Journal of Advances in Modeling Earth Systems

Supporting Information for

Using simple, explainable neural networks to predict the Madden-Julian oscillation

Zane K. Martin[1]

Elizabeth A. Barnes[1]

Eric Maloney[1]

1 Department of Atmospheric Science, Colorado State University, Fort Collins, CO

Contents of this file:

1. Text S1 - Sensitivity Tests
2. Figures S1-S5

Text S1. Sensitivity Tests

In developing both the regression and classification ANN architectures, we conducted many tests exploring sensitivity to the processing of the input data, the ANN architectures, and the nature of model output. We show results from some of these sensitivity tests in Figures S1-S5 below: here we provide additional methodological detail regarding those tests. All tests below are shown for models during winter and, unless noted, for models that input OLR and zonal wind at 850 and 200 hPa.

In Figure S1, we compare the regression and classification ANN skill in a model trained using all-year data evaluated over winter and summer periods, versus the models trained in winter and summer respectively. While changes are modest, we found season-specific training to be somewhat advantageous in improving skill.

In Figure S2, we show the sensitivity to a change in how the regression ANN is trained. Rather than training the regression ANN on all winter days, we instead train the model on all active MJO days and a random subset of inactive MJO days such that weak MJO days are 1/9 of the overall training datasets. This is analogous to how the classification model is trained (see Section 3.1.1), and provides the regression ANN with more strong MJO samples at all lead times. While it marginally improves the accuracy of the regression model when active MJO days are considered (Fig. S2), it does not have a large change on the overall accuracy or the BCC; the regression model still shows poor performance forecasting active MJO events at leads longer than a few days.

In Figure S3 we show accuracy over active MJO days from the classification ANN at lead times of 0, 2, 5, 10, 15, and 20 days from a range of sensitivity tests. For the "control" test, the model is the same as that discussed in Section 4.1, with the range across 10-member ANN

ensemble shown to capture spread due to random initial ANN weights. For the sensitivity tests only one ANN was trained at each lead time.

The first set of sensitivity tests shown in Figure S3 are slight changes to the ANN architecture. For the "high_ridge" test the ridge regression penalty was increased from 0.25 to 1, and for the "low_ridge" test the penalty was decreased from 0.25 to 0.1. For the "wide_net" test the number of nodes in the hidden layer was increased from 16 to 64, and in the "deep_net" the single, 16 node single layer was replaced with 2 fully connected layers of 16 nodes each. Note across these tests, large changes relative to the control are not observed, and typically fall within the control spread (Fig. S3a).

The second set of tests explore changes to the model input. The "30NS" and "15NS" experiments alter the latitude bands over which the input data is retained. The "lat_avg" model takes the 15N-15S average of the input before feeding it into the neural network, such that the input is a function only of longitude (e.g. a vector of length 144 per variable). Further, in the "lat_avg" model the learning rate is increased to 0.001 from the 0.0005 value used in the control. The "prior_days" test includes not only the variables from forecast day 0 in the input, but also includes forecast day -5, doubling the size of the input vector.

Model performance in all of these tests lies within the range of the control, with the exception of the latitudinal averaging at lead times of less than 5 days, which shows notably higher accuracy. Because the RMM index takes 15N-15S averaged variables as input, this increase in accuracy at short leads is likely due to the fact that the input is more closely associated with the output (i.e. how the RMM is computed), making it easier for the ANN to learn the relationship between the latitudinally-averaged input and the RMM phase. The fact that this increase relative to the control fades at longer lead times suggests, consistent with the discussion in Section 4.2,

69 that identifying the MJO at short leads is a different task than predicting MJO behavior. Because
70 the improvement is only seen at short leads, and because we are interested in how the 2-D structure
71 of input variables informs the ANN (e.g. for in the LRP plot in Figures 9 and 10), we prioritize the
72 2-D input approach. Many additional sensitivity tests were performed during model development,
73 and similar tests were performed for the regression model, but for brevity are not shown here, as
74 results are comparable to those discussed.

75 A third sensitivity test, shown in Figure S4, quantifies sensitivity to training the regression
76 ANN using a longer training dataset than NOAA OLR and ERA5 data allow. For this, we use ERA
77 20th century reanalysis daily OLR and zonal wind at 850 and 200 hPa data (Poli et al. 2016), which
78 we obtained over the full period of availability from January 1, 1901 to October 31, 2010. ERA-
79 20C input data is processed identically to the input for ERA-5 described in Section 2.1. The RMM
80 index is calculated from ERA-20C using the method described in Wheeler and Hendon (2004);
81 over the period in which the ERA-20C data overlaps with the observed RMM index, we found the
82 correlation between our calculated ERA-20C RMM1/2 and the observed RMM1/2 values to be
83 approximately .89, indicating good agreement in how the RMM index is formed.

84 We train a regression ANN with an architecture identical to that discussed in Section 3.1.1
85 but using ERA-20C data instead of ERA-5 data. The validation period is January 1, 2001 to
86 October 31, 2010. We explored varying the training dataset to see whether model performance
87 improved if much more training data was included. For lead times of 0, 5, 10, and 15 days we
88 trained separate models for 11 different training periods. All training periods end December 31,
89 1999, but start dates vary across June 1 of: 1994, 1989, 1984, 1979, 1974, 1969, 1959, 1949, 1929,
90 1909, and 1901. To facilitate comparison to ERA5, we trained an additional model on

NOAA/ERA5 data from June 1979 to December 1999, and validated on NOAA/ERA5 data from January 1, 2001 to October 31, 2010.

Results in Figure S4 show generally comparable performance between ERA5 and ERA20C when the same period is used for validation and training. For reasons that we did not explore in depth, the ERA20C model shows higher BCC values at 0 and 5 days than NOAA/ERA5, comparable performance at 10 days, and worse performance at 15 days. More importantly, Figure S4 indicates that training the simple ANN on ERA20C with significantly more data does not lead to substantial improvement in the BCC at any lead time after between 120 and 200 months. Further tests with wider or deeper ANNs using the full 1901-1999 period of training also did not show improved performance.

The final sensitivity test, shown in Figure S5, explores sensitivity to including four or more additional input variables, following the same procedure as described in the manuscript in Section 4.2. Legend conventions in Figure S5 follow Figure 8, except “d” denotes divergence. Overall, no substantial increase in skill is seen in models with four or more variables.

Supplemental Figures.

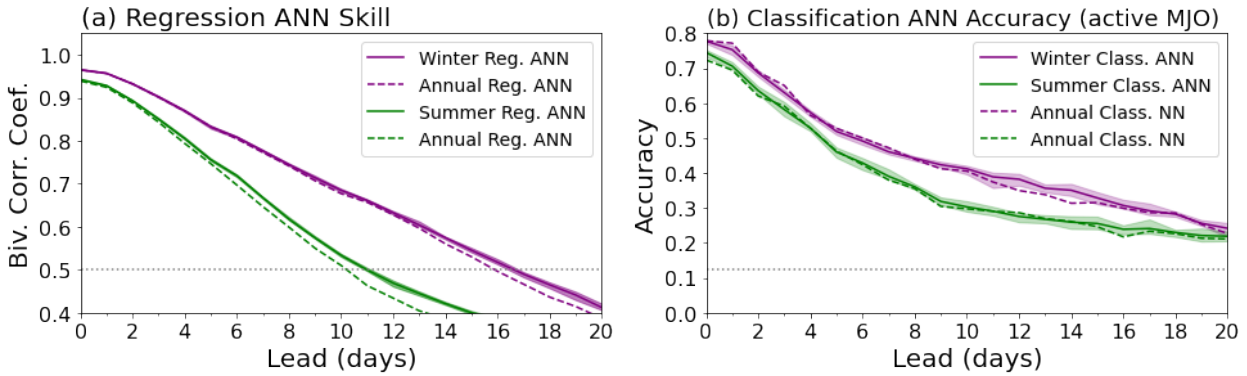


Figure S1 Regression ANN (panel a) and classification ANN (panel b) performance, similar to Figures 4 and 6, for ANNs trained specifically on winter and summer seasons (solid lines) versus a model trained on all seasons and evaluated separately in summer and winter (dashed line). The shading shows the seasonal model range across 10 ensemble models; for the annual model only one ensemble is considered.

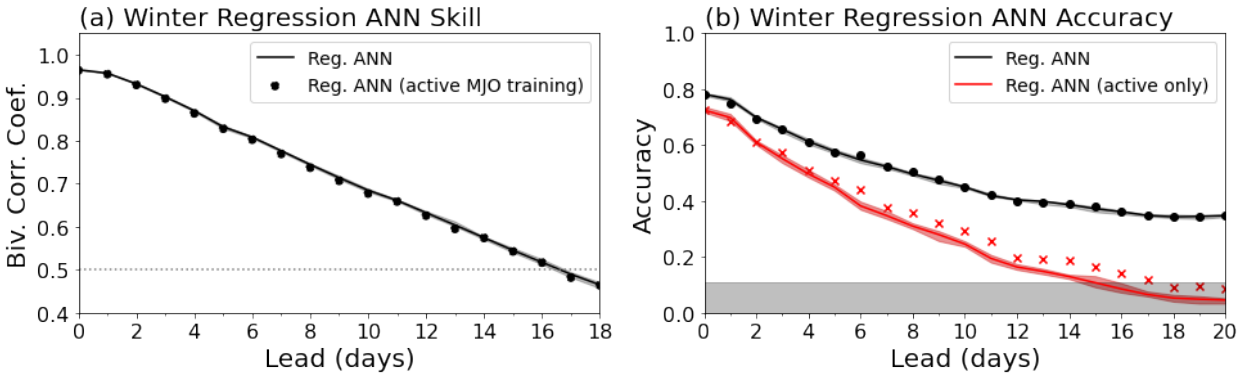
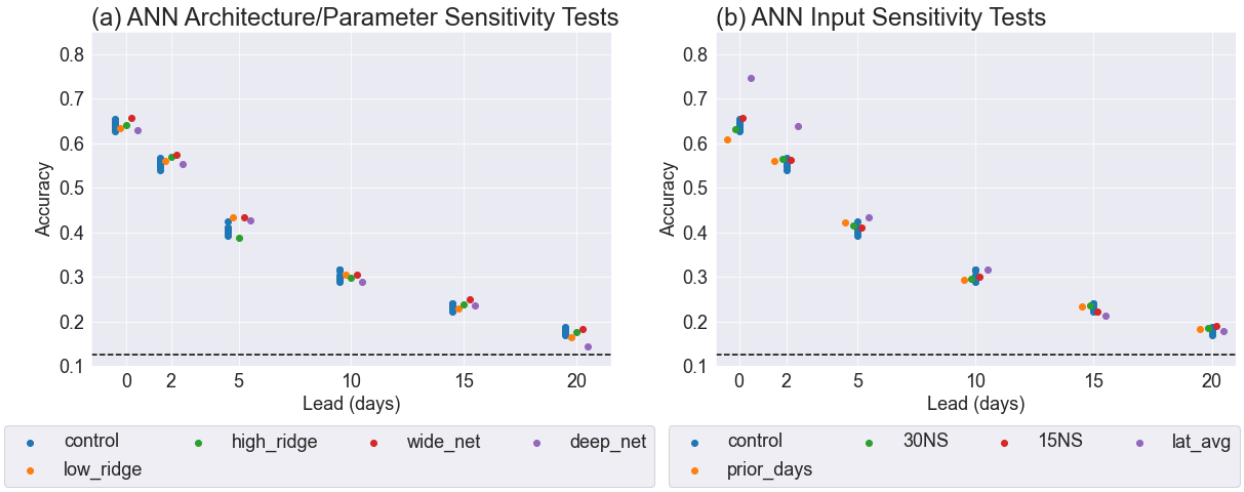


Figure S2 Winter regression ANN skill (panel a as in Figure 4) or accuracy of MJO phase (right; similar to Figure 6) for regression ANNs trained on all MJO days (lines) versus ANNs trained using fewer weak MJO days (as described in Supplemental Text S1; dots or x's). In the panel (b), black curves/dots are regression model accuracy evaluated over all MJO days, and red curves/x's are regression model accuracy evaluated only for active MJO days. Note the poor performance for active days, caused by the inability of the regression model to predict strong amplitude events.



126

127 **Figure S3** Model accuracy over active MJO days for sensitivity tests varying the architecture or

128 hyperparameters of the ANN (panel a) and varying the model input (panel b). Tests are indicated

129 by the legend as described in Supplemental Text S1. For the control, a 5 ANN ensemble was used.

130

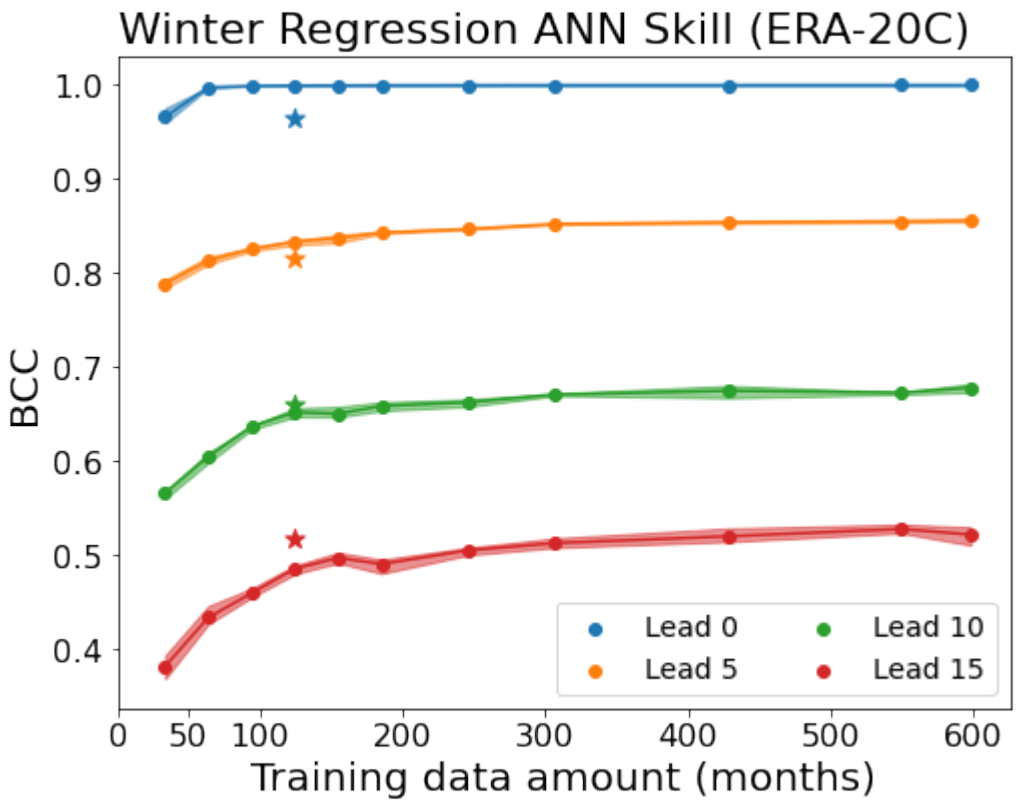


Figure S4 Winter regression ANN skill at lead times of 0, 5, 10, and 15 days (colors), trained using ERA-20C data. Models are trained using larger amounts of training data (dots; see Supplemental Text S1), and the *x*-axis shows the number of months in the training data period. Shading shows the range across 5 ANNs with different random starting weights. Stars show results using ERA5/NOAA winds and OLR data, as described in Supplemental Text S1.

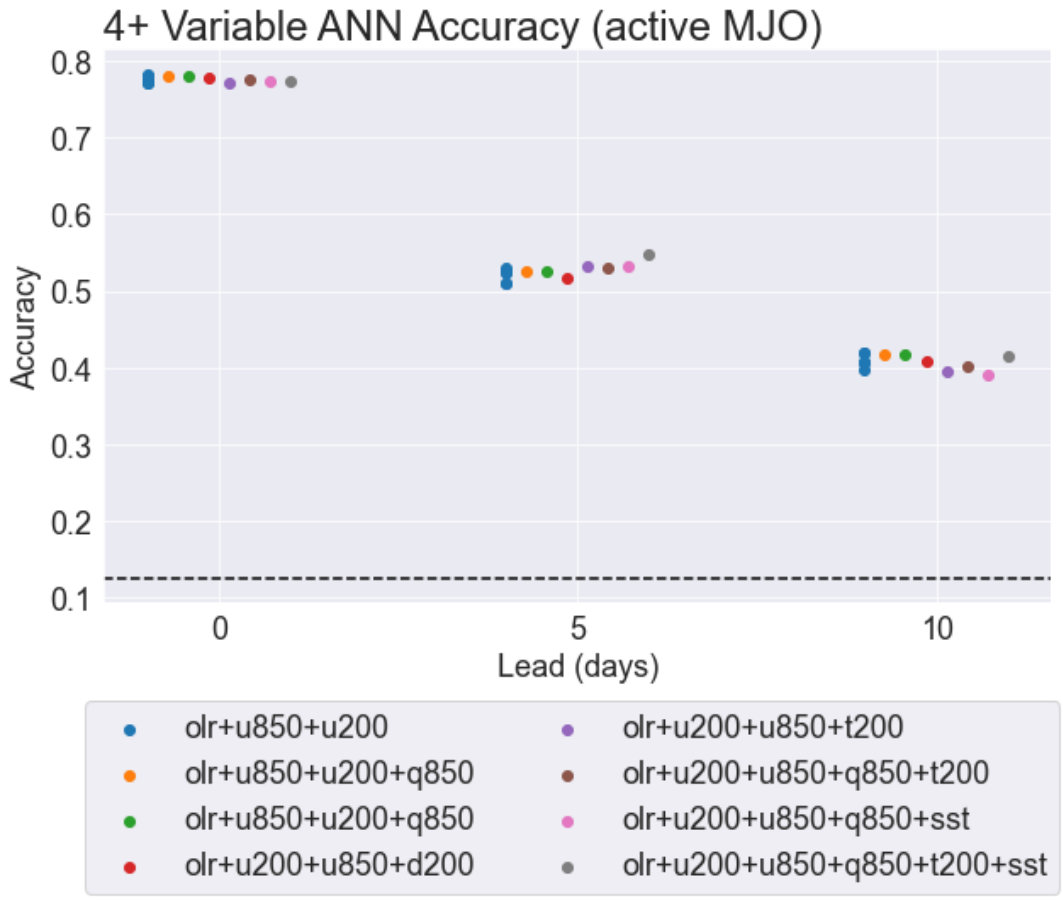


Figure S5 Similar to Figures 8 and S3, but for a series of tests with 4, 5, or 6 input variables. The blue "olr+u850+200" model is the same as in Figure 8c; other models have only 1 ensemble member.