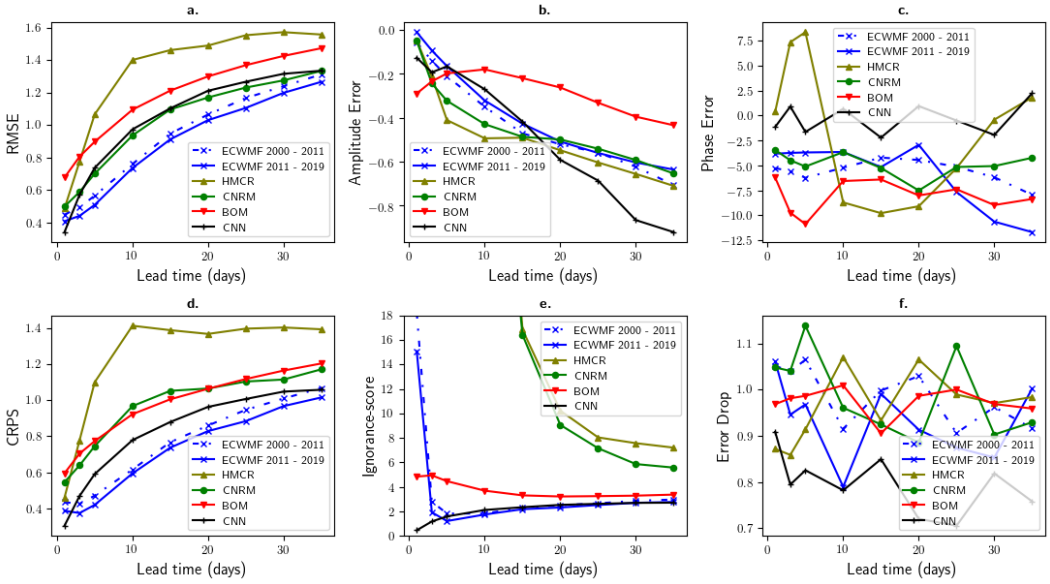


Figure 1. Skill of CNN (black), compared to forecasts from the subseasonal-to-seasonal prediction project (colours) as a function of lead time. (a) Root mean square error. (b) Amplitude error. (c) Phase error. (d) Continuous Ranked Probability Score. (e) Log-score. CNRM and HMCR scores before day-15 were too high to be shown. (f) Error-Drop. For all scores, a value closer to zero indicates a more skilful forecast. Forecasts from different models cover: ECMWF 2000-2019; HMCR 1985-2010; CNRM 1993-2017; BOM 1982-2013; CNN 2011-2019. The ECMWF data was split into two to allow direct comparison with the CNN over 2011-2019, and to give an indication of sampling uncertainty.

An alternative score is the 'Ignorance' or log-score (Roulston & Smith, 2002) (Panel e). This score is local, derived from information theory, and easily generalises to multivariate predictions (Roulston & Smith, 2002; Bjerregård et al., 2021). It is also consistent with the loss function used to train the network. According to the log-score, the CNN is one of the two models with the best forecast skill at lead times of 5–35 days. At shorter lead times, it outperforms all dynamical models. The poor performance of dynamical models at these short lead times is due to overconfident forecasts (Bjerregård et al., 2021), which are penalised by the log-score. In contrast, the CNN is able to balance the loss in accuracy with an increasing predicted uncertainty as the lead time increases.

For probabilistic forecasts to be useful, observations should behave as if they were drawn from the forecast probability distribution. For this to hold, a smaller forecast spread should indicate a smaller root mean squared error (RMSE) in the forecast mean on av-

erage. We assess this property using Error-Spread diagrams (Leutbecher & Palmer, 2008) shown in Figure 2. The RMSE is a measure of predictability of the atmosphere: high RMSE indicates lower predictability. The spread indicates the forecast model's belief about the predictability. For well calibrated forecasts, RMSE and spread should be correlated, and the observed RMSE should equal the predicted standard deviation, with scattered points lying on the one-to-one line. None of the dynamical models have this property: their error distributions are independent of the forecast spread, such that the spread gives no indication of the true predictability of the MJO on that day. In contrast, if the CNN forecast spread is low, the RMSE is smaller than if the spread is high. The probabilistic forecasts produced by the CNN are a dynamic indicator of the certainty in the MJO forecasts, and therefore the instantaneous predictability of the MJO. The aleatoric uncertainty predicted by the CNN is substantially greater than the epistemic uncertainty, indicating that while the MJO exhibits chaotic unpredictability, the CNN weights are well constrained by the available data.

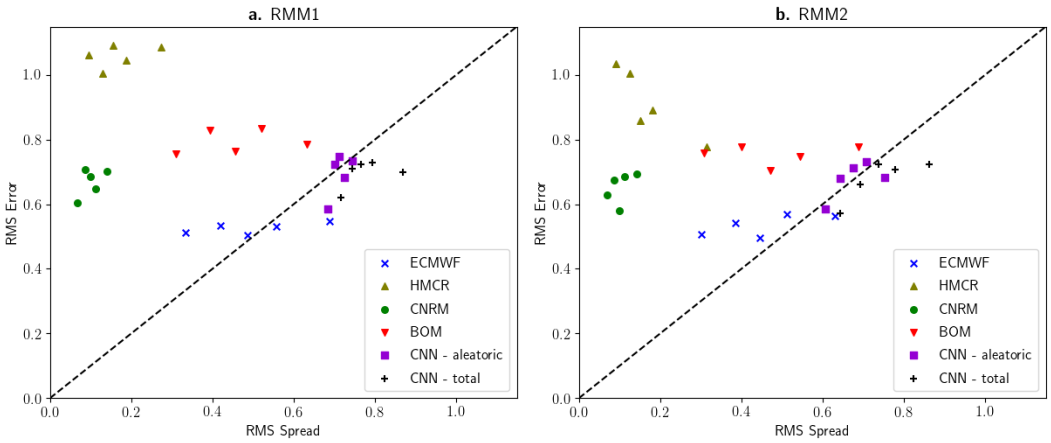


Figure 2. Error-Spread Diagrams for (a) RMM1 and (b) RMM2 at a lead time of ten days. The data are sorted according to the predicted spread before being split into five quintiles. The figure shows the average spread and RMSE for each quintile. Well calibrated forecasts lie on the one-to-one dashed line.

To quantify this property across many lead times, we incrementally remove the days with the highest predicted variance for each lead time and RMM index before computing the RMSE in the forecast of the remaining days. This produces the confidence curve (Scalia et al., 2019). If the forecast correctly ranks different days in terms of forecast un-

certainty, the confidence curve should be strictly decreasing. The error-drop (Figure 1(f)), is the ratio between the last and first points on the confidence curve (Scalia et al., 2019). The smaller the error-drop, the greater the reduction in RMSE when test days are sorted by the forecast uncertainty. The CNN performs better than all dynamical models. It can distinguish between predictable and unpredictable days at all lead times. While an under-dispersive ensemble spread can be corrected to improve the log-score of dynamical models (Figure 1), the ability to sort days according to their predictability cannot be introduced by statistical post-processing.

3.2 Interpretation to validate network behaviour

Before using the CNN to understand sources of uncertainty in the evolution of the MJO, we must understand how the CNN can make skilful forecasts of the MJO. This is necessary, as it reveals any concerning behaviour or spurious correlations (e.g. Lapuschkin et al., 2019), lending confidence to the predictions.

To interpret the CNN mean forecasts, we use the PatternNet algorithm (Kindermans et al., 2017) to derive signal maps for each forecast. These indicate where information is detected by the CNN in each input field. Because the different input variables are introduced as separate channels into the CNN, weights are shared across all variables for much of the network: the CNN distinguishes between variables in the first layer only. It is therefore useful to consider both the signal maps averaged over all variables (the *signal mean*) and the difference between the signal map for each variable and the signal mean map (the *signal anomalies*).

Since propagation over the Maritime Continent is a source of error in MJO forecasts in many models (Kim et al., 2016), we contrast one event which propagated over the Maritime Continent (28/02/2012), and one which decayed (25/02/2006) to validate the CNN's behaviour. Supporting Figure S1 shows the observed RMM indices for these two events, and the corresponding mean forecasts initialised in phase 3, which capture the observed behaviour.

Figure 3(a–b) shows the SHUM400 input fields averaged over all days in RMM phase 3 for the decaying and the propagating events respectively. Panels (c–d) show the signal means for RMM1 for the associated ten-day CNN forecasts initialised in phase 3. (The signal means for the decaying RMM2 are much smaller, consistent with the pre-

diction that day-10 RMM2 is close to zero for the events selected: see Supporting Figure S5). For both events, the CNN signal mean maps show that the CNN integrates over a large region spanning the Indian and Pacific Oceans, rather than tightly focusing on the active MJO region: the CNN also derives information from the input fields in regions of suppressed convection (Feng et al., 2015; Barrett et al., 2021).

Figure 3 (e–f) show the corresponding PatternNet signal anomalies for SHUM400, highlighting the relative information provided by this input field. We see a large reduction in signal over the Pacific (150°E–90°W), and an enhancement over the Maritime Continent (90°E–110°E) co-located with enhanced SHUM400. Supporting Figures S6–S7 show the equivalent figure for OLR. The RMM1 signal anomaly is greater than for SHUM400, and it is stronger over the Pacific than was the case for SHUM400. Both Feng et al. (2015) and Barrett et al. (2021) found OLR precursors in this region which distinguished between propagating and non-propagating MJO events. We conclude that the CNN has identified true predictive features of MJO propagation, giving us confidence in the network.

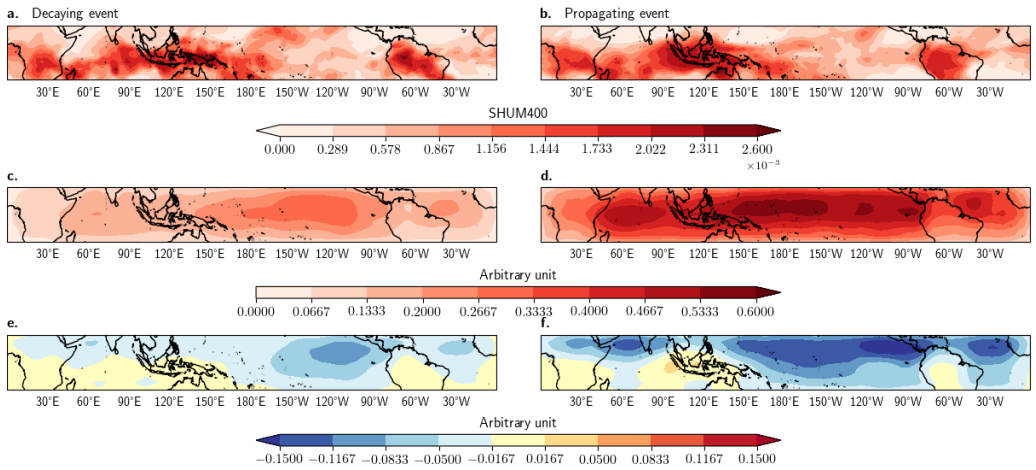


Figure 3. Interpretation of CNN mean forecasts. (a–b) Composite maps of phase-3 SHUM400 for an MJO event which (a) decays and (b) propagates over the Maritime Continent. (c–d) PatternNet RMM1 signal means (averaged over all variables) for ten-day CNN forecasts for the decaying and propagating event respectively. (e–f) RMM1 signal anomalies in SHUM400 for the decaying and propagating events respectively.

3.3 Predictors of uncertainty in MJO forecasts

The ability of the CNN to rank days by uncertainty enables us to investigate drivers of short-term predictability of the MJO. We consider cases in Boreal winter, and separate MJO events into 4 categories according to the CNN's 10-day forecast. We first categorise according to strength: for each day, an event is weak (strong) if the initial observed RMM amplitude is less than (greater than) 1.0. The data are then divided into certain and uncertain forecasts. To study the uncertainty that is directly linked to the MJO initial conditions, we use the network's predicted aleatoric uncertainty. An event is certain (uncertain) if both the RMM1 and RMM2 forecast aleatoric uncertainties are under (over) their respective 30% (70%) percentiles. For each initial observed phase and input feature, we compute the difference between certain and uncertain days, separately for weak and strong events.

Figure 4 shows results for SHUM400 for events starting in phases 3 and 7. For phase 3, the initial conditions of 'certain' forecasts have reduced humidity at the equator in the central Pacific (150°E-120°W) and Indian Ocean (45°E-100°E), combined with off-equatorial regions of enhanced humidity over the Maritime Continent and Australia (100°E-160°E). Before concluding that this 'fingerprint' is an indicator of high certainty, there are two possible confounding factors to consider: the initial strength of the signal, and the forecast strength at day-10. The difference maps for weak and strong events are similar to each other, indicating the fingerprint is independent of initial strength. However, there is a correlation between the forecast uncertainty and the forecast strength at day-10: $\sim$ 65% of 'certain' events are forecast as weak by day-10, while $\sim$ 80% of 'uncertain' events are forecast strong at day-10 (Supporting Table S3). Therefore sorting the data by forecast certainty unintentionally also sorts by forecast strength. To remove this confounding factor, we further stratified the events by strength at day-10. The moisture signal was muted if all events forecast as weak at day-10 were removed from the composites, whereas if only events forecast as transitioning from strong to weak were considered, the signal became more intense (not shown). This confirms that the fingerprint is primarily an indicator of forecast strength at day-10, consistent with the conclusions of (Jiang et al., 2020) who found that this structure hinders the eastward propagation of the MJO.

For events initialised in phase 7, uncertain events show reduced moisture over the Maritime Continent in the MJO suppressed region (90°E-120°E), and enhanced mois-

ture over the MJO active region (150°E-150°W), when compared to certain events. This signature of an enhanced MJO signal in the initial conditions for unpredictable events is observed for other variables for phase 7, particularly OLR (Supporting Figure S8). For events initialised in phase 7, 85% of uncertain forecasts are also likely to be strong at day-10, whereas that drops to 40% for certain forecasts (Supporting Table S4). However, if we further stratify the forecasts by final strength, we find the signature persists (not shown). Thus we conclude that an initially stronger MJO signal is associated with more uncertainty in the forecast.

Finally, we find that MJO predictability is affected by the background state through which it propagates. In particular, for events classified as certain, Z850 shows an enhanced gradient between the Eastern Pacific and the Maritime Continent for all forecasts initialised in phases 4–7 (i.e. all events crossing the Pacific: Supporting Figure S9–S10). An enhanced Z850 gradient is consistent with a higher Southern Oscillation index and a stronger Walker circulation cell over the Pacific. Further stratification by strength at day-10 indicates that this signal is unrelated to forecast strength. An enhanced (neutral or weakened) Walker circulation therefore leads to enhanced (reduced) certainty in the MJO.

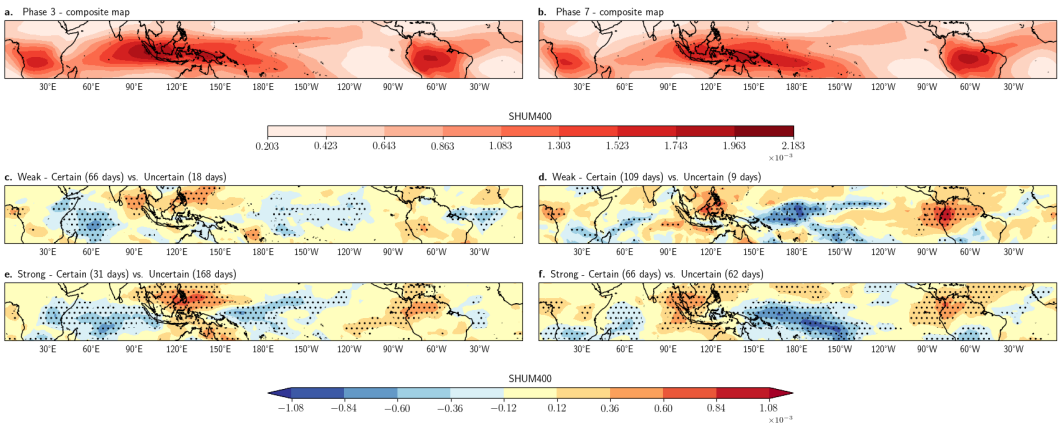


Figure 4. Interpretation of CNN uncertainty forecasts. (a-b) Composite maps of specific Humidity at 400hPa (SHUM400) for extended Boreal winter MJO events in (a) phase 3 and (b) phase 7. (c-f) Difference between input maps for predictable and unpredictable events as classified by ten-day forecasts using the CNN. (c) Weak phase 3 events (d) Weak phase 7 events. (e) Strong phase 3 events (f) Strong phase 7 events. Stippling denotes areas where anomalies are significant at the 95% level using the Student's t-test.

4 Discussion and Conclusions

We presented a CNN which produces probabilistic forecasts of the MJO in terms of means and variances of the bivariate RMM index. The skill of the CNN is competitive with models from the S2S database. Moreover, the CNN outperforms all S2S models for one key forecast property: it can rank start dates according to the forecast uncertainty associated with the initial conditions. In other words, the CNN forecast spread is a dynamic indicator of the uncertainty in the MJO forecast on a given day.

Since the CNN exhibits state-dependent reliability, we identify 'certain' CNN forecasts with predictable states of the MJO and use the CNN forecasts to probe associated sources of predictability. We do this by considering composites of initial conditions which the CNN indicated led to 'certain' and 'uncertain' ten-day forecasts. We found that for forecasts initialised in phase 3, reduced humidity on the equator increases the likelihood of a decaying MJO event, which is associated with high forecast certainty. However, enhanced humidity on the equator increases the likelihood of MJO propagation over the MC, but it does not guarantee propagation, leading to high uncertainty in the forecast and low medium-range predictability.

The CNN also used background state information to determine the MJO's instantaneous predictability. A reduced gradient in Z850 was linked to more forecast uncertainty for all MJO phases approaching the Pacific. This change in Z850 reflects a weaker Walker circulation, associated with El-Niño events. However, we found no consistent signal in East Pacific SST across these phases (Supporting Figures S11–S12). There is substantial debate about the dependency of the MJO on the state of the El Niño-Southern Oscillation (ENSO) (e.g. Ling et al., 2017). The Eastward extent of MJO activity is greater in El Niño years, (Kessler, 2001), and the MJO lifetime and propagation speed is also modulated by ENSO, though it shows sensitivity to the season of interest and type of ENSO event (Pohl & Matthew, 2007; Pang et al., 2016). In contrast, the overall amplitude of MJO activity appears unrelated to ENSO (J. M. Slingo et al., 1999; Kessler, 2001). While the dependency of the MJO on the back-ground state is usually considered in terms of SST, our results demonstrate ENSO could primarily influence the MJO via changes to the atmospheric dynamical background associated with El Niño and La Niña.

Our CNN approach is complementary to earlier MJO predictability studies (e.g. Neena et al., 2014; Wu et al., 2016; Kim et al., 2018). Instead of quantifying the poten-

tial predictability *limit* using our model, we assess relative predictability in the medium-range across different initial conditions. We can only do this because the CNN produces state dependent reliable probabilistic forecasts. Our focus was on forecasts at a lead time of 10-days. Longer lead time forecasts may show a different signal of predictability in the initial conditions: for example, while we found that a weak MJO event predictably decays over a 10-day period, the situation after those 10-days is likely to be more unpredictable than for events where the MJO persists beyond the 10-day period.

The CNN is competitive with the best available dynamical models at predicting the MJO. However CNNs are complementary to dynamical models, and further improvements to MJO forecasting may be achieved through a blend of dynamical and machine learning approaches (Kim et al., 2021). Nevertheless, developing a stand-alone CNN facilitates interpretation, enabling us to probe the performance of the CNN and develop new physical understanding, e.g. the role of different input features. This framework of combining state-dependent uncertainty estimates from neural networks with interpretation techniques could be applied to other climate phenomena, allowing us to quantify the diverse range of sources of uncertainty in the Earth System.

5 Open Research

Data related to this paper can be downloaded from the ERA5 Copernicus database (`https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels`, `https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels`) and the S2S project archive (`http://s2sprediction.net`) via the ECMWF portal: (`https://apps.ecmwf.int/datasets/data/s2s-reforecasts-instantaneous-accum-ecmf/`; `https://apps.ecmwf.int/datasets/data/s2s-reforecasts-instantaneous-accum-rums/`; `https://apps.ecmwf.int/datasets/data/s2s-reforecasts-instantaneous-accum-lfpw/`; `https://apps.ecmwf.int/datasets/data/s2s-reforecasts-instantaneous-accum-ammc/`). The CNN forecasts produced for this paper can be downloaded from `10.5281/zenodo.5175837`. The RMM indices were computed using the CLIVAR diagnostics package (`https://www.ncl.ucar.edu/Applications/mjoclivar.shtml`). PyTorch (`https://www.pytorch.org`) and DropBlock (`https://github.com/miguelvr/dropblock`) libraries were implemented to build and train the CNN model. PatternNet code was adapted from `https://github.com/TNTLFreiburg/pytorch_patternnet`. The

codes used in the current analysis are available at `https://github.com/antoine-delaunay/DeepLearningMJO/` .

Acknowledgments

H.M.C. was funded by Natural Environment Research Council grant number NE/P018238/1. Thanks to D.J. Gagne and an anonymous reviewer for their comments which improved this manuscript.

References

Arcomano, T., Szunyogh, I., Pathak, J., Wikner, A., Hunt, B. R., & Ott, E. (2020). A machine learning-based global atmospheric forecast model. *Geophysical Research Letters*, *47*(9).

Barrett, B. S., Densmore, C. R., Ray, P., & Sanabia, E. R. (2021, March). Active and weakening MJO events in the Maritime Continent. *Climate Dynamics*.

Bauer, P., Thorpe, A., & Brunet, G. (2015, sep). The quiet revolution of numerical weather prediction. *Nature*, *525*(7567), 47–55. Retrieved from `http://www.nature.com/doifinder/10.1038/nature14956`

Bjerregård, M. B., Møller, J. K., & Madsen, H. (2021, June). An introduction to multivariate probabilistic forecast evaluation. *Energy and AI*, *4*, 100058. Retrieved 2021-06-17, from `https://linkinghub.elsevier.com/retrieve/pii/S2666546821000124` doi: 10.1016/j.egyai.2021.100058

Bolin, D., & Wallin, J. (2019). Local scale invariance and robustness of proper scoring rules. , 1–26. Retrieved from `http://arxiv.org/abs/1912.05642`

Cassou, C. (2008). Intraseasonal interaction between the Madden-Julian Oscillation and the North Atlantic Oscillation. *Nature*, *455*(7212), 523–527.

Christensen, H. M., Moroz, I. M., & Palmer, T. N. (2015). Evaluation of ensemble forecast uncertainty using a new proper score: Application to medium-range and seasonal forecasts. *Quarterly Journal of the Royal Meteorological Society*, *141*(687), 538–549. doi: 10.1002/qj.2375

F. Vitart et al. (2017, January). The Subseasonal to Seasonal (S2S) Prediction Project Database. *Bulletin of the American Meteorological Society*, *98*(1), 163–173. (`ftp://s2sidx:s2sidx@acquisition.ecmwf.int/RMMS`)

Feng, J., Li, T., & Zhu, W. (2015). Propagating and nonpropagating MJO events

over maritime continent. *Journal of Climate*, *28*(21), 8430–8449. doi: 10.1175/JCLI-D-15-0085.1

Ferranti, L., Palmer, T., Molteni, F., & Klinker, E. (1990). Tropical-extratropical interaction associated with the 30–60 day oscillation and its impact on medium and extended range prediction. *Journal of Atmospheric Sciences*, *47*(18), 2177–2199.

Gal, Y. (2016). *Uncertainty in Deep Learning* (Unpublished doctoral dissertation). Cambridge University.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning* (pp. 1050–1059).

Ghiasi, G., Lin, T.-Y., & Le, Q. V. (2018, October). DropBlock: A regularization method for convolutional networks. *arXiv:1810.12890 [cs]*. (arXiv: 1810.12890)

H. Hersbach et al. (2018a). ERA5 hourly data on pressure levels from 1979 to present. *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*. (Accessed on 01-03-2021, `https://doi.org/10.24381/cds.bd0915c6`)

H. Hersbach et al. (2018b). ERA5 hourly data on single levels from 1979 to present. *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*. (Accessed on 01-03-2021, `https://doi.org/10.24381/cds.adbb2d47`)

Ham, Y.-G., Kim, J.-H., & Luo, J.-J. (2019). Deep learning for multi-year enso forecasts. *Nature*, *573*(7775), 568–572.

Henderson, G. R., Barrett, B. S., & M. Lafleur, D. (2014). Arctic sea ice and the Madden–Julian Oscillation (MJO). *Climate Dynamics*, *43*(7-8), 2185–2196.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, *15*(5), 559–570. doi: 10.1175/1520-0434(2000)015⟨0559:DOTCRP⟩2.0.CO;2

Hersbach, H., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. doi: 10.1002/qj.3803

Huntingford, C., Jeffers, E. S., Bonsall, M. B., Christensen, H. M., Lees, T., & Yang, H. (2019). Machine learning and artificial intelligence to aid climate change research and preparedness. *Environmental Research Letters*, *14*(12). doi: 10.1088/1748-9326/ab4e55

Jiang, X., Maloney, E., & Su, H. (2020, December). Large-scale controls of propagation of the Madden-Julian Oscillation. *npj Climate and Atmospheric Science*, *3*(1), 29.

Kessler, W. S. (2001). EOF representations of the Madden-Julian and its connection with ENSO. *Journal of Climate*, *14*(13), 3055–3061. doi: 10.1175/1520-0442(2001)014⟨3055:EROTMJ⟩2.0.CO;2

Kim, H., Ham, Y. G., Joo, Y. S., & Son, S. W. (2021, December). Deep learning for bias correction of MJO prediction. *Nature Communications*, *12*(1), 3087.

Kim, H., Kim, D., Vitart, F., Toma, V. E., Kug, J.-S., & Webster, P. J. (2016, June). MJO Propagation across the Maritime Continent in the ECMWF Ensemble Prediction System. *Journal of Climate*, *29*(11), 3973–3988.

Kim, H., Vitart, F., & Waliser, D. E. (2018, December). Prediction of the Madden–Julian Oscillation: A Review. *Journal of Climate*, *31*(23), 9425–9443. Retrieved from `https://journals.ametsoc.org/doi/10.1175/JCLI-D-18-0210.1` doi: 10.1175/JCLI-D-18-0210.1

Kindermans, P.-J., Schütt, K. T., Alber, M., Müller, K.-R., Erhan, D., Kim, B., & Dähne, S. (2017, October). Learning how to explain neural networks: PatternNet and PatternAttribution. *arXiv:1705.05598 [cs, stat]*. Retrieved 2021-05-18, from `http://arxiv.org/abs/1705.05598` (arXiv: 1705.05598)

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K. R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, *10*(1), 1–8. Retrieved from `http://dx.doi.org/10.1038/s41467-019-08987-4` doi: 10.1038/s41467-019-08987-4

Leutbecher, M., & Palmer, T. (2008, March). Ensemble forecasting. *Journal of Computational Physics*, *227*(7), 3515–3539. Retrieved 2021-05-27, from `https://linkinghub.elsevier.com/retrieve/pii/S0021999107000812` doi: 10.1016/j.jcp.2007.02.014

Li, X., Yin, M., Chen, X., Yang, M., Xia, F., Li, L., ... Zhang, C. (2020). Impacts of the Tropical Pacific–Indian Ocean Associated Mode on Madden–Julian Oscillation over the Maritime Continent in Boreal Winter. , 14.

Lim, Y., Son, S.-W., & Kim, D. (2018, May). MJO Prediction Skill of the Subseasonal-to-Seasonal Prediction Models. *Journal of Climate*, *31*(10),

4075–4094. Retrieved 2021-05-05, from http://journals.ametsoc.org/doi/10.1175/JCLI-D-17-0545.1 doi: 10.1175/JCLI-D-17-0545.1

Ling, J., Li, C., Li, T., Jia, X., Khouider, B., Maloney, E., ... Zhang, C. (2017). Challenges and opportunities in MJO studies. *Bulletin of the American Meteorological Society*, *98*(2), ES53–ES56. doi: 10.1175/BAMS-D-16-0283.1

Madden, R. A., & Julian, P. R. (1971). Detection of a 40–50 day oscillation in the zonal wind in the Tropical Pacific. *Journal of the Atmospheric Sciences*, *28*, 702–708. doi: 10.1017/CBO9781107415324.004

Marshall, A. G., Hendon, H. H., & Hudson, D. (2016). Visualizing and verifying probabilistic forecasts of the Madden-Julian Oscillation. *Geophysical Research Letters*, *43*(23), 12,278–12,286. doi: 10.1002/2016GL071423

Martin, Z., Barnes, E., & Maloney, E. (2021, March). *Predicting the MJO using interpretable machine-learning models* (Tech. Rep.). Atmospheric Sciences. Retrieved 2021-04-01, from http://www.essoar.org/doi/10.1002/essoar.10506356.1 doi: 10.1002/essoar.10506356.1

Neena, J. M., Lee, J. Y., Waliser, D., Wang, B., & Jiang, X. (2014). Predictability of the Madden-Julian oscillation in the Intraseasonal Variability Hindcast Experiment (ISVHE). *Journal of Climate*, *27*(12), 4531–4543. doi: 10.1175/JCLI-D-13-00624.1

Palmer, T. N. (2000). Predicting uncertainty in forecasts of weather and climate. *Reports on Progress in Physics*, *63*(2), 71–116. doi: 10.1088/0034-4885/63/2/201

Pang, B., Chen, Z., Wen, Z., & Lu, R. (2016). Impacts of two types of El Niño on the MJO during boreal winter. *Advances in Atmospheric Sciences*, *33*(8), 979–986. doi: 10.1007/s00376-016-5272-2

Pohl, B., & Matthew, A. J. (2007). Observed changes in the lifetime and amplitude of the Madden-Julian oscillation associated with interannual ENSO sea surface temperature anomalies. *Journal of Climate*, *20*(11), 2659–2674. doi: 10.1175/JCLI4230.1

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. Retrieved from http://dx.doi.org/10.1038/s41586-019-0912-1 doi:

10.1038/s41586-019-0912-1

Roulston, M. S., & Smith, L. A. (2002). Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, *130*(6), 1653–1660. doi: 10.1175/1520-0493(2002)130⟨1653:EPFUIT⟩2.0.CO;2

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... others (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*(3), 211–252.

Scalia, G., Grambow, C. A., Pernici, B., Li, Y.-P., & Green, W. H. (2019, October). Evaluating Scalable Uncertainty Estimation Methods for DNN-Based Molecular Property Prediction. *arXiv:1910.03127 [cs, stat]*. (arXiv: 1910.03127)

Schultz, M., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L., ... Stadtler, S. (2021). Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society A*, *379*(2194).

Slingo, J., & Palmer, T. (2011). Uncertainty in weather and climate prediction. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *369*(1956), 4751–4767.

Slingo, J. M., Rowell, D. P., Sperber, K. R., & Nortley, F. (1999). On the predictability of the interannual behaviour of the Madden-Julian Oscillation and its relationship with El Nino. *Quarterly Journal of the Royal Meteorological Society*, *125*(554), 583–609. doi: 10.1256/smsqj.55410

Toms, B. A., Kashinath, K., Prabhat, & Yang, D. (2020). Testing the reliability of interpretable neural networks in geoscience using the madden-julian oscillation. *Geoscientific Model Development Discussions*, *2020*, 1–22. Retrieved from `https://gmd.copernicus.org/preprints/gmd-2020-152/` doi: 10.5194/gmd-2020-152

Translational Neurotechnology Lab. (2019). PatternNet GitHub Repository. (University of Freiburg, `https://github.com/TNTLFreiburg/pytorch\_patternnet`)

Vitart, F. (2017, July). Madden—Julian Oscillation prediction and teleconnections in the S2S database. *Quarterly Journal of the Royal Meteorological Society*, *143*(706).

Vitart, F., & Robertson, A. W. (2018). The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. *npj Climate and Atmo-*

spheric Science, *1*(1), 1–7. Retrieved from http://dx.doi.org/10.1038/s41612-018-0013-0 doi: 10.1038/s41612-018-0013-0

Wheeler, M. C., & Hendon, H. H. (2004). An All-Season Real-Time Multivariate MJO Index: Development of an Index for Monitoring and Prediction. *Monthly Weather Review*, *132*, 16.

Wu, J., Ren, H. L., Zuo, J., Zhao, C., Chen, L., & Li, Q. (2016). MJO prediction skill, predictability, and teleconnection impacts in the Beijing Climate Center Atmospheric General Circulation Model. *Dynamics of Atmospheres and Oceans*, *75*, 78–90. Retrieved from http://dx.doi.org/10.1016/j.dynatmoce.2016.06.001 doi: 10.1016/j.dynatmoce.2016.06.001

Yoo, C., Lee, S., & Feldstein, S. B. (2012). Arctic response to an mjo-like tropical heating in an idealized gcm. *Journal of Atmospheric Sciences*, *69*(8), 2379–2393.

Zhang, C. (2013). Madden-Julian Oscillation: Bridging weather and climate. *Bulletin of the American Meteorological Society*, *94*(12), 1849–1870. doi: 10.1175/BAMS-D-12-00026.1

Supporting Information for "Interpretable Deep Learning for Probabilistic MJO Prediction"

Antoine Delaunay[1]and Hannah M. Christensen[2]

[1]Department of Applied Mathematics, Ecole Polytechnique, Palaiseau, France

[2]Department of Physics, University of Oxford, Oxford, UK

Contents of this file

1. Text S1 to S6
2. Figures S1 to S12
3. Tables S1 to S4

Introduction

In this Supporting Information we provide further details on the methodology used in our study. Text S1 and Table S1 provide details of the Subseasonal-to-seasonal (S2S) prediction project data used as a benchmark for the CNN performance. Text S2 describes the observational data and preprocessing used to train the model. Text S3 provides more details of the CNN forecasting model, focusing on the techniques used to represent epistemic and aleatoric uncertainty, and Figure S2 shows the CNN architecture. Figure S3 shows sensitivity of the CNN performance to chosen input fields. Text S4 details the

PatternNet algorithm. Text S5 details the validation metrics. Text S6 proves the validity of the DropBlock approach in place of standard dropout for convolutional layers.

We also provide further results to support our conclusions. Figure S4 shows the bivariate correlation skill for the CNN and S2S models. Figure S1 shows the phase diagram corresponding to the decaying and propagating events analysed in Section 3.2 of the manuscript, while Figures S5–S7 show further interpretation of the CNN forecasts for those events. Figures S8–S12 show further results concerning predictors of uncertainty in MJO forecasts for outgoing longwave radiation (OLR), 850 hPa geopotential (Z850), and sea surface temperature (SST).

Text S1. *Subseasonal-to-Seasonal forecast model data*

We select four representative models from the Subseasonal-to-Seasonal (S2S) prediction project database (F. Vitart et al., 2017) for comparison with the CNN. The database consists of near real-time operational ensemble forecasts and reforecasts from 11 centres. As the operational models are continuously improved, the skill of the forecasts evolves in time. For that reason, we select the reforecasts for comparison with the CNN. Reforecasts are forecasts made retrospectively using a single up-to-date version of the dynamical model.

Details of the available reforecast data for the selected models are presented in Table S1. Some observations have to be made: first, all models do not have the same number of members, so to be consistent we decided to restrict the number of members to 10. Second, as the computational cost is heavy, the reforecasts are not made every day and consequently each model has a different reforecasting period and time range. This is an issue which is difficult to overcome but with a large enough number of days, we should still be able to make fair comparisons.

Text S2. *Observational data and preprocessing*

We train the CNN using atmospheric data from the ECMWF ERA5-Reanalysis dataset (H. Hersbach et al., 2018a), (H. Hersbach et al., 2018b). The inputs are maps of daily averaged fields from 1979 to 2019 with a spatial coverage of 0 – 360°E, 20°S – 20°N on a 2.5° x 2.5° grid. We only use ERA5 data over the satellite era for which accurate estimates of OLR are available. Selected variables are: zonal wind at 200 hPa and 850 hPa (UA200, UA850), Outgoing Long-Wave Radiation (OLR), Sea Surface Temperature

(SST), Specific Humidity at 400 hPa (SHUM400), Geopotential at 850 hPa (Z850), and Downward Long-Wave Radiation at the surface (DLR). For UA200, UA850 and OLR, we apply the RMM preprocessing transform of (Wheeler & Hendon, 2004) to leave only subseasonal anomalies: the time mean, the first three Fourier harmonics and the 120-day running mean are removed sequentially. For SST, we subtract the climatological mean (for each date of the calendar year, we compute the average over the same date for all the years in the training dataset), and set all inland grid points to zero. The raw data is used for SHUM400, Z850, and DLR, allowing the network to learn seasonal variations in MJO predictability. Finally for every variable, we rescale the inputs to between 0 and 1 independently at every gridpoint with a Min-Max scaling to ensure the stability of the training.

Our choice of input fields for the CNN was guided by an iterative procedure. The first network we trained took as input the three variables used to define the RMM index: UA200, UA850 and OLR subseasonal anomalies. Subsequent networks were trained using one or more additional variables, and the predictive performance of the network assessed. Supporting Figure S3 shows the relative benefit of including each of the additional input variables selected for the final network: sea surface temperature anomalies (SST), daily downwelling long-wave radiative forcing (DLR), daily geopotential at 850 hPa (Z850), and specific humidity at 400 hPa (SHUM400). We compared the performance of the final network to a network trained on DLR, Z850 and SHUM400 *anomalies* instead of *means*, but found this degraded performance. We also considered including the values of fields

at earlier timesteps (5, 10 days before), but found this did not improve the network's performance, and instead led to overfitting.

Text S3. *The CNN forecasting model*

For an initial date t and a forecast range τ, let $\mathbf{x}_t$ be the input at the date t, and $\mathbf{y}_{t+\tau}$ be the observed RMM indices, $\mathbf{y}_{t+\tau} = (\mathrm{RMM1}_{t+\tau}, \mathrm{RMM2}_{t+\tau})$ at the chosen lead time, τ. The input $\mathbf{x}_t$ is a series of gridded maps representing physical quantities (variables) for each date t as a function of latitude and longitude. We train a separate network for each forecast range τ, where τ takes discrete values: $\tau = 1, 3, 5, 10, 15, 20, 25, 30, 35$ days.

Aleatoric uncertainty is caused by the chaotic nature of the system. Physically, we recognise that the input variables supplied to the CNN are a subset of all possible variables, and only include information on scales larger than the resolution of the input maps, such that the future state of the MJO is not a deterministic function of these inputs[1]. This uncertainty is a property of the data and thus irreducible, regardless of the model's training. It is also heteroscedastic, or state-dependent. The predicted aleatoric uncertainty is included as an output of the CNN. We assume the RMM indices follow a Gaussian bivariate distribution with a null correlation between RMM1 and RMM2 (Wheeler & Hendon, 2004). The probabilistic network therefore has a 4-neuron output consisting of the forecast mean, $\boldsymbol{\mu}_{t+\tau}$, and variance $\boldsymbol{\sigma}^2_{a\ t+\tau}$, where the first and second entries of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2_a$ correspond to RMM1 and RMM2 respectively. Aleatoric uncertainty is accounted for in the loss function: the model is trained by maximising the log-likelihood:

$$L = \frac{1}{N}\sum_{t=1}^{N} -\frac{1}{2}[ln(|\mathbf{\Sigma}_{t+\tau}|) + (\mathbf{y}_{t+\tau} - \boldsymbol{\mu}_{t+\tau})^T \mathbf{\Sigma}^{-1}_{t+\tau}(\mathbf{y}_{t+\tau} - \boldsymbol{\mu}_{t+\tau})] + ln(2\pi) \qquad (1)$$

where $\mathbf{\Sigma}_t$ is the diagonal covariance matrix and N is the number of samples per batch.

The epistemic uncertainty in the forecast is due to uncertainty on the CNN's weights θ. We recognise that the training dataset (X, Y) is a sample from the true joint distribution of inputs, X, and outputs, Y. We therefore seek the distribution $p(\theta \mid X, Y)$ over θ instead of a single estimate. The Monte-Carlo dropout method approximates $p(\theta \mid X, Y)$ by a parametric distribution $q_\Phi(\theta)$, where Φ is a vector of parameters to tune. Following (Scalia et al., 2019), we model $q_\Phi(\theta)$ as a Bernoulli distribution, β_Φ. In other words, for a given set of input fields, each of the CNN's weights is deactivated with a probability set by the vector Φ, representing the dropout rate of each layer. For the j^{th} parameter this gives $\theta_j \sim \hat{\theta}_j * \beta_{\Phi\, j}$.

Dropout is applied to the network at both training and testing time. During training, dropout prevents overfitting by randomly deactivating some neurons at each epoch. It ensures the predictive capability of the network is distributed across all neurons, instead of converging to a solution in which certain neurons dominate. During testing, we use dropout to produce M Monte-Carlo forecasts, $(\theta^{(i)}, \boldsymbol{\mu}^{(i)}_{t+\tau}, \boldsymbol{\sigma}^{(i)\,2}_{a\;t+\tau})$. We chose $M = 10$ for consistency with the ensemble size of dynamical MJO forecasts, though the computational efficiency of the CNN would enable vastly larger ensemble sizes than this. For the linear layers of the CNN, we apply standard dropout with a dropout rate of 0.3. However, this is not suitable for convolutional layers, because neighbouring points in the feature maps for each layer are often highly correlated (Ghiasi et al., 2018). Instead, we use a DropBlock approach, with a dropout rate of 0.1 for the first convolutional layer and 0.3 for subsequent convolutional layers. In DropBlock, a fraction of points of the maps are randomly set to zero, before all their neighbouring points are also deactivated (Ghiasi

et al., 2018). In this way, DropBlock introduces a correlation between inactive points. However, in contrast to standard dropout, DropBlock disables points on the feature maps and not the weights directly. In Text S6 we demonstrate that deactivating points on the input maps as is carried out in DropBlock is equivalent to the weight deactivation applied in standard dropout, for the case of convolutional layers without bias. This allows us to combine the dropout and DropBlock techniques to represent epistemic uncertainty in the CNN.

Finally, the estimated aleatoric and epistemic uncertainties are combined to give the final predicted mean and total variance:

$$\boldsymbol{\mu}_{t+\tau} = \frac{1}{M}\sum_i \boldsymbol{\mu}^{(i)}_{t+\tau} \tag{2}$$

$$\boldsymbol{\sigma}^2_{tot\ t+\tau} = \frac{1}{M}\sum_i \boldsymbol{\sigma}^{(i)\ 2}_{a\ t+\tau} + \mathrm{Var}(\boldsymbol{\mu}^{(i)}) \tag{3}$$

The network's architecture is shown in supporting figure S2. Our network has three convolutional layers without bias. For all of them, we used Leaky ReLU with $\alpha = 0.003$ as activation function to avoid vanishing gradients. Each of the two first convolutional layers have a (5,5) kernel and are followed by average pooling with a (3,3) kernel size and a (2,1) stride. The third convolutional layer has a (3,3) kernel size. Convolutional layers are followed by two fully-connected layers with 1920 and 200 neurons. The output layer has 4 neurons: the forecast means and aleatoric variances of RMM1 and RMM2. To ensure the output variances are positive, we apply the function $f : x \mapsto \log(1 + \exp(x))$ which we found more stable than ReLU. We train the network with batches of 50 samples up to 35 epochs. Given the large amount of data required to train a deep learning model,

the amount of data needed to make reasonable comparisons with dynamic models and the fact that the Outgoing Longwave Radiation data goes back to 1979 at the most, we made the choice to keep only one train set and one test set and to tune the network parameters (kernel, strides, etc.) on the train set. In order to prevent the overfitting that could result from this choice, we used an L2 regularization in addition to dropout and DropBlock. We observed some sensitivity of the network with respect to the regularization coefficient λ on the train set performance. Hence to avoid overoptimistic results as much as possible, we kept $\lambda = 0.01$, the L2 coefficient that had the best performance on the train set amongst the values of λ high enough to prevent overfitting.

Text S4. *PatternNet*

PatternNet propagates the estimated signal from the output to the input space. Instead of weights, each convolutional or feed-forward layer in PatternNet consists of statistical attribution vectors, which are chosen to maximise certain functions of the covariance between the signal and the output (Kindermans et al., 2017). These vectors are computed layerwise during a training phase, using input fields and corresponding CNN forecasts from the training dataset, and with knowledge of the CNN network weights. Once these vectors are computed, PatternNet is a backpropagation algorithm. The signal s^l at layer l coming from the neuron i is obtained by multiplying the signal s_i^{l+1} of neuron i in the previous layer, $l+1$, with the attribution vector $\mathbf{a}^l$. The signal s_j^l of neuron j in layer l is then the sum of all the signals of its input neurons from layer $l+1$: $s_j^l = \sum_i s_i^{l+1,\,\mathrm{j}}$.

We used the PyTorch implementation of PatternNet by (Translational Neurotechnology Lab, 2019). During backpropagation, when a ReLU layer is encountered, the signal is

backpropagated without modification if the neuron was active during the forward pass and set to zero otherwise. However, the case of Average Pooling and Leaky ReLU layers has not been addressed (Kindermans et al., 2017; Translational Neurotechnology Lab, 2019). For Average Pooling layers the output neuron is an average of input neurons: for such layers we backpropagate the output neuron signal to the inputs without modification. For Leaky ReLU layers, the signal is backpropagated as follows: if the input was positive in the forward pass, the signal is backpropagated without modification, otherwise it is multiplied by the parameter of the Leaky ReLU function (4).

$$s_j^l = \begin{cases} s_j^{l+1} \text{ if positive input in the forward pass} \\ \\ \alpha s_j^{l+1} \text{ otherwise} \end{cases} \tag{4}$$

When using the PatternNet, we use the forecasts of the single CNN member without dropout to simplify the computation. For a given input and corresponding forecast from the CNN, PatternNet provides signals S1 and S2 for each pixel in the input fields, corresponding to RMM1 and RMM2 respectively. The signals S1 and S2 can take any real value. We are interested in signal amplitude and not direction, and so take the absolute value, and then rescale the signals to between 0 and 1. The Signal Mean Maps in Supporting Figures 5–7 are computed with the signals from the test dataset.

Text S5. *Validation Metrics*

CNN and S2S dynamical model forecasts were validated using days with initial observed amplitude above 1.0 (Lim et al., 2018). Three deterministic metrics were considered. The Root Mean Square Error between the forecast and observed RMMs is defined as

$$RMSE(\tau) = \sqrt{\frac{1}{N}\sum_{t=1}^{N}[(f_1(t,\tau) - v_1(t))^2 + (f_2(t,\tau) - v_2(t))^2]} \tag{5}$$

where f_1, f_2 are the forecast mean RMM indices for start date t at lead time τ, v_1, v_2 are the verification RMM indices at that time, and N is the total number of start dates.

The MJO amplitude is defined as

$$A(t,\tau) = \sqrt{(RMM_1(t,\tau)^2 + RMM_2(t,\tau)^2)} \tag{6}$$

The MJO Bivariate Correlation (Supplementary Figure S4) is defined as

$$BV(t,\tau) = \frac{\sum_{t=1}^{N} f_1(t,\tau)v_1(t) + f_2(t,\tau)v_2(t)}{\sqrt{\sum_{t=1}^{N}(f_1(t,\tau)^2 + f_2(t,\tau)^2)} + \sqrt{\sum_{t=1}^{N} v_1(t)^2 + v_2(t)^2}} \tag{7}$$

The amplitude error can then be written

$$ERR_A = \frac{1}{N}\sum_{t=1}^{N}(A_f - A_v) \tag{8}$$

where A_f and A_v are the forecast and verification amplitudes respectively.

Following (Kim et al., 2018), the MJO phase is defined as

$$ERR_P = \frac{1}{N}\sum_{i=1}^{N} \text{atan}(\frac{v_1(t)f_2(t,\tau) - v_2(t)f_1(t,\tau)}{v_1(t)f_1(t,\tau) + v_2(t)f_2(t,\tau)})) \tag{9}$$

Three further scoring rules were used to assess the probabilistic skill of the forecasts. The Continuous Ranked Probability Score (CRPS, (Hersbach, 2000)) is widely used to validate ensemble forecasts.

$$\text{CRPS}(P_f, v) = \int_{-\infty}^{\infty} \left[P_f(x) - \Theta(x - v)\right]^2 \mathrm{d}x, \tag{10}$$

where P_f is the forecast cumulative distribution function, and $\Theta(x - v)$ is the observed cumulative distribution function, which is equal to the Heaviside step function centres on the verification, v. Gaussianity is assumed for forecasts. Following Marshall et al. (2016),

the CRPS of a given day is computed as the sum of the CPRS for RMM1 and RMM2 separately. Then the resulting CRPS are averaged across the whole dataset.

For consistency with the loss function, the log-score, or Ignorance score (Roulston & Smith, 2002), score (equation (1)) was also used to assess forecast skill, where the number of samples N was the number of days in the test data set.

To assess the ability of forecasts to discern predictable from unpredictable days, we compute Error-Spread diagrams following (Leutbecher & Palmer, 2008). For each day we have a data triplet consisting of a predicted mean, variance, and an observed value for RMM1 and RMM2 respectively. We first sort the triplets according to the predicted variance into 5 equally-populated bins. Then for each bin we compute the root mean variance, and the root mean-squared error between the forecast mean and the observation. This process is repeated for RMM1 and RMM2 separately, for each forecast lead time, and for each type of uncertainty (epistemic, aleatoric, and total). For well calibrated forecasts, the average RMSE in each bin should equal the root mean variance.

We also compute the confidence curve $C(\alpha)$ and Error-Drop for each model. For each lead time and each RMM index in turn, we remove the $\alpha\%$ most uncertain cases, and compute the RMSE between the forecast mean and the observed RMM index for the remaining data. We repeat this process setting α to be each of the 20 evenly-spaced quantiles of the RMSE in turn. The Error-Drop is computed from the confidence curve as:

$$\text{Error-Drop} := \frac{C(\alpha_{max})}{C(\alpha_{min})}, \tag{11}$$

where α_{min} and α_{max} correspond to the minimum and maximum fraction of days removed respectively. We set $\alpha_{min} = 0.0$ and $\alpha_{max} = 0.95$.

Text S6. *Monte-Carlo Dropout for DropBlock*

Here, we prove that we can use the Monte-Carlo Dropout method with DropBlock. In our model, the DropBlock is applied after each convolutional layer and these layers do not have bias. Considering a specific convolutional layer layer l, if we denote X the input of this layer, (Gal, 2016) showed that the convolutional operation could be seen as a matrix multiplication $W^T X$ where W is a convolutional weight matrix, rewritten to match the matrix multiplication operation. If we denote m the number of lines of W^T, we can rewrite $W^T X$ with the dot products $W^T X = (W_1^T X, W_2^T X, ..., W_m^T X)^T$.

As we have considered the convolutional operation as a matrix multiplication, $W^T X$ is a column vector of size m. We must rewrite this vector as an output feature map of shape $n \times p$ denoted F, such that $n * p = m$. Each coefficient F_{ij} is equal to a one of the dot products $W_k^T X$. n and p depend on the convolutional parameters (kernel size, stride, dilation).

Then we apply DropBlock. In a first step, each F_{ij} is independently multiplied by a Bernoulli $\beta(p)$. Then in a second time, for each F_{ij}, we consider all its neighbours. We denote d_{ij} the number of neighbours of F_{ij} (in particular, there are less neighbours on the edges than in the center). F_{ij} is disabled if one of its neighbours (or itself) has been disabled during the first step. It is equivalent as considering that F_{ij} has been multiplied by a Bernoulli $\beta(p^{d_{ij}})$. Hence we can write that after the DropBlock,

$$F_{ij}^{*} \;=\; F_{ij}\beta(p^{d_{ij}}) = W_k^T \beta(p^{d_k})X \tag{12}$$

$$W_k^{T\,*} \;=\; W_k^T \beta(p^{d_k}) \tag{13}$$

Thus we conclude that DropBlock applied after a convolutional layer without bias is equivalent to a standard Dropout with a distinct dropout probability for each weight, which can be achieved using the Monte-Carlo Dropout method.

Notes

1. Note that even if the network were supplied with the highest resolution observational data available, these estimates of the observed Earth System would have a finite resolution and would contain errors, thus aleatoric uncertainty remains.

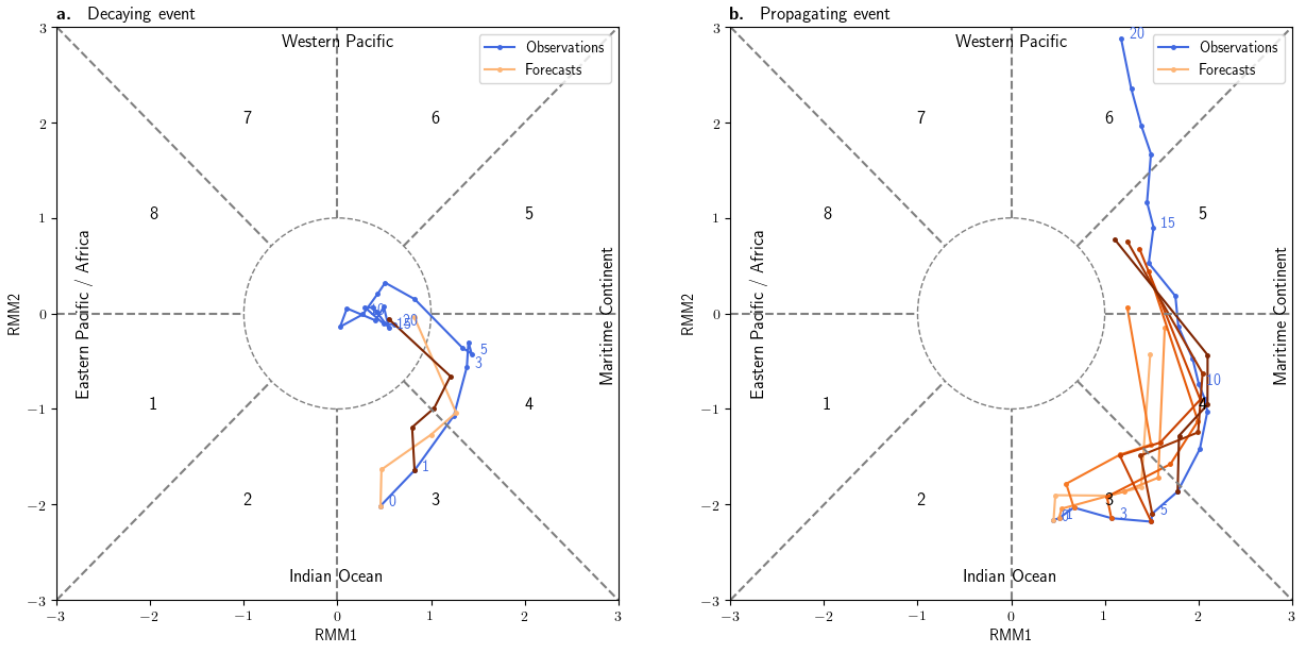


Figure S1. Phase diagrams for a decaying and a propagating event for forecasts initialised in phase 3.

Observations from the first day - 25/02/2006 (**a.**) and 28/02/2012 (**b.**) - are represented up to day-10 in blue. All forecasts (day-1, 3, 5, 10) which began in initial observed phase 3 for each chosen event are represented in shades of orange.

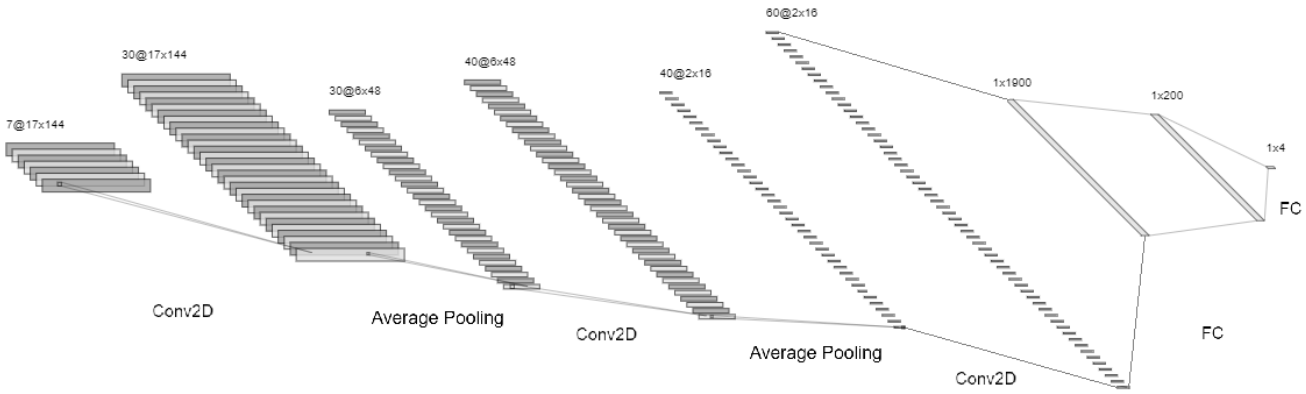


Figure S2. CNN Architecture. Leaky ReLU was used as the activation function of the convolutional layers 1 and 2.

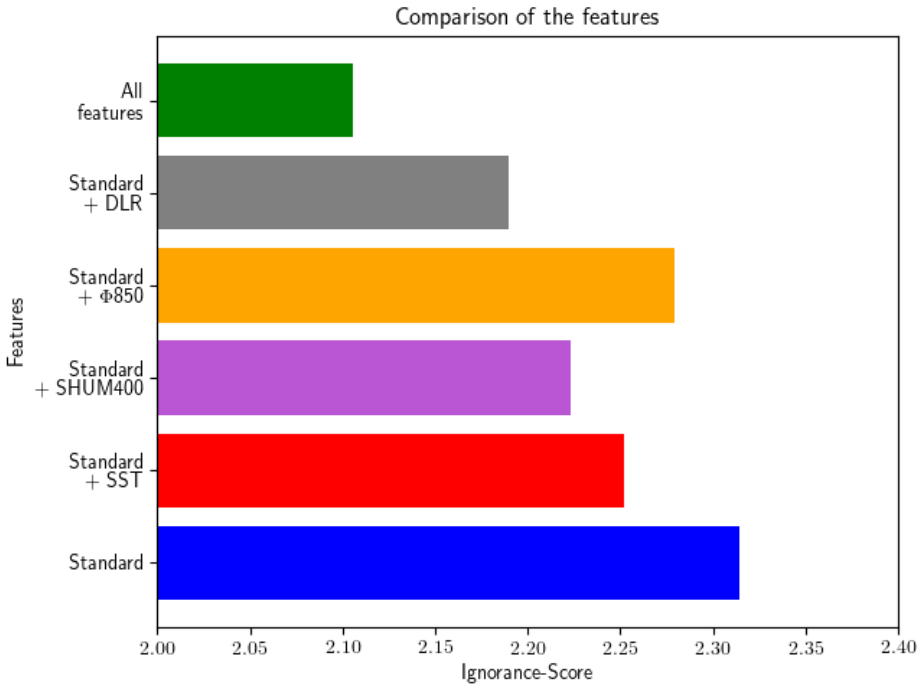


Figure S3. Comparison of the features' performance. Log-score is computed for a CNN trained on different subsets of input features for day-10 forecasts. Days used have initial amplitude above 1.0. Standard stands for "UA200 + UA850 + OLR".

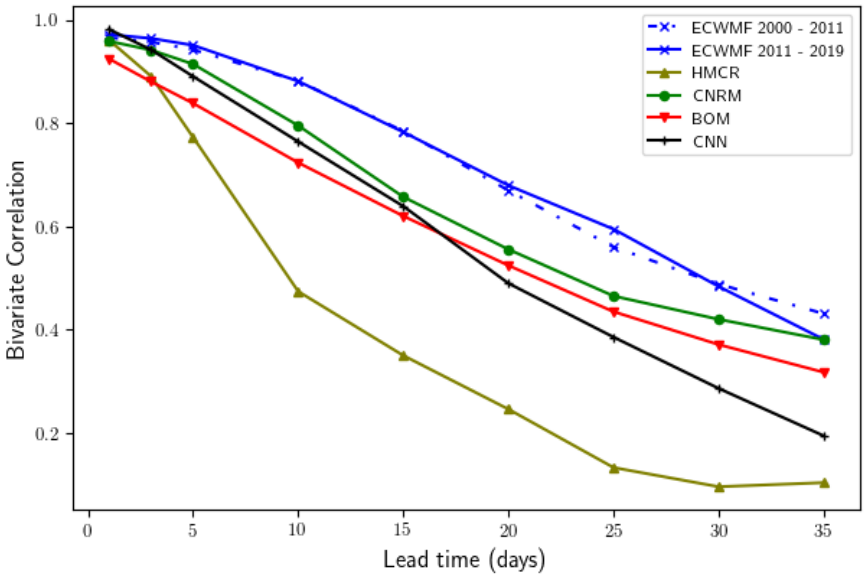


Figure S4. Bivariate correlation computed for RMM1 and RMM2 as a function of lead time. Note that forecasts from different models cover different dates: ECMWF 2000-2019; HMCR 1985-2010; CNRM 1993-2017; BOM 1982-2013; CNN 2011-2019. The ECMWF data was split into two periods to allow direct comparison with the CNN over 2011-2019, and to give an indication of sampling uncertainty.

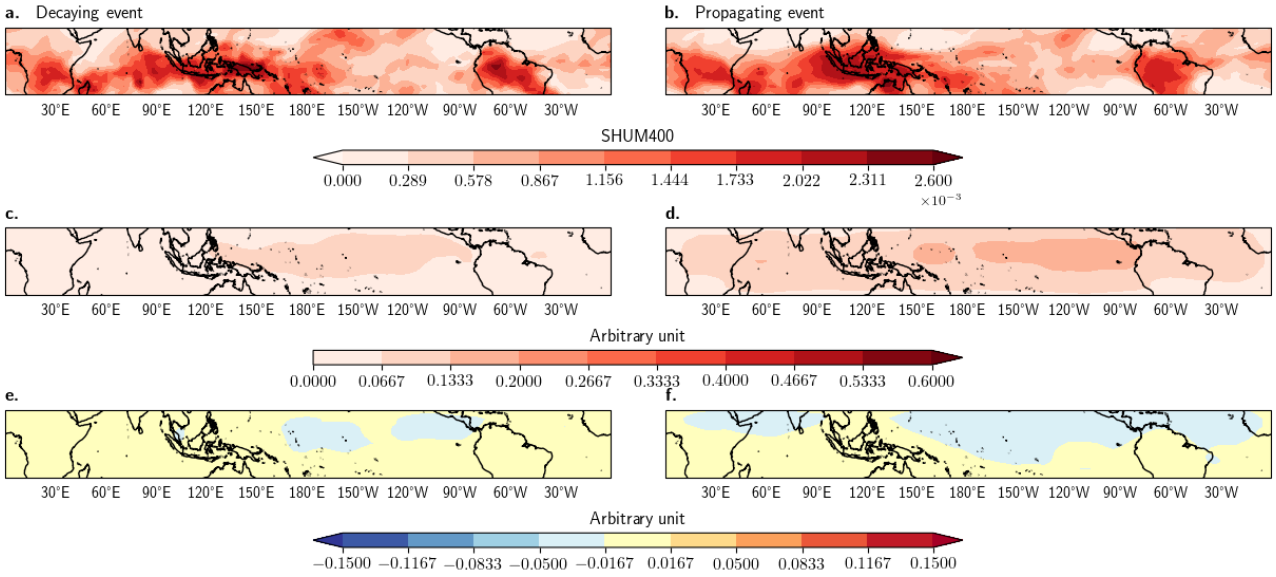


Figure S5. Interpretation of the CNN mean forecasts. (a–b) Composite maps of phase-3 SHUM400 for an MJO event which (a) decays and (b) propagates over the Maritime Continent. (c–d) PatternNet **RMM2** signal mean maps (signal maps averaged over all variables) corresponding to ten-day CNN forecasts for the decaying and propagating event respectively. (e–f) **RMM2** signal anomalies in SHUM400 for the decaying and propagating events respectively. The signal anomalies show a greater focus over the Maritime Continent region for this input variable.

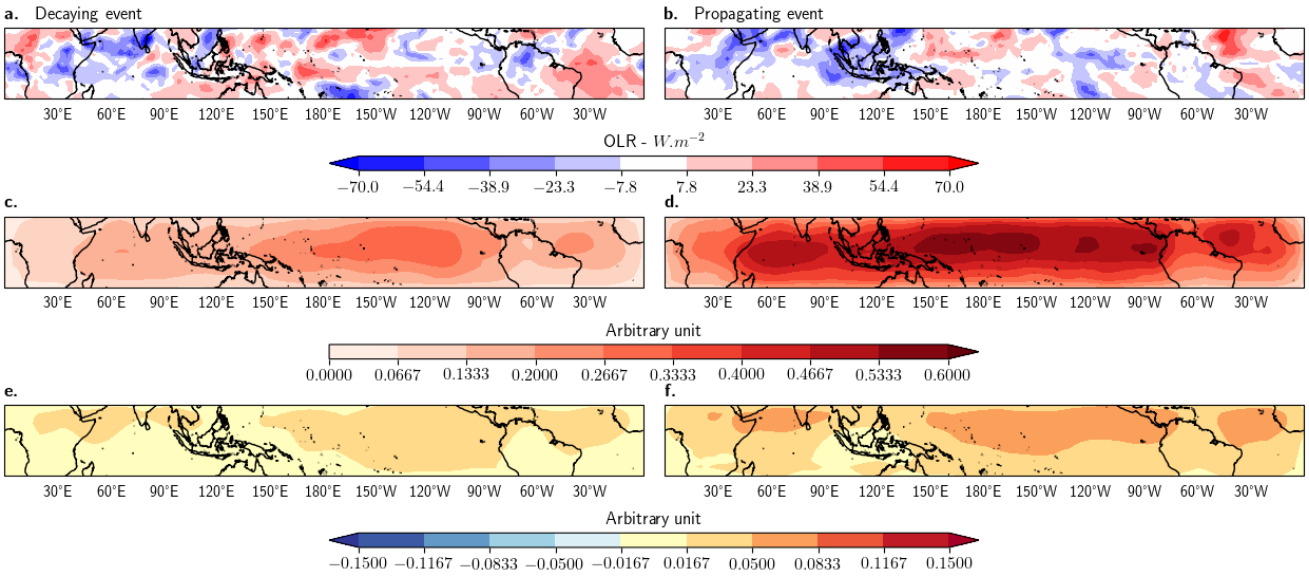


Figure S6. Interpretation of the CNN mean forecasts. (a–b) Composite maps of phase-3 OLR for an MJO event which (a) decays and (b) propagates over the Maritime Continent. (c–d) PatternNet **RMM1** signal mean maps (signal maps averaged over all variables) corresponding to ten-day CNN forecasts for the decaying and propagating event respectively. (e–f) **RMM1** signal anomalies in OLR for the decaying and propagating events respectively. The signal anomalies show a greater focus over the Maritime Continent region for this input variable.

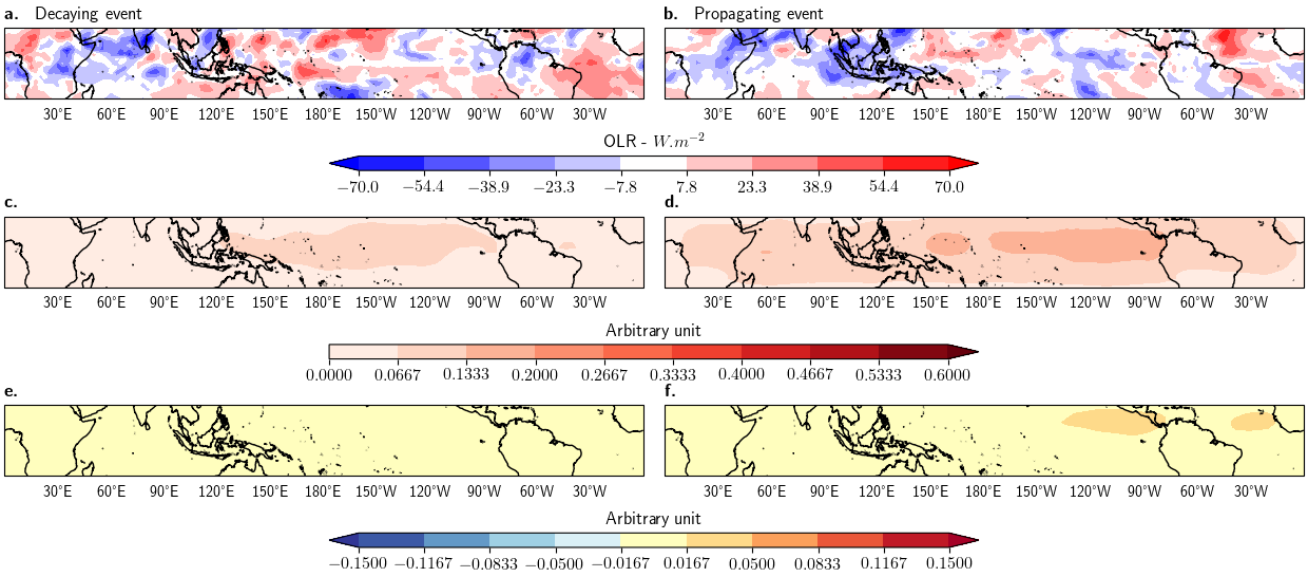


Figure S7. Interpretation of the CNN mean forecasts. (a–b) Composite maps of phase-3 OLR for an MJO event which (a) decays and (b) propagates over the Maritime Continent. (c–d) PatternNet **RMM2** signal mean maps (signal maps averaged over all variables) corresponding to ten-day CNN forecasts for the decaying and propagating event respectively. (e–f) **RMM2** signal anomalies in OLR for the decaying and propagating events respectively. The signal anomalies show a greater focus over the Maritime Continent region for this input variable.

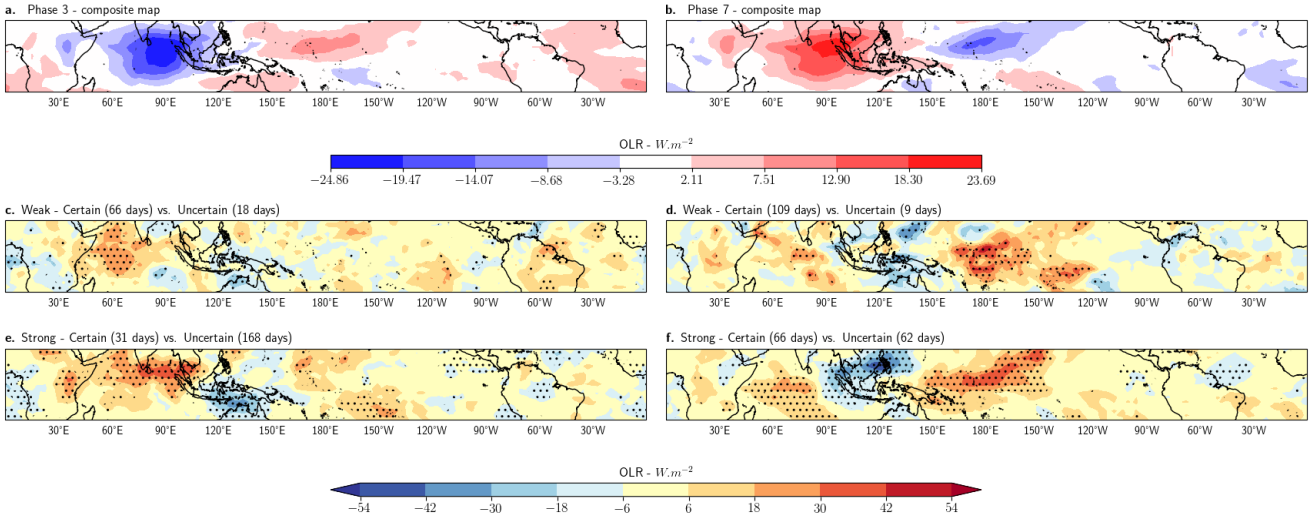


Figure S8. OLR uncertainty interpretation of the CNN MJO forecasts. **a.** and **b.** Composite maps of OLR in initial phases 3 and 7 for day-10 forecasts. Maps have been rescaled using MinMax scaling at each grid point before being fed to the CNN. **c. to f.** Anomalies maps between Weak (Strong) Predictable minus Weak (Strong) Unpredictable events. Weak events have an amplitude below (above) 1.0. Predictable (Unpredictable) events have RMM1 and RMM2 aleatoric uncertainties both inferior (superior) to their 30% (70%) percentiles. Stippling denotes areas where anomalies are significant at the 95% level using the Student's t-test.

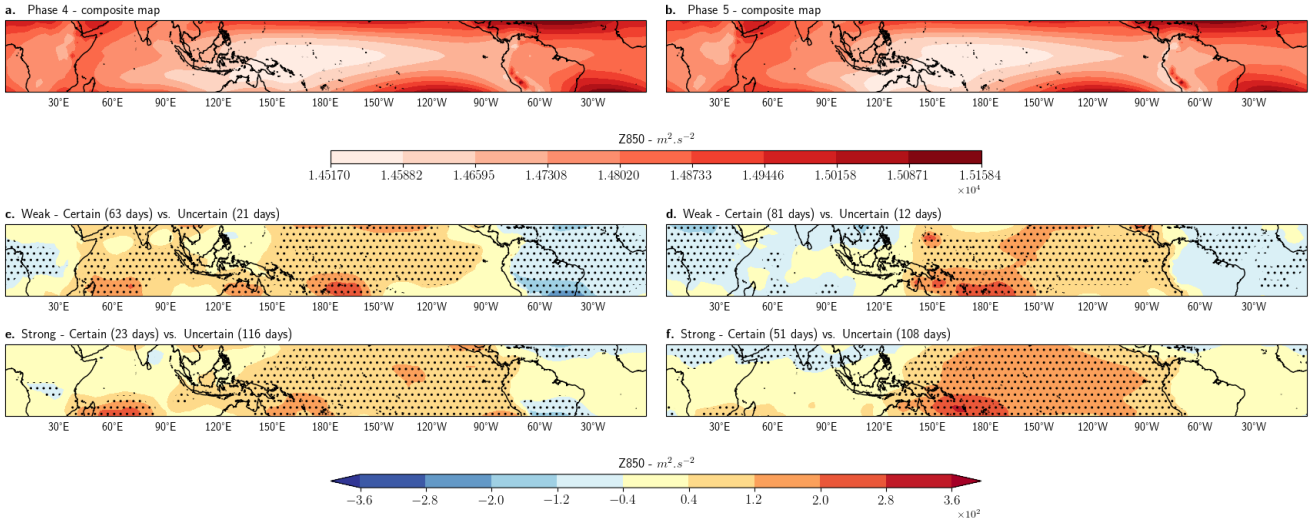


Figure S9. Z850 uncertainty interpretation of the CNN MJO forecasts. **a.** and **b.** Composite maps of geopotential at 850hPa (Z850) in initial phases 4 and 5 for day-10 forecasts. Maps have been rescaled using MinMax scaling at each grid point before being fed to the CNN. **c. to f.** Anomalies maps between Weak (Strong) Predictable minus Weak (Strong) Unpredictable events. Weak events have an amplitude below (above) 1.0. Predictable (Unpredictable) events have RMM1 and RMM2 aleatoric uncertainties both inferior (superior) to their 30% (70%) percentiles. Stippling denotes areas where anomalies are significant at the 95% level using the Student's t-test.

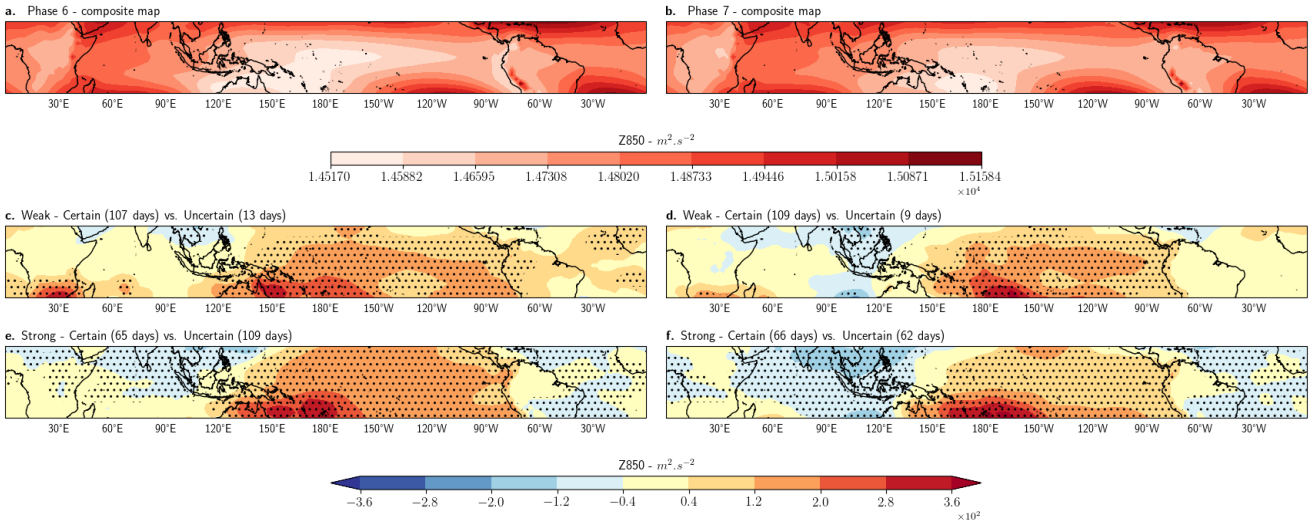


Figure S10. Z850 uncertainty interpretation of the CNN MJO forecasts. **a.** and **b.** Composite maps of geopotential at 850hPa (Z850) in initial phases 6 and 7 for day-10 forecasts. Maps have been rescaled using MinMax scaling at each grid point before being fed to the CNN. **c.** to **f.** Anomalies maps between Weak (Strong) Predictable minus Weak (Strong) Unpredictable events. Weak events have an amplitude below (above) 1.0. Predictable (Unpredictable) events have RMM1 and RMM2 aleatoric uncertainties both inferior (superior) to their 30% (70%) percentiles. Stippling denotes areas where anomalies are significant at the 95% level using the Student's t-test.

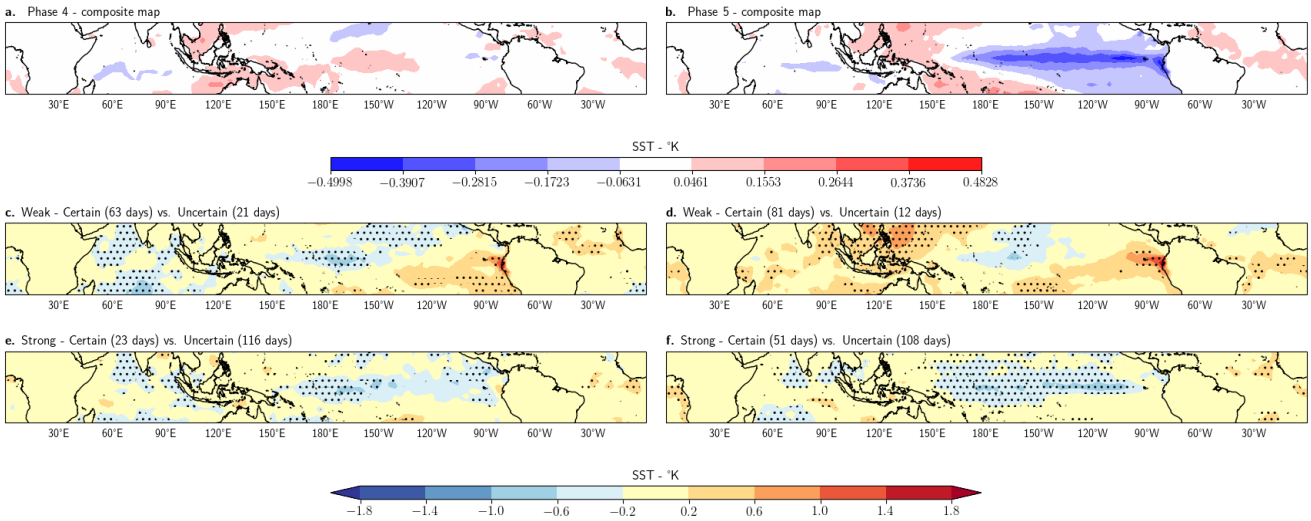


Figure S11. SST uncertainty interpretation of the CNN MJO forecasts. **a.** and **b.** Composite maps of Sea Surface Temperatures (SST) in initial phases 4 and 5 for day-10 forecasts. Maps have been rescaled using MinMax scaling at each grid point before being fed to the CNN. **c. to f.** Anomalies maps between Weak (Strong) Predictable minus Weak (Strong) Unpredictable events. Weak events have an amplitude below (above) 1.0. Predictable (Unpredictable) events have RMM1 and RMM2 aleatoric uncertainties both inferior (superior) to their 30% (70%) percentiles. Stippling denotes areas where anomalies are significant at the 95% level using the Student's t-test.

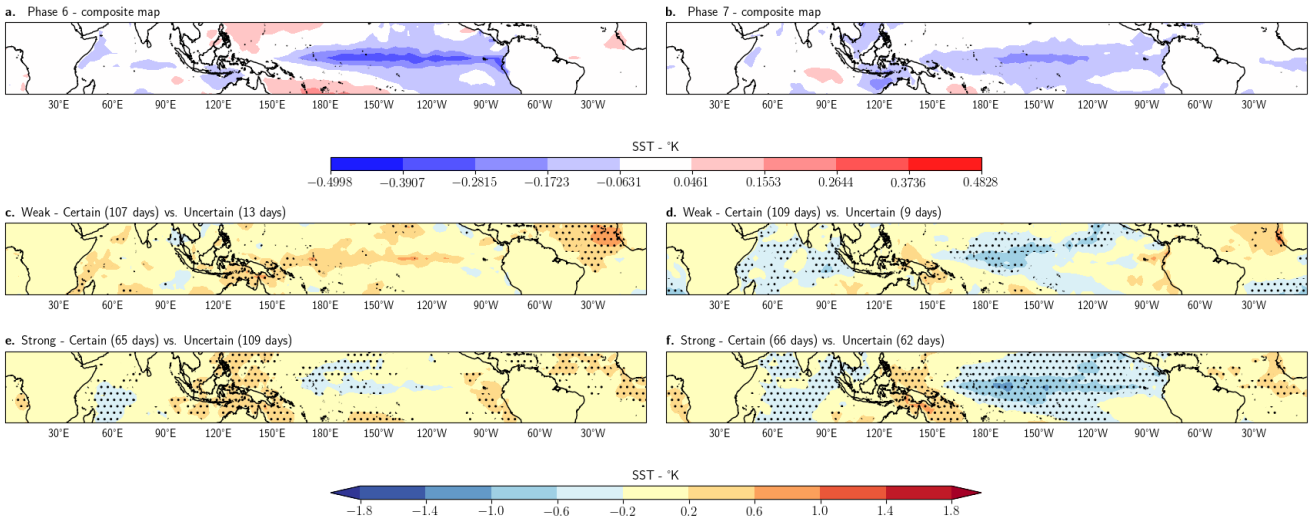


Figure S12. SST uncertainty interpretation of the CNN MJO forecasts. **a.** and **b.** Composite maps of Sea Surface Temperatures anomalies (SST) in initial phases 6 and 7 for day-10 forecasts. Maps have been rescaled using MinMax scaling at each grid point before being fed to the CNN. **c. to f.** Anomalies maps between Weak (Strong) Predictable minus Weak (Strong) Unpredictable events. Weak events have an amplitude below (above) 1.0. Predictable (Unpredictable) events have RMM1 and RMM2 aleatoric uncertainties both inferior (superior) to their 30% (70%) percentiles. Stippling denotes areas where anomalies are significant at the 95% level using the Student's t-test.

Table S1. Description of the dynamical forecast models used for comparison.[a]

Model	Time Range	Reforecast frequency	Ensemble size	Model year
ECMWF	02/01/2000 - 30/11/2019	Twice weekly	11	2020
CNRM	07/01/1993 - 28/12/2017	Weekly	10	2019
BOM	01/01/1982 - 26/12/2013	Twice weekly	33	2014
HMCR	02/01/1985 - 31/12/2010	Weekly	10	2020
CNN (Train)	01/05/1979 - 18/10/2011	Daily	10	2021
CNN (Test)	19/10/2011 - 30/11/2019	Daily	10	2021

[a] The most recent model version were selected according to their availability. The reforecasts are available at `ftp://s2sidx:s2sidx@acquisition.ecmwf.int/RMMS/`

Table S2. All initial phases

	Certain		Uncertain		
	Strong at t	Weak at t	Strong at t	Weak at t	Total
Strong at $t+10$	142	213	707	74	1136
Weak at $t+10$	193	418	117	38	766
Total	335	631	824	112	1902

Table S3. Initial Phase 3

	Certain		Uncertain		
	Strong at t	Weak at t	Strong at t	Weak at t	Total
Strong at $t+10$	16	24	135	13	188
Weak at $t+10$	15	42	33	5	95
Total	31	66	168	18	283

Table S4. Initial Phase 7

	Certain		Uncertain		
	Strong at t	Weak at t	Strong at t	Weak at t	Total
Strong at $t+10$	26	43	56	4	129
Weak at $t+10$	40	66	6	5	117
Total	66	109	62	9	246

References

F. Vitart et al. (2017, January). The Subseasonal to Seasonal (S2S) Prediction Project Database. *Bulletin of the American Meteorological Society*, *98*(1), 163–173. (`ftp://s2sidx:s2sidx@acquisition.ecmwf.int/RMMS`)

Gal, Y. (2016). *Uncertainty in Deep Learning* (Unpublished doctoral dissertation). Cambridge University.

Ghiasi, G., Lin, T.-Y., & Le, Q. V. (2018, October). DropBlock: A regularization method for convolutional networks. *arXiv:1810.12890 [cs]*. (arXiv: 1810.12890)

H. Hersbach et al. (2018a). ERA5 hourly data on pressure levels from 1979 to present. *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*. (Accessed on 01-03-2021, https://doi.org/10.24381/cds.bd0915c6)

H. Hersbach et al. (2018b). ERA5 hourly data on single levels from 1979 to present. *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*. (Accessed on 01-03-2021, https://doi.org/10.24381/cds.adbb2d47)

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, *15*(5), 559–570. doi: 10.1175/1520-0434(2000)015⟨0559:DOTCRP⟩2.0.CO;2

Kim, H., Vitart, F., & Waliser, D. E. (2018, December). Prediction of the Madden–Julian Oscillation: A Review. *Journal of Climate*, *31*(23), 9425–9443. Retrieved from https://journals.ametsoc.org/doi/10.1175/JCLI-D-18-0210.1 doi: 10.1175/JCLI-D-18-0210.1

Kindermans, P.-J., Schütt, K. T., Alber, M., Müller, K.-R., Erhan, D., Kim, B., & Dähne, S. (2017, October). Learning how to explain neural networks: PatternNet and PatternAttribution. *arXiv:1705.05598 [cs, stat]*. Retrieved 2021-05-18, from http://arxiv.org/abs/1705.05598 (arXiv: 1705.05598)

Leutbecher, M., & Palmer, T. (2008, March). Ensemble forecasting. *Journal of Computational Physics*, *227*(7), 3515–3539. Retrieved 2021-05-27, from https://linkinghub.elsevier.com/retrieve/pii/S0021999107000812 doi: 10.1016/j.jcp.2007.02.014

Lim, Y., Son, S.-W., & Kim, D. (2018, May). MJO Prediction Skill of the Subseasonal-to-Seasonal Prediction Models. *Journal of Climate*, *31*(10), 4075–4094. Retrieved 2021-05-05, from http://journals.ametsoc.org/doi/10.1175/JCLI-D-17-0545.1 doi: 10.1175/JCLI-D-17-0545.1

Marshall, A. G., Hendon, H. H., & Hudson, D. (2016). Visualizing and verifying probabilistic forecasts of the Madden-Julian Oscillation. *Geophysical Research Letters*, *43*(23), 12,278–12,286. doi: 10.1002/2016GL071423

Roulston, M. S., & Smith, L. A. (2002). Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, *130*(6), 1653–1660. doi: 10.1175/1520-0493(2002)130⟨1653:EPFUIT⟩2.0.CO;2

Scalia, G., Grambow, C. A., Pernici, B., Li, Y.-P., & Green, W. H. (2019, October). Evaluating Scalable Uncertainty Estimation Methods for DNN-Based Molecular Property Prediction. *arXiv:1910.03127 [cs, stat]*. (arXiv: 1910.03127)

Translational Neurotechnology Lab. (2019). PatternNet GitHub Repository. (University of Freiburg, https://github.com/TNTLFreiburg/pytorch_patternnet)

Wheeler, M. C., & Hendon, H. H. (2004). An All-Season Real-Time Multivariate MJO Index: Development of an Index for Monitoring and Prediction. *Monthly Weather Review*, *132*, 16.