Using Neural Networks for Data Cleaning in Weather Datasets

Jack R. P. Hanslope [1]   **Laurence Aitchison** [1]

Abstract

In climate science, we often want to compare across different datasets. Difficulties can arise in doing this due to inevitable mismatches that arise between observational and reanalysis data, or even between different reanalyses. This misalignment can raise problems for any work that seeks to make inferences about one dataset from another. We considered tropical cyclone location as an example task with one dataset providing atmospheric conditions (ERA5) and another providing storm tracks (IBTrACS). We found that while the examples often aligned well, there were a considerable proportion (around 25%) which were not well aligned. We trained a neural network to map from the wind field to the storm location; in this setting misalignment in the datasets appears as "label noise" (i.e. the labelled storm location does not correspond to the underlying wind field). We found that this neural network trained only on the often noisy labels from IBTrACS had a denoising effect, and performed better than the IBTrACS labels themselves, as measured by human preferences. Remarkably, this even held true for training points, on which we might have expected the network to overfit to the IBTrACS predictions.

1. Introduction

In the field of climate science, there are many different datasets from many different sources. There are often times when we would like to compare across different datasets. For example, we might want to predict tropical cyclone tracks (Richman & Leslie, 2012; Camargo & Wing, 2016; Camargo et al., 2019), flooding (Mosavi et al., 2018; Park & Lee, 2020; Motta et al., 2021) or the power output of a renewable power station (Foley et al., 2012; Kumar &

[1]Department of Computer Science, University of Bristol, Bristol, UK. Correspondence to: Jack Hanslope <jack.hanslope@bristol.ac.uk>.

Kalavathi, 2018; Sharadga et al., 2020) from atmospheric condition data.

We might, when performing these comparisons, encounter discrepancies between datasets (Greatbatch & Rong, 2006; Li et al., 2022). These discrepancies can arise for a number of reasons. One reason is that some datasets are reanalyses (ECMWF, 2023). In reanalysis the goal is often to produce an accurate, globally consistent representation of the atmospheric conditions. However, as we cannot directly measure the atmospheric conditions, we must use computational modelling to estimate them from indirect observations. This estimation procedure can introduce mismatches to direct observational data (e.g. the output of a renewable energy plant), or between different reanalyses. These discrepancies can make it is difficult to draw connections between variables in different datasets, which causes significant problems if the two variables of interest in our scientific analysis are only available in different datasets.

We studied how to mitigate these issues of mismatch between different datasets. As an example, we studied tropical cyclones. We had two reanalyses, one providing the location of tropical cyclones (IBTrACS; Knapp et al., 2010; 2018) and one providing the atmospheric conditions (ERA5; Hersbach et al., 2018; 2020). Initially, we were interested in predicting the future location of a tropical cyclone, when provided with the wind field in the present. However, preliminary analyses indicated that while the track data and the wind field often aligned well, there were many occasions where this was not the case, even when comparing the wind field and tropical cyclone location at the same time. There were a number of ways in which this misalignment appeared. In some instances, ERA5 would show very obvious signs of a tropical cyclone in one location while IBTrACS would indicate a tropical cyclone in a different location. Sometimes, there would appear to be more storms within the wind field than the number indicated by the storm location data. If we consider the ERA5 wind field as the input and the IBTrACS storm location as the output, then we can consider this to be an issue of having noisy labels.

Given that there is this misalignment between the IBTrACS storm locations and the ERA5 wind field, it may seem that the only solution would be to hand-label the storm locations, based on the ERA5 wind fields. We considered how-

ever, whether we could leverage the fact that, in around 75% of instances, there was a very close match between the two datasets. In other words, instances where the IB-TrACS tropical cyclone location would match the ERA5 wind field. We therefore took a naive approach: training a neural network to predict the IBTrACS storm location from the ERA5 wind field using a U-Net architecture (Ronneberger et al., 2015). Surprisingly, despite being trained on often noisy/mismatched data, we found that the neural network almost always matched human judgements of storm location based on the ERA5 wind field. Indeed, even in instances where the two datasets were misaligned, the neural network usually gave a much better predication of the location than the tropical cyclone location dataset.

2. Related Work

There has been much research in the field of noisy label machine learning (Natarajan et al., 2013; Han et al., 2019; Song et al., 2023). However, none of this work has studied whether neural networks are able to address the issues raised by the noisy labels that are inevitably observed in weather datasets.

3. Methodology

3.1. Data Description

3.1.1. ERA5

For this work, we need a dataset providing the atmospheric conditions. Such data includes variables like pressure, temperature, wind speed and direction, and atmospheric pressure. We use ERA5 to provide this data (Hersbach et al., 2018; 2020). ERA5 is provided by the Copernicus Climate Change Service (C3S) at the European Centre for Medium-Range Weather Forecasts (ECMWF).

It is fairly obvious that it is impossible to obtain measurements of every variable we might be interested in at every location (latitude, longitude, altitude) within the Earth's atmosphere. As such, ERA5 is a reanalysis dataset. A reanalysis first takes a weather forecasting model and uses it to create a prediction for a certain weather variable (for example, temperature) based on the value of that variable at some earlier time. It then corrects this prediction using known, but indirect, observations; this step is called data assimilation (Dee et al., 2011). A reanalysis provides a globally consistent picture of the atmospheric conditions at any given time, but it is ultimately an estimate based on indirect measurements and modelling.

The ERA5 reanalysis provides many different variables which are either provided at a single level or at different pressure levels. The different pressure levels are at 37 pressure levels from $1000\,\mathrm{hPa}$ to $1\,\mathrm{hPa}$ (hectopascals). Pressure varies inversely with altitude from sea level, that is, as altitude increases, pressure decreases. Most of the ERA5 variables are provided hourly, at a $31\,\mathrm{km}$ resolution.

Using all of the variables from the dataset would require an enormous amount of storage, so instead we use only a selection. We use wind data because this is, intuitively, the most indicative of a hurricane. We have u (eastward) and v (northward) components of the wind. We use the wind at $850\,\mathrm{hPa}$ which corresponds with an altitude of approximately $1500\,\mathrm{m}$.

It is worth noting that the aim of this work is not to produce the most accurate storm location model possible. If that were the goal, then we would almost certainly be able to achieve a higher accuracy if we used more variables. The goal of this work is to explore whether we are able to use a neural network to mitigate the mismatch between different reanalysis datasets.

3.1.2. IBTRACS

In addition to a dataset providing us with the atmospheric conditions, we also require a dataset providing us with locations of tropical cyclones. The dataset we selected is the International Best Track Archive for Climate Stewardship (IBTrACS; Knapp et al., 2010; 2018).

This dataset combines best track datasets from many different national agencies around the world. An agency's best track dataset provides the best location estimate for each point in the storm's lifetime, as well as storm intensity at each of those points.

The IBTrACS dataset provides many variables but the only information we use from this dataset is the latitude and longitude of the tropical cyclone centre. This information is usually provided at 3 hour intervals, with some additional timestamps (for example, when a tropical cyclone makes landfall).

3.2. Neural Network Setup

The neural network takes as input the ERA5 u and v channels of $850\,\mathrm{hPa}$ wind speeds at one timestamp. The network output is an estimate of the location of the tropical cyclone, and we train the network to match the locations given by IBTrACS. While we could return this estimate as a continuous latitude and longitude value (i.e. using regression), we instead ask the neural network to classify the grid box that the tropical cyclone is located within, as this allows the network to naturally return an estimate of the probability that the tropical cyclone is within a given grid box.

We restrict the data to only the North Indian basin and only instances where there is exactly one storm in the basin. The basin is constrained by the equator to the south (all latitudes

are greater than 0) and longitudes are between 30° and 100° east. Only tropical storms since 1980 were used, since prior to then, identification of storms was less reliable. Since then, satellites have been used to locate storm centres. We discard any time steps that do not fall on a whole number of 3 hours. In order to reduce the storage requirements of the data, we coarsen the data to the nearest degree (coarsening by a factor of about 4).

3.3. Neural Network Architecture

The model we trained for storm location is a U-Net (Ronneberger et al., 2015). These typically perform well on tasks with image-like outputs; our network outputs an "image" giving the probability of the tropical cyclone being within each grid box.

U-Nets have an encoder-decoder structure and the encoder half of our network has 4 downsampling layers. Each has two 3×3 convolutions followed by ReLUs and 2×2 max pooling. Note that the final layer has no max pooling. The decoder layers consist of interpolation upsampling concatenated to the output of the corresponding layer in the encoder (skip connections) and two 3×3 convolutions, each followed by a ReLU. The number of filters in the encoder layers are 64, 128, 256 and 512; similarly, the decoders have 512, 256, 128 and 64 filters.

We use the U-Net implementation provided in the GitHub repository attached to Wolny et al. (2020).

3.4. Training Setup

The batch size was 32 and we trained for 100 epochs. We used Adam (Kingma & Ba, 2014) with a learning rate of 10^{-3} on a cross entropy loss. We split the data into training, validation and test with 15% each for the test and validation sets and 70% for the training set. At each learning step, the network would output a logit for each of the $32 \times 56 = 1792$ grid boxes, describing the probability of the tropical cyclone being within that grid box. We calculate the cross entropy loss by comparing this to the target cell from IBTrACS.

Post-hoc, we apply temperature scaling (Guo et al., 2017). This method is intended to bring the confidences of the network closer to the true probabilities. It scales all of the network's logits by a single number, the temperature, which is tuned on the validation set. The relative order of the class scores is maintained.

4. Results

Now, we compare the original IBTrACS storm locations to the locations provided by our trained network. Overall, we find that when they suggest different locations, usually the neural network gives a better location.

4.1. Qualitative Results

Figure 1 shows four wind fields The wind direction and magnitude is shown by the direction and size of the arrows on each point; this data comes from ERA5. The location of the storm centre, provided by IBTrACS is shown by the orange cross. The probabilistic predictions of the model are shown as a colourmap, ranging from white (low probability) to dark blue (high probability).

The particular storms were chosen by hand to show a mismatch between IBTrACS and ERA5. In most of these instances, we see that the model will usually put most of its confidence in the middle of whichever cyclone has the strongest wind. Intuitively, this is where we, as humans, would expect the storm to be. In some of these wind fields, the ground truth label does not appear to be in the middle of a cyclone, and in some cases does not appear to be in a cyclone at all.

4.2. Quantitative Results

Of course, it is not possible to reach any definitive conclusions based on just those examples. At the same time, it is difficult to know how to quantitatively compare the neural network outputs with those of IBTrACS in the absence of any "ground truth" data. As such, to determine whether the locations provided by the U-Net are indeed better than the locations provided by IBTrACS we devised a test where one of the authors was shown a wind field with two storm locations. One of the locations was from the U-Net and the other from IBTrACS but it was not indicated which was which (i.e. the author was blinded). The author was asked which, if either, they preferred. An example is shown in Figure 2.

200 wind fields were shown from each of the train and test set, with the storms being chosen randomly (with a seed of 0). Note that there are many cases where the locations cocoincided, so no preference could be made. The results are summarised in Table 1, and clearly show an improvement for the U-Net. In particular, the human preferred the U-Net 46 out of 59 times where there was a preference for the test set, and 49 of 64 times for the training set.

Using the 59 and 64 instances where there was a preference, we conducted one-tailed binomial tests. The null hypothesis was that *when* there was a preference of one location, both locations are equally likely. The alternative hypothesis was that when there was a preference of one location, the U-Net was more likely to be preferred. The p-values are 9.6×10^{-6} and 1.2×10^{-5} which both strongly suggests that when there is a noticeable difference between the two locations, the U-Net is providing a better location. Note that it is remarkable that we see improvements even on the training set, as we might have expected the neural network to overfit
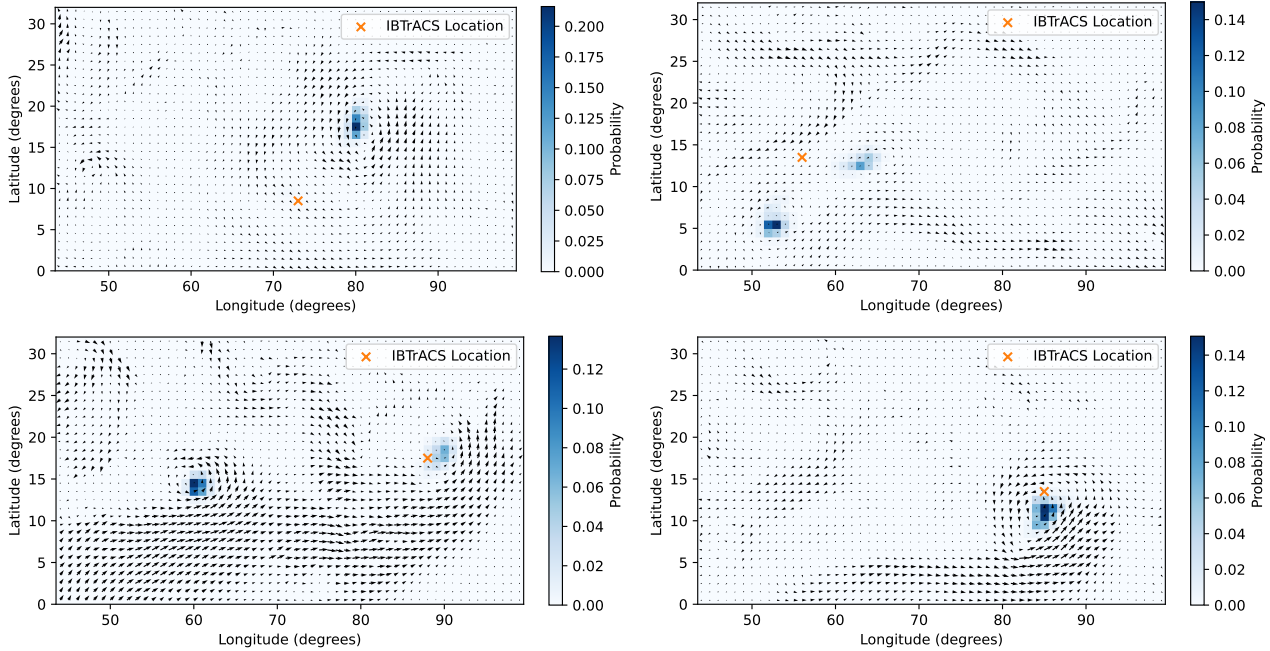
Figure 1. Wind fields for 4 hand chosen storms. IBTrACS location of storm centre marked by orange cross and probability predicted by the U-Net shown by the colourmap. Top left 1980-10-19 03:00; top right 1986-11-11 09:00; bottom left 1987-06-02 21:00; bottom right 1996-11-28 18:00.
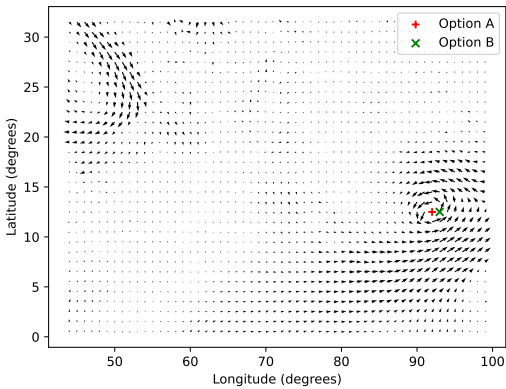

Figure 2. Example of a wind field shown to the author for comparison.

to the training set, and therefore to simply give the same location as the original IBTrACS location that the network was trained on.

5. Conclusion

In summary, we've shown that there are sometimes discrepancies between climate science datasets, and these discrepancies can often make it difficult to draw conclusions about a target variable in one dataset from data in the other dataset. We have shown that it is sometimes possible to mitigate

Table 1. Preferences of U-Net location, IBTrACS location and no preference for test set and train set. 200 images were shown for each set. The p-value is calculated as a one tailed binomial test using only the instances where one or other location was preferred.

PREFERENCE	TEST SET	TRAIN SET
U-NET	46	49
IBTRACS	13	15
NEITHER	141	136
TOTAL	200	200
p-VALUE	9.6×10^{-6}	1.2×10^{-5}

these mismatches by exploiting the denoising properties inherent in neural networks.

References

Camargo, S. J. and Wing, A. A. Tropical cyclones in climate models. *Wiley Interdisciplinary Reviews: Climate Change*, 7(2):211–237, 2016.

Camargo, S. J., Camp, J., Elsberry, R. L., Gregory, P. A., Klotzbach, P. J., Schreck III, C. J., Sobel, A. H., Ventrice, M. J., Vitart, F., Wang, Z., et al. Tropical cyclone prediction on subseasonal time-scales. *Tropical Cyclone Research and Review*, 8(3):150–165, 2019.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, d. P., et al. The era-interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the royal meteorological society*, 137(656):553–597, 2011.

ECMWF. Fact sheet: Reanalysis, 2023. URL https://www.ecmwf.int/en/about/media-centre/focus/2023/fact-sheet-reanalysis. Accessed: 2024-05-23.

Foley, A. M., Leahy, P. G., Marvuglia, A., and McKeogh, E. J. Current methods and advances in forecasting of wind power generation. *Renewable energy*, 37(1):1–8, 2012.

Greatbatch, R. J. and Rong, P.-p. Discrepancies between different northern hemisphere summer atmospheric data products. *Journal of Climate*, 19(7):1261–1273, 2006.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Han, J., Luo, P., and Wang, X. Deep self-learning from noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N. Era5 hourly data on pressure levels from 1940 to present., 2018. URL https://cds.climate.copernicus.eu/doi/10.24381/cds.bd0915c6.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Knapp, K. R., Kruk, M. C., Levinson, D. H., Diamond, H. J., and Neumann, C. J. The international best track archive for climate stewardship (ibtracs) unifying tropical cyclone data. *Bulletin of the American Meteorological Society*, 91(3):363–376, 2010.

Knapp, K. R., Diamond, H. J., Kossin, J. P., Kruk, M. C., and Schreck, C. J. International best track archive for climate stewardship (ibtracs) project, version 4, 2018. URL https://data.nodc.noaa.gov/cgi-bin/iso?id=gov.noaa.ncdc:C01552. Accessed: 2023-06-01.

Kumar, K. R. and Kalavathi, M. S. Artificial intelligence based forecast models for predicting solar power generation. *Materials today: proceedings*, 5(1):796–802, 2018.

Li, X., Qin, X., Yang, J., Xu, W., et al. Comparison of the applicability of two reanalysis products in estimating tall tower wind based on multiple linear regression and artificial neural network in south china. *Advances in Meteorology*, 2022, 2022.

Mosavi, A., Ozturk, P., and Chau, K.-w. Flood prediction using machine learning models: Literature review. *Water*, 10(11):1536, 2018.

Motta, M., de Castro Neto, M., and Sarmento, P. A mixed approach for urban flood prediction using machine learning and gis. *International journal of disaster risk reduction*, 56:102154, 2021.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. Learning with noisy labels. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/3871bd64012152bfb53fdf04b401193f-Paper.pdf.

Park, S.-J. and Lee, D.-K. Prediction of coastal flooding risk under climate change impacts in south korea using machine learning algorithms. *Environmental Research Letters*, 15(9):094052, 2020.

Richman, M. B. and Leslie, L. M. Adaptive machine learning approaches to seasonal prediction of tropical cyclones. *Procedia Computer Science*, 12:276–281, 2012.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. (arXiv:1505.04597), May 2015. doi: 10.48550/arXiv.1505.04597. URL http://arxiv.org/abs/1505.04597. arXiv:1505.04597 [cs].

Sharadga, H., Hajimirza, S., and Balog, R. S. Time series forecasting of solar power generation for large-scale photovoltaic plants. *Renewable Energy*, 150:797–807, 2020.

Song, H., Kim, M., Park, D., Shin, Y., and Lee, J.-G. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8135–8153, 2023. doi: 10.1109/TNNLS.2022.3152527.

Wolny, A., Cerrone, L., Vijayan, A., Tofanelli, R., Barro, A. V., Louveaux, M., Wenzl, C., Strauss, S., Wilson-Sánchez, D., Lymbouridou, R., Steigleder, S. S., Pape, C., Bailoni, A., Duran-Nebreda, S., Bassel, G. W., Lohmann, J. U., Tsiantis, M., Hamprecht, F. A., Schneitz, K., Maizel, A., and Kreshuk, A. Accurate and versatile 3d segmentation of plant tissues at cellular resolution. *eLife*, 9:e57613, jul 2020. ISSN 2050-084X. doi: 10.7554/eLife.57613. URL https://doi.org/10.7554/eLife.57613.