Neural Processes for Short-Term Forecasting of Weather Attributes

Benedetta L. Mussati [1] Helen McKay [2] Stephen Roberts [1, 2]

[1]University of Oxford [2]Mind Foundry Ltd

Problem setting

Forecasting from time series data is of key relevance in the field of climate science. Modelling short-term dynamics is of particular interest to detect unexpected weather conditions and potential precursors of extreme events. Modelling with limited data is important since the sensors collecting data can be prone to failure.

Objective

- Machine learning models offer a data-driven approach, relying on statistical approaches to identify patterns in historical climate data to aid future predictions.
- We cast time series forecasting as a multi-task meta-learning problem, where each weather attribute corresponds to a task.
- We investigate how neural processes (NPs) can be used for short-term forecasting with incomplete context data and, if by simultaneously learning a set of correlated tasks, the model can leverage information from their joint distribution.

Neural Processes

Neural processes (NPs) are a family of meta-learning algorithms developed as a neural network approximation to Gaussian processes (GPs) [1, 2]. We consider predicting at target locations T by learning the mapping $f : \mathcal{X} \to \mathcal{Y}$ that generated a labelled set of data, the context set C. Assuming $f \sim P$, the objective is to learn a model whose parameters maximise the likelihood of P [3].

Fig. 1 shows the computational diagram for the multi-task neural process (MTNP) [4]. The task-specific latent variables $v^1, \dots, v^N$ are conditioned on the global latent variable z to express inter-task correlation. The single-task neural process (STNP) has N independent latent variables $v^1, \dots, v^N$.

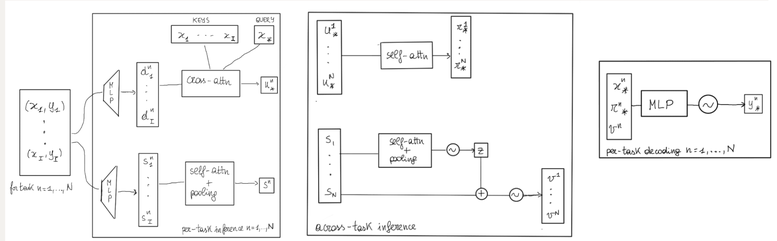


Figure 1. Computation diagram of the multi-task neural process (MTNP), adapted from [4], showing the computations per-task n and the across-task inference phase. x_*, y_* are the target location and predicted output.

References & Acknowledgments

[1] M. Garnelo, D. Rosenbaum, C. Maddison, T. Ramalho, D. Saxton, M. Shanahan, Y. W. Teh, D. J. Rezende, and S. M. A. Eslami. Conditional neural processes. In J. G. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1690–1699. PMLR, 2018.

[2] M. Garnelo, J. Schwarz, D. Rosenbaum, F. Viola, D. J. Rezende, S. M. A. Eslami, and Y. W. Teh. Neural processes. *CoRR*, abs/1807.01622, 2018.

[3] S. Jha, D. Gong, X. Wang, R. E. Turner, and L. Yao. The neural process family: Survey, applications and perspectives. *CoRR*, abs/2209.00517, 2022.

[4] D. Kim, S. Cho, W. Lee, and S. Hong. Multi-task neural processes. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022.

Benedetta L. Mussati acknowledges funding and support from Mind Foundry Ltd. and the EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines and Systems (Grant No. EP/S024050/1).

Data and preprocessing

The *chimet* dataset is a week-long dataset collected from Chichester Harbour in the United Kingdom, in August 2007. The recorded weather attributes are: *air temperature (C)*, *max wave height (m)*, *mean wave height (m)*, *sea temperature (C)*, *tide height (m)* (shown in Fig. 2), *wind gust speed (kn)* and *wind speed (kn)*.

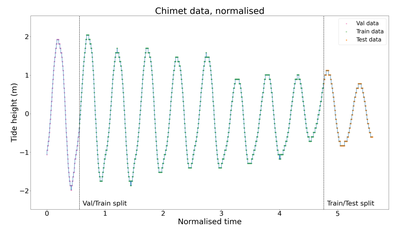


Figure 2. *Tide height (m)* weather attribute time series, with the train-validation-test splits. The timestamps have been converted to float values in the range $[0, 6]$ and the datapoints have been normalised.

Proposed approach

The functions making up a dataset are fixed-length slices of the original data with a fixed time lag between the start of two consecutive slices. We experiment with datasets of 2-hour long slices with 5 minute intervals and with datasets of 6-hour long slices with 30 minute intervals.

Given a function draw, the context points' locations are the same for all tasks. To simulate incomplete context data, some context points are randomly dropped with *missing rate* $\gamma \in [0, 1]$ independently for each weather attribute [4]. For each task n, the set C^n of a function's context points and the target locations T^n are given as input to the MTNP/STNP.

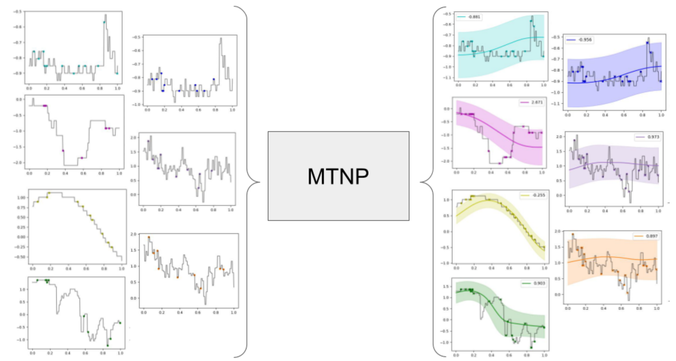


Figure 3. Forecasting on one function draw for all N tasks. On the left, the coloured points are the context points C^n that the model receives as input for task n. On the right, the mean of the predictive distribution $p(y_*^n | x_*^n, C^n)$ with 95% confidence intervals is plotted for task $n = 1, \dots, N$. The crosses are the context points that were randomly dropped in each task.

The models are assessed using negative log-likelihood (NLL) and mean squared error (MSE) averaged across all N tasks and M points.

$$\text{NLL} = \frac{\sum_{n=1}^{N} \text{NLL}^n}{N} = -\frac{\sum_{n=1}^{N}\sum_{m=1}^{M} \log p(\hat{y}_m^n | x_m^n, C^n)}{N \times M}$$

$$\text{MSE} = \frac{\sum_{n=1}^{N} \text{MSE}^n}{N} = \frac{\sum_{n=1}^{N}\sum_{m=1}^{M} (\hat{y}_m^n - y_m^n)^2}{N \times M}$$

Results: 2-hour forecast horizon

Terminology: C^n is the context set and T^n the target locations for task n. γ_{train} is the missing rate at training phase and γ_{eval} at evaluation phase. cs_{eval} is the size of the context set at evaluation.

$\gamma_{train}, \gamma_{eval}$	Model	NLL	MSE
0.2	STNP	**−0.5913**	**0.0343**
	MTNP	−0.4438	0.0390
0.5	STNP	**−0.4467**	**0.0443**
	MTNP	−0.2967	0.0484
0.8	STNP	0.1910	**0.0724**
	MTNP	**0.1089**	0.0870

Table 1. Comparison of cumulative NLL and MSE with varying $\gamma_{train}, \gamma_{eval}$ and fixed $cs_{eval} = 10$.

- As shown in Table 2 and Fig. 4, the MTNP leverages inter-task correlation to maximise $p(f(T^n)|C^n, T^n)$ for task n when C^n is small (small cs_{eval} and $\gamma_{train}, \gamma_{eval} = 0.8$).
- However, the STNP performs better than the MTNP when provided with sufficient context information (see Table 1).

cs_{eval}	Model	NLL	MSE
5	STNP	0.9839	**0.1455**
	MTNP	**0.5502**	0.1515
10	STNP	0.1910	**0.0724**
	MTNP	**0.1089**	0.0870
20	STNP	**−0.3140**	**0.0505**
	MTNP	−0.1668	0.0694

Table 2. Comparison of cumulative NLL and MSE on the 2-hour horizon, with varying cs_{eval} and fixed $\gamma_{train}, \gamma_{eval} = 0.8$.

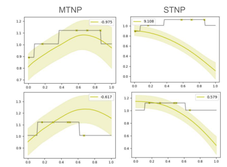


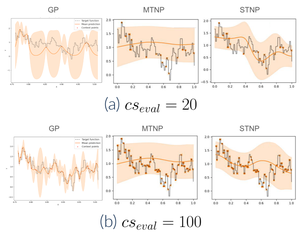
Figure 4. Comparison of MTNP's and STNP's predictions on evaluation functions of *Tide height (m)*, with $cs_{eval} = 5$ and $\gamma_{eval} = 0.8$.

Results: 6-hour forecast horizon

- Fix $\gamma_{train}, \gamma_{eval} = 0.5$ and compare the MTNP and the STNP to an ensemble of GPs.
- With a longer forecast horizon, the STNP outperforms the MTNP and GP when cs_{eval} is small (Table 3). When the number of context points is large, we find the NPs are outperformed by the GP approach (Fig. 5b).
- With less context information, the GP is less certain of its prediction (see Fig. 5a), and its predictive mean does not approximate the true function as well as the STNP in areas with no context points.

cs_{eval}	Model	NLL	MSE
10	MTNP	0.5542	0.1740
	STNP	**0.1449**	**0.0933**
	GP	0.4252	0.2955
20	MTNP	0.3058	0.1139
	STNP	**-0.0911**	**0.0725**
	GP	0.0348	0.1667
50	MTNP	0.2561	0.1086
	STNP	-0.2281	0.0632
	GP	**-0.5705**	**0.0475**
100	MTNP	0.2493	0.1069
	STNP	-0.2649	0.0616
	GP	**-0.9895**	**0.0203**
200	MTNP	0.2188	0.1044
	STNP	-0.2660	0.0613
	GP	**-1.3624**	**0.0080**

Table 3. Comparison of cumulative NLL and MSE, with varying cs_{eval}. The missing rates of the NPs are fixed $\gamma_{train}, \gamma_{eval} = 0.5$.



(a) $cs_{eval} = 20$

(b) $cs_{eval} = 100$

Figure 5. Comparison of GP's, MTNP's and STNP's predictions on the same evaluation function of the *Wind speed (kn)* weather attribute.

Conclusions

We present preliminary results detailing how NPs can be used as a data-driven approach to learn short-term dynamics for forecasting purposes. When limited context information is provided, the MTNP leverages inter-task knowledge and outperforms the STNP. The STNP outperforms both the MTNP and the GPs ensemble when context information is sufficient but not excessive.

blmussati@robots.ox.ac.uk