sbbd:41

Machine learning derived sub-seasonal to seasonal extremes

Daniel Salles Civitarese[1], **Bianca Zadrozny**[1]

[1]IBM Research
Av. Chile, 330 – Rio de Janeiro – RJ – Brazil

{sallesd,biancaz}@br.ibm.com

Abstract. *Improving the accuracy of sub-seasonal to seasonal (S2S) extremes can significantly impact society. Providing S2S forecasts in the form of risk or extreme indices can aid in disaster response, especially for drought and flood events. Additionally, it can provide updates on disease outbreaks and aid in predicting the occurrence, duration, and decline of heatwaves. This work uses a transformer model to predict the daily temperature distributions in the S2S scale. We analyze how the model performs in extreme temperatures by comparing its output distributions with those obtained from ECMWF forecasts across different metrics. Our model produces better responses for temperatures in average and extreme regions. Also, we show how our model better captures the heatwave that hit Europe in the summer of 2019.*

1. Introduction

In 2010, there were 874 disasters related to weather and climate, which caused 68,000 deaths and $99 billion in damage worldwide [Robertson and Vitart 2018]. The 2019 European heatwave caused over 4,000 deaths, with record-breaking temperatures in France, Belgium, Germany, and the Netherlands. Better prediction models for extreme weather will improve early warning systems for preparedness.

Sub-Seasonal to Seasonal climate prediction has long been a gap in operational weather forecasts [Robertson and Vitart 2018]. Its timescale varies from two weeks to an entire season, although some authors have recently used the term S2S more broadly to include seasonal forecasts up to 12 months ahead [Board et al. 2016]. Although S2S forecasts offer many benefits, they are considered more challenging than both numerical weather prediction (NWP) (1-15 days) and seasonal forecasts (2-6 months) due to limited predictive information from land and ocean and a weak predictive signal from the atmosphere [Board et al. 2016].

One way to simplify the problem is to compute indices that aggregate the forecasts in quantiles and compute the probability that a climate variable will be in a specific quantile in the forthcoming weeks, fortnights, or months. It is possible to use the hindcast distribution to establish quantile boundaries. A common way to communicate S2S extreme weather events predictions is by showing how the frequency of an event has changed over a sufficiently large area and period compared to climatology [Robertson and Vitart 2018].

In this work, we use the temporal fusion transformer (TFT) architecture to forecast daily 2-meter temperature quantiles up to 46 days ahead. The TFT encodes a multi-horizon sequence of historical information and forecast outlooks and can handle static data such as spatiotemporal features (e.g., location, climate modes, and land characteristics). Thus, it is a ready-to-use architecture for hybrid ”statistical-dynamical”modeling.

2. Related work

[Civitarese et al. 2021] propose using a TFT model to forecast maximum daily precipitation quantiles up to 6 months in advance. Results show that TFT outperforms a calibrated ECMWF SEAS5 ensemble forecast in quantile risk, but it is hard to link to extreme weather events. Here we use a similar TFT model but generalize it through an overall methodology that includes post-processing to generate extreme indices.

[Akram Zaytar et al. 2022] proposes an ML-based daily forecast model that predicts global 2-meter temperature and total precipitation. It combines multimodal data into feature vectors given as inputs to three ML models: XGBoost, U-Net, and NGBoosting. The method consistently outperforms both ECMWF forecasts and climatology regarding the Ranked Probability Skill Score (RPSS), but no assessment of specific extreme weather events prediction accuracy is done.

[Patel et al. 2022] identified extreme weather indices for improved communication of climate risk in S2S scales. These indices can be calculated from probabilistic daily forecasts or trained ML-models. We apply the first approach of directly calculating extreme weather indices from a Probabilistic TFT model.

3. Methodology

We developed a pipeline to preprocess forecasts (Figure 1), train and run probabilistic TFT models, calculate output distributions, and validate results using means and specific indices. The Dataset box (yellow) includes time-dependent data and static covariates. The light blue Preprocessing Featurization box includes all the steps for generating embeddings to feed into predictive models. The Probabilistic TFT box (green) represents the model that predicts climate variables, including temperature. This model is shown in more detail on the right panel of the Figure and will be described in the next paragraph. The light blue boxes calculate extreme indices or event likelihood based on our model's forecasts. It depends on an accurate probabilistic distribution. Red boxes evaluate model performance by measuring the skill of probabilistic forecasters in predicting extreme weather using quantifiable methods. In this work, we focus specifically on extreme temperatures above the 90^{th} percentile.

The right panel of Figure 1 shows the architecture of our Probabilistic TFT in detail. The input structure is similar to the one presented in [Lim et al. 2021], where the authors consider three kinds of input: (a) static $\mathbf{s} \in \mathbb{R}^{m_s}$, (b) known future $\mathbf{x} \in \mathbb{R}^{m_x}$, and (c) historical $\mathbf{z} \in \mathbb{R}^{m_z}$ (Figure 1 – yellow, blue, and red, respectively). Static information, such as location, land cover or altitude, does not change through time, so we repeat these values for all time steps for the encoder and the decoder. We use TFT's "known inputs" to embed actual known future variables, such as day-of-the-year and external forecasts. Some examples of the latter are geopotential and column of water. Although these forecasts are not "known information" and may have errors, their use follows the same procedure as future inputs. Historical inputs refer to past observations for the target series or other exogenous factors, such as surface pressure or soil moisture.

We produce daily forecasts for each location in the form $\hat{y}_{lat,lon}(q, t, \tau)$, where each time-step $t \in [0, T]$ refers to a specific day in the time series, and $\tau \in [1, 46]$ is the lead time in days. The index $(lat, lon) = i$ refers to the location in the globe, and it is associated with a set of static inputs $\mathbf{s}_i$, as well as the time-dependent inputs $\mathbf{z}_{i,t}$ and $\mathbf{x}_{i,t}$.

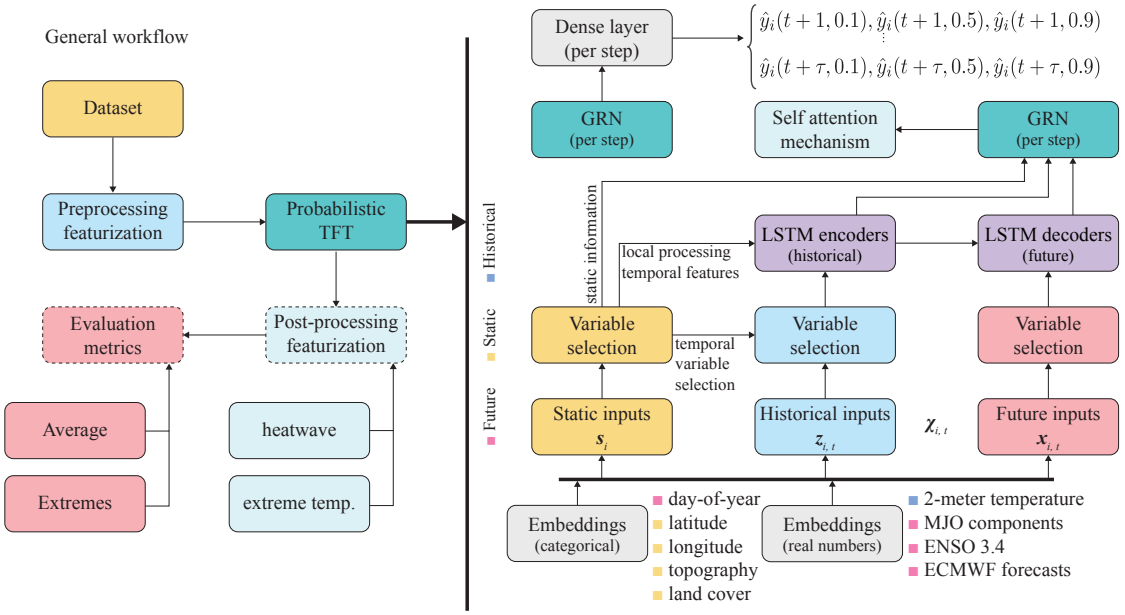

Figura 1. System pipeline overview.

In our experiments, we predict the quantiles 0.02, 0.1, 0.25, 0.5, 0.75, 0.9, and 0.98 of the 2-meter temperature for each day up to 46 days ahead. Each quantile is represented by $\hat{y}_i(q,t,\tau) = f_q(\tau, y_{i,t-k:t}, z_{i,t-k:t}, x_{i,t-k:t+\tau}, s_i)$. Finally, we approximate a normal distribution using the produced quantiles, i.e, $\hat{y}_i(q,t,\tau) \xrightarrow{g} \mathcal{N}_{i,t}(\mu_{i,t}, \sigma^2_{i,t})$, where g is the fitting function.

As in [Lim et al. 2021], we train the TFT model by jointly minimizing the quantile loss, summed across all quantile outputs:

$$\mathcal{L}(\Omega, \mathbf{W}) = \sum_{y_t \in \Omega} \sum_{q \in \mathcal{Q}} \sum_{\tau=1}^{\tau_{max}} \frac{QL\left(y_t, \hat{y}(q, t-\tau, \tau), q\right)}{M\tau_{max}} \tag{1}$$

$$QL(y, \hat{y}, q) = q(y - \hat{y})_+ + (1-q)(\hat{y} - y)_+ \tag{2}$$

where Ω is the domain of the training data, which contains M samples, $\mathbf{W}$ refers to the TFT weights, $\mathcal{Q}$ is the set of output quantiles, and $(.)_+ = max(0, .)$. The normalized quantile loss (q-risk) is used for validation and testing:

$$\text{q-risk} = \frac{2 \sum_{y_t \in \tilde{\Omega}} \sum_{\tau=1}^{\tau_{max}} QL\left(y_t, \hat{y}(q, t-\tau, \tau), q\right)}{\sum_{y_t \in \tilde{\Omega}} \sum_{\tau=1}^{\tau_{max}} |t_t|} \tag{3}$$

where $\tilde{\Omega}$ is the domain of the validation or test samples.

4. Experiments and discussions

4.1. Data

Our experiments use six data sources: GMTED, C3S-LC, CPC, ECMWF extended-range, ENSO 3.4, and MJO. The Global Multi-resolution Terrain Elevation Data (GMTED 2010) is a global elevation model [Poppenga et al. 2010]. It has a global accuracy of 6 m RMSE

and an RMSE range of 25 to 42 meters depending on the resolution. The C3S land cover provides global land surface maps with 22 classes [Store 2019]. CPC[1] provides global observation data for 2-meter temperature and total precipitation, gridded using the Shepard Algorithm. ECMWF extended-range forecasts provide a 46-day outlook on forthcoming weather conditions. The forecasts are based on data available at a resolution of 1.5º from the S2S challenge[2]. ENSO 3.4 index refers to the sea surface temperature anomaly in Niño 3.4 region (5ºN-5ºS, 120º-170ºW). The Madden-Julian Oscillation (MJO) is a major source of variability in the tropical atmosphere, moving west to east in 30-60 days and impacting weather across various latitudes. We downloaded both ENSO and MJO from the NOAA website. These are the variable that we use in our experiments:

- **historical**: 2-meter temperature (target), total precipitation from CPC;
- **known**: day-of-year (sin/cos), 2-meter temperature, precipitation, geopotential, surface pressure, soil temperature top 20 and 100 cm, soil moisture top 20 and 100 cm, total cloud cover, total column water, and time integrated top net thermal radiation from ECMWF, ENSO 3.4, and MJO components amplitude, phase, RMM1, and RMM2;
- **static**: latitude, longitude (sin/cos), topography, and land cover.

The final dataset includes 2.7 TB of global data from 2010 to 2020, stored in Zarr format for parallel reading. Xarray and Dask Python libraries are used to preprocess the files. Finally, we store the training, validation, and testing datasets in Feather format using Pandas.

4.2. Training

We used PyTorch Forecasting, PyTorch, Lightning, and Optuna for HPO and final training. We ran them on a 32-CPU machine with 128 GB and an NVIDIA A100 with 80 GB. Over 50 candidates were generated in 24 hours. The final training took less than 20 minutes over five epochs. Inference time was under two seconds for 46 days per site.

The hyper-parameter optimization selects the best architecture settings, including the input length (28, 35, 42, 84), the number of attention heads $[1, 8]$, dropout rate $[0.1, 0.3]$, learning rate $[10^{-4}, 0.1]$, gradient clipping $[10^{-2}, 1.0]$, and channels for categorical and continuous encoders $[30, 128]$. The best configuration has the following parameters: input 42, 1 attention head, learning rate of 8.7×10^{-3}, clipping of 2.4×10^{-2}, encoder size 56. After selecting the optimal topology and training configuration, we train the model for 50 epochs with early stopping, allowing for five epochs of patience.

4.3. Results

We evaluate our model based on the expected forecast value ($\mu_{i,j}$) and extreme temperatures (above the 90^{th} percentile). In the former, we compute the anomaly correlation coefficient (ACC), the Continuous Ranked Probability Score (CRPS), and the mean absolute error (MAE), whereas in the latter, we compute the reliability and sharpness.

On average, our model outperformed climatology and the current ECMWF forecasts in all metrics across all locations, forecast times, and lead times. Table 1 shows the

[1]CPC Global Unified Temperature data provided by the NOAA PSL, Boulder, Colorado, USA, from their website at https://psl.noaa.gov

[2]https://s2s-ai-challenge.github.io

Tabela 1. Averaged results over all dates, lead times and locations

	Climo	ECMWF	TFT
ACC	-	0.067	0.074
CRPS	2.140	2.539	2.024
MAE	4.782	3.360	2.692

results for ACC, CRPS, and MAE. Notice that our CRPS is lower than the climatology, indicating a positive skill score.

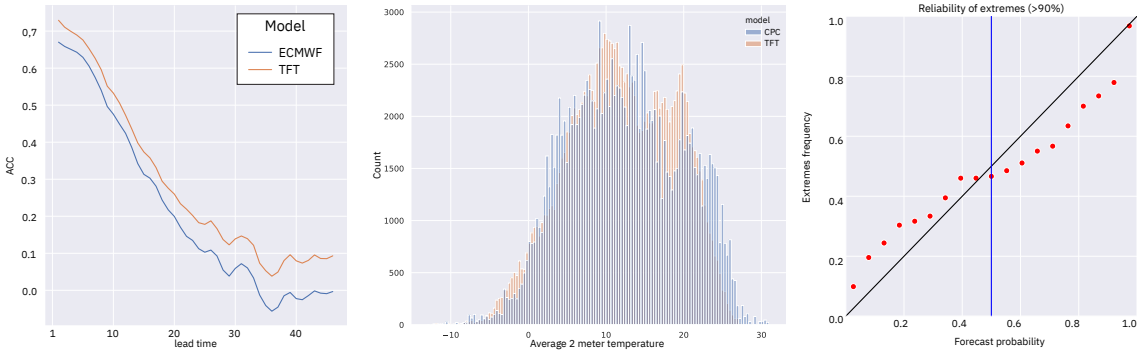

Figura 2. Anomaly correlation coefficient (left), sharpness (center), reliability over quantile 0.9. (right).

In Figure 2, we see the ACC, the sharpness, and the reliability from left to right. The TFT produces a better ACC than the ECMWF for all lead times, but it depends highly on the forecasts. However, this is not a problem since our goal is to improve the forecasts, not replace them. The sharpness and reliability plots demonstrate that the model's predictions closely match the observations and produce distributions centered on lower and higher values, which is highly desirable.

Figure 3 compares the ECMWF and TFT forecasts with the European heatwave that struck the region in 2019. In the plots, the blue dots represent observations from CPC, while the red, green, and purple lines show the minimum, mean, and maximum values of the ECMWF ensemble. The orange line and shades represent the TFT output quantiles. In France, Germany, and England, we notice that the TFT assigns higher probabilities for observations above the limits of ECMWF ensemble members, i.e., the original forecasts did not capture these values. Conversely, Moldova experienced a colder season, and the TFT captured the lower values the original forecasts did not.

5. Conclusion

In this study, the TFT model shows potential in predicting daily temperature distributions on the S2S scale. It outperformed both climatology and current ECMWF forecasts, demonstrating its ability to improve weather forecasting accuracy. Providing risk or extreme indices as S2S forecasts can aid in disaster response. The model's ability to capture extreme temperatures, as seen in the 2019 European heatwave and colder season in Moldova, highlights its usefulness in predicting unusual weather events. Future work aims to investigate how surface variables from various global regions, such as sea level and sea surface temperature, would impact the model's performance.

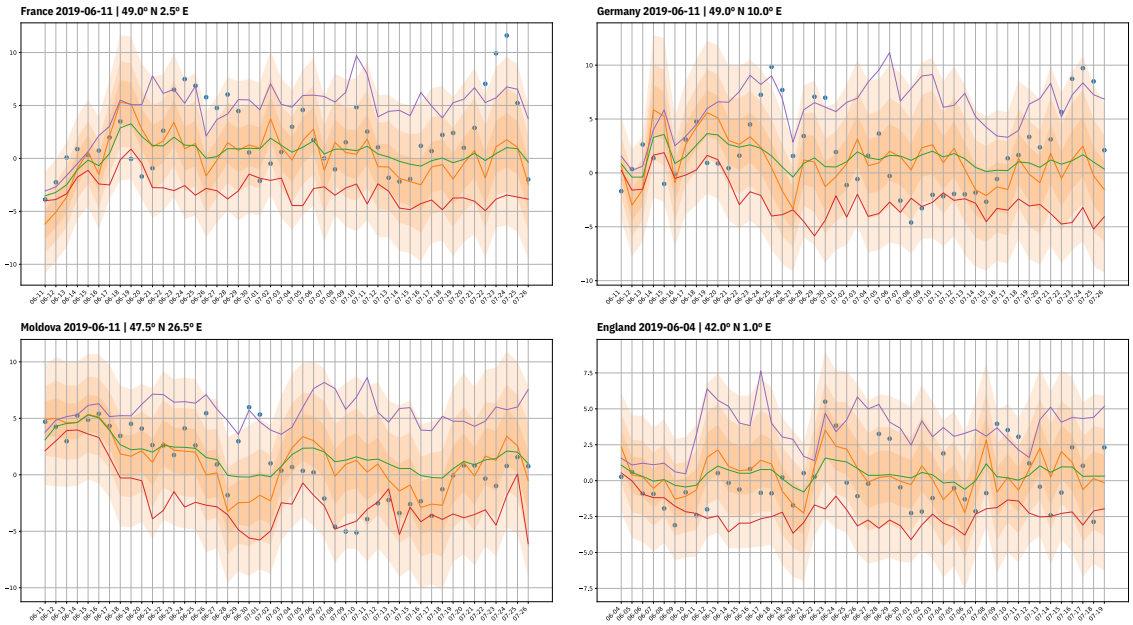

Figura 3. Four locations that were affected by the temperature extremes in 2019.

Referências

Akram Zaytar, M., Zadrozny, B., Watson, C., Salles Civitarese, D., Eben Vos, E., Mathonsi, T. M., and Lukhetho Mashinini, T. (2022). ML-based Probabilistic Prediction of 2m Temperature and Total Precipitation. In *EGU General Assembly Conference Abstracts*, EGU General Assembly Conference Abstracts, pages EGU22–11063.

Board, O. S., of Sciences Engineering, N. A., Medicine, et al. (2016). *Next generation earth system prediction: strategies for subseasonal to seasonal forecasts*. National Academies Press.

Civitarese, D. S., Szwarcman, D., Zadrozny, B., and Watson, C. (2021). Extreme precipitation seasonal forecast using a transformer neural network.

Lim, B., Arık, S. Ö., Loeff, N., and Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764.

Patel, Z., Baloyi, G., Watson, C., Zaytar, A., Zadrozny, B., Civitarese, D., Makhanya, S., and Vos, E. (2022). S2S Extreme Weather Featurization: A Global Skill Assessment Study. In *EGU General Assembly Conference Abstracts*, EGU General Assembly Conference Abstracts, pages EGU22–12461.

Poppenga, S. K., Evans, G., Gesch, D., Stoker, J. M., Queija, V. R., Worstell, B., Tyler, D. J., Danielson, J., Bliss, N., and Greenlee, S. (2010). Topographic science. Technical report, US Geological Survey.

Robertson, A. and Vitart, F. (2018). *Sub-seasonal to seasonal prediction: the gap between weather and climate forecasting*. Elsevier.

Store, C. C. D. (2019). Land cover classification gridded maps from 1992 to present derived from satellite observations.