Medium-range weather forecasts with probabilistic machine learning methods

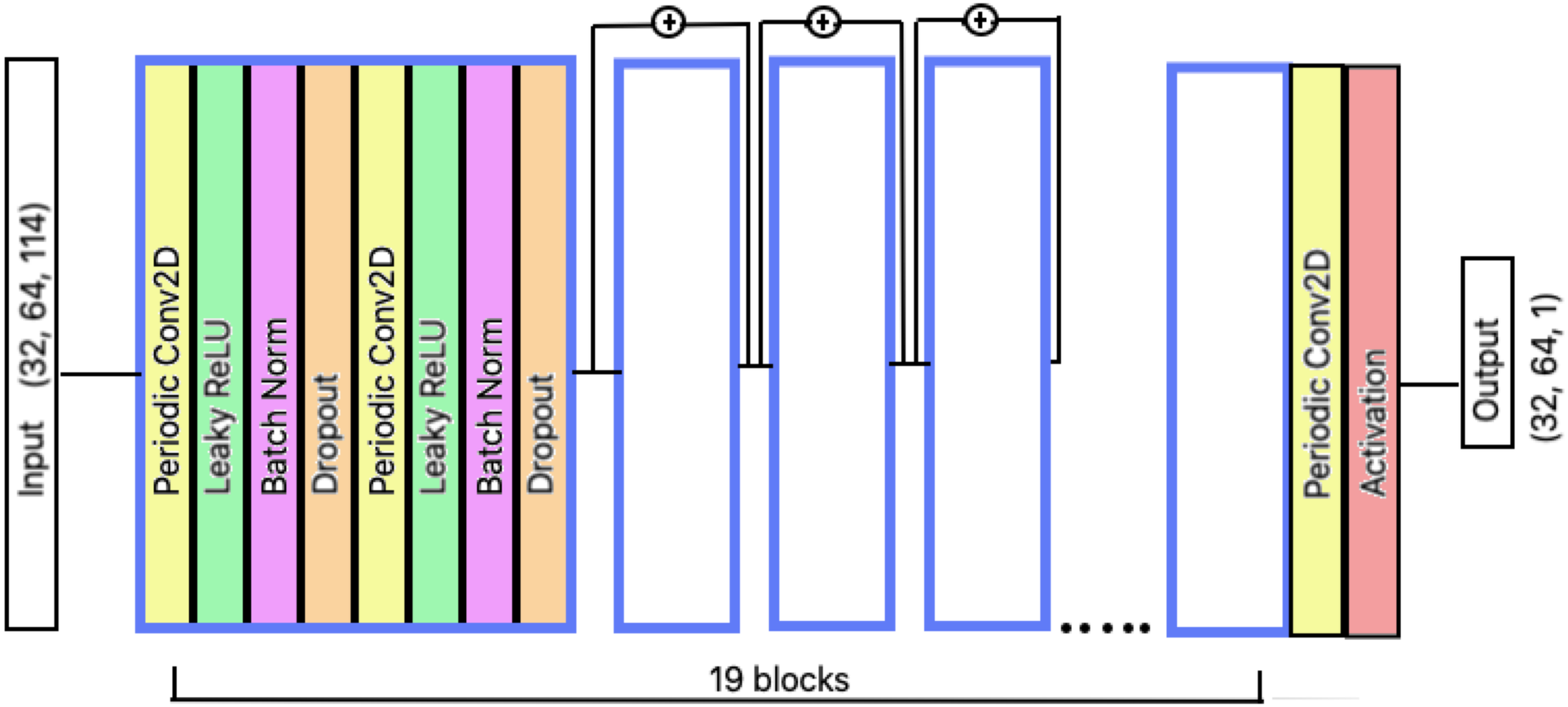
Sagar Garg[1], Stephan Rasp[2], Nils Thuerey[1]
[1] Dept. of Informatics, Technical University of Munich, [2] ClimateAi Inc.

Supplementary Material

Data

- ERA5 **Reanalysis** Data by ECMWF—> Regridded at 5.625 degrees,
  i.e, 32x64 lat-lon grid, (~625x625 sq km at equator)
-
- Inputs (114): several 3D and 2D fields describing atmospheric flow + auxiliary. 7 vertical levels, 3 timesteps
  - Geopotential, Temperature, wind-velocity, specific humidity, 2-metre temperature, total precipitation, total incident solar radiation, land-sea mask, orography, latitude
- Outputs (4)
  - Upper-level atmosphere: Geopotential at 500hPa (**z**), Temperature at 850hPa (**t**)
  - Surface: 2-metre temperature (**t2m**), total precipitation (**tp**)
- Training: 1979-2015. Validation : 2016. Test: 2017-2018. Data at 2 hours' interval.

Primary Model Architecture -ResNet



- For Longitudes: Periodic Conv2D- ‘soft’ cylindrical boundary condition in longitude direction
- For latitudes: All loss-functions and evaluations metrics are area-weighted
- Training on single GPU takes ~1 day. Trained with Adam Optimizer using early stopping, L2 regularisation, LR decay

Evaluation Metrics

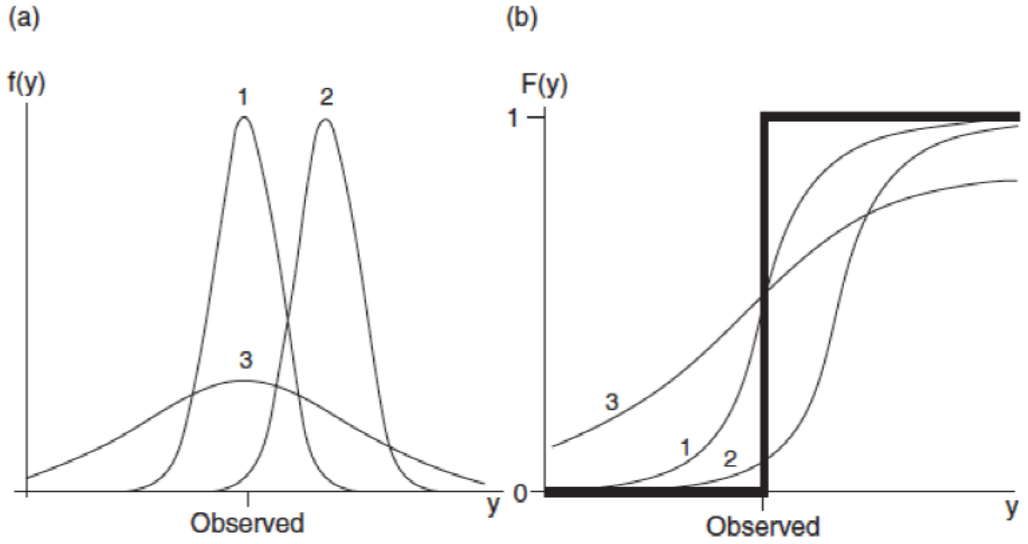
- RMSE of ensemble mean
- Spread-Skill Ratio (Ideally=1)
- Continuous Ranked Probability Score
- Rank Histograms

Continuous Ranked Probability Score

- Strictly proper scoring rule to compare a cumulative probability distribution to a point observation. (s*maller is better*)

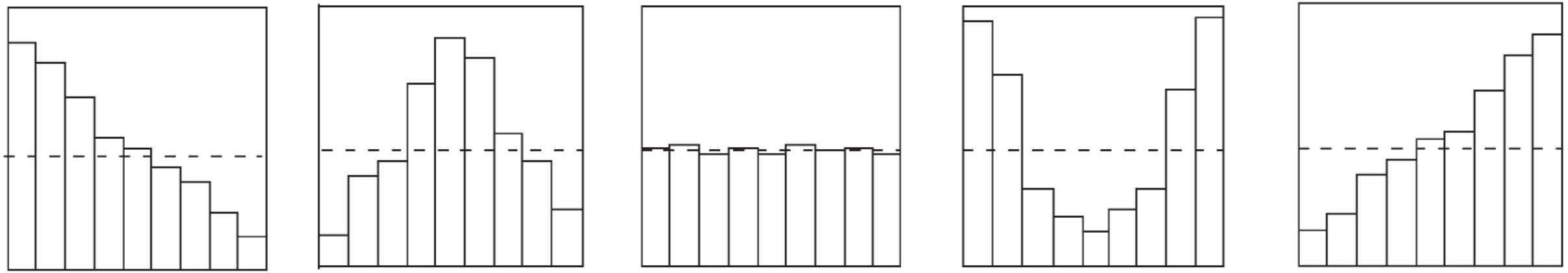
$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} [F(z) - 1(y \leq z)]^2 dz$$

- Maximizes **sharpness** and **calibration**
- Reduces to Mean Absolute Error for Deterministic forecast. (*easy* to *compare*)
- Can be used as an evaluation metric or loss function (*as in Parametric Forecast*)



Rank Histograms

- ‘reliable’ forecast - observation equally likely to fall between any two members
- Construction:
  - 1. For each observation point, rank the N ensemble members from lowest to highest value, representing N+1 bins.
  - 2.Identify the ‘rank’ of the observation, i.e., which bin it falls into.
  - 3.Tally over many observations to create a rank histogram



Overforecasting Bias

Overdispersion
(Underconfident)

Rank Uniformity

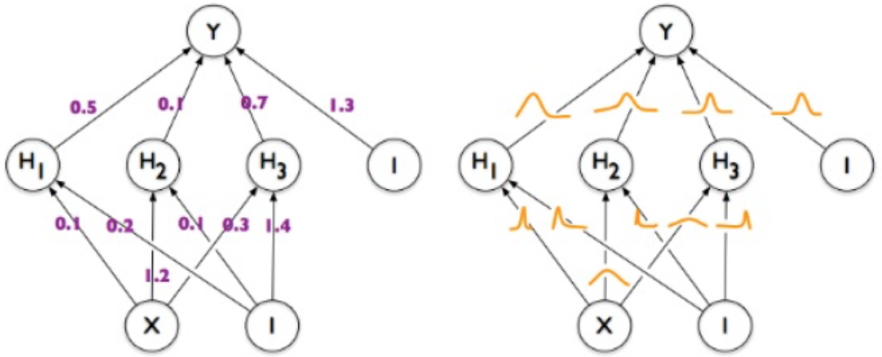
Underdispersion
(Overconfident)

Underforecasting Bias

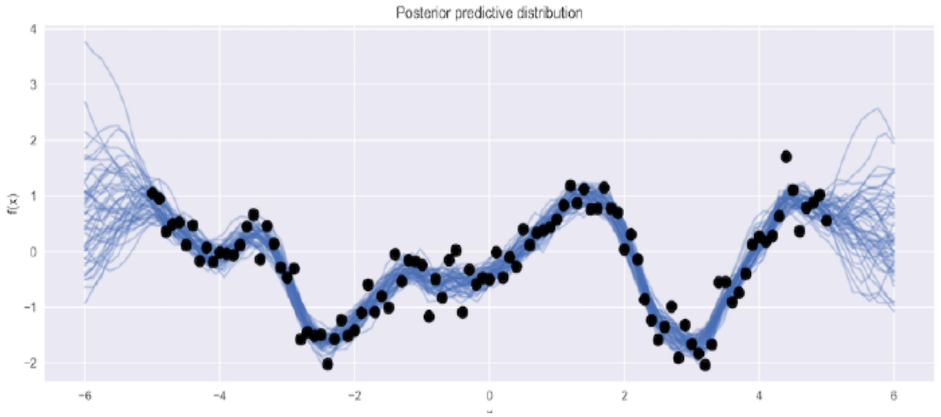
Exp. 1: Monte-Carlo Dropout

- In Bayesian NNs,
  - weights follow some distribution.
  - (Intractable) Predictive Posterior Distribution

$$p\left(x_{new} \mid X\right) = \int p\left(x_{new} \mid \omega\right) p(\omega \mid X) d\omega$$



- In Monte-Carlo Dropout:
  - 'Dropout' training interpreted as approximate variational inference to Bayesian NN.
  - Train a model and simply switch 'on' dropout during test-time
  - Make 50 stochastic forward passes, use as ensemble members

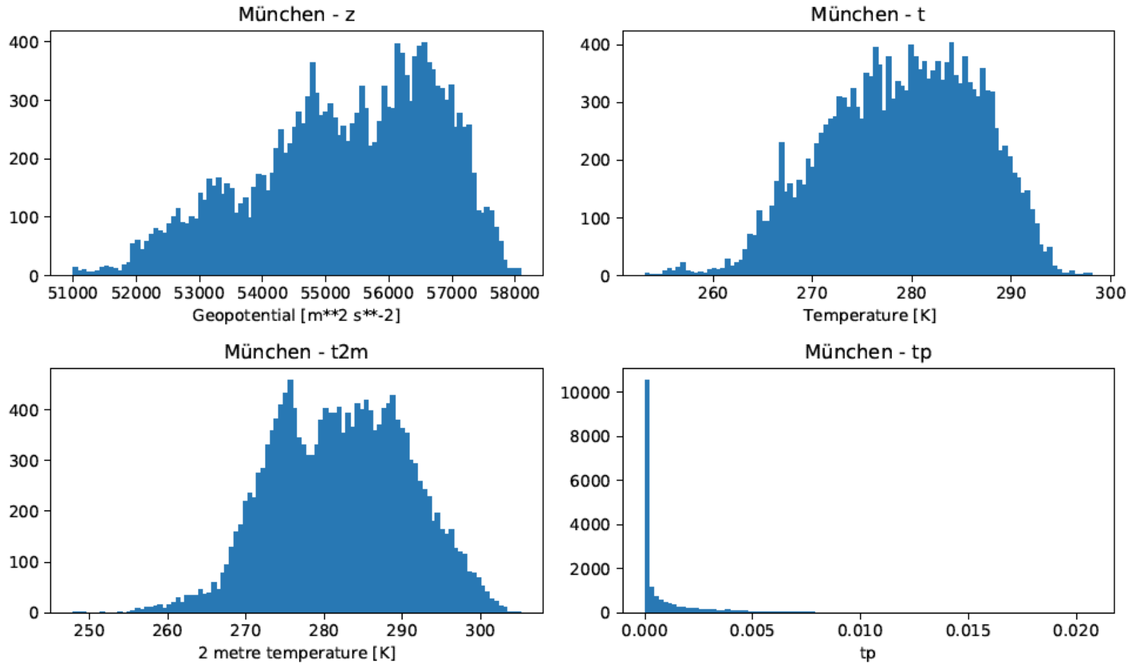


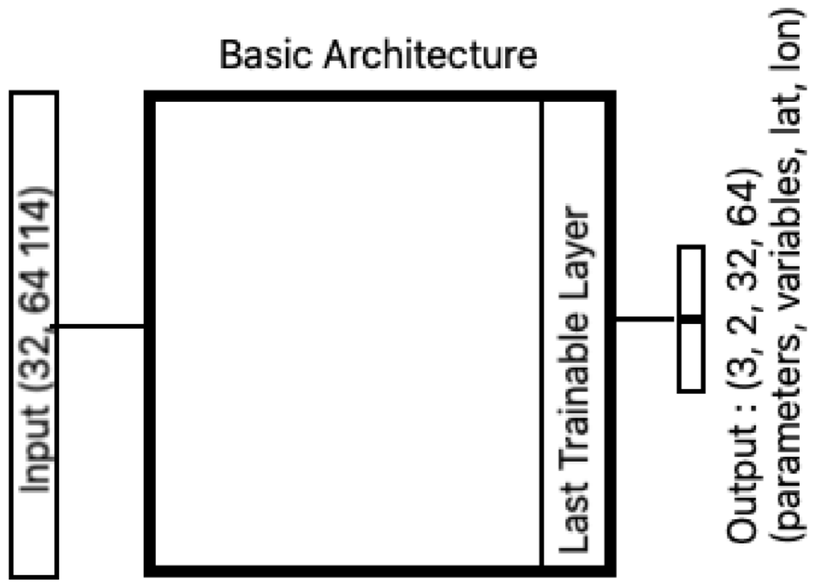
Note: In original paper [Y. Gal, Z. Ghahramani],

- **mean and variance of forward passes are estimated, by optimizing a length-scale parameter τ. Also, dropout is applied after conv, but we apply it before.**
- **New research shows better implementation for ResNet-with stochasticity in image patches, features, entire layers**

Exp. 2: Parametric Forecast

- Z, T, T2M - Gaussian distributions.
  TP- Generalized Extreme Value*
- Turns into a Regression problem
  - Output: Mean, Variance.
  - Ensemble: draw 50 samples from PDF
- Loss function: CRPS, a probabilistic score
  - maximizes **sharpness** and **calibration**

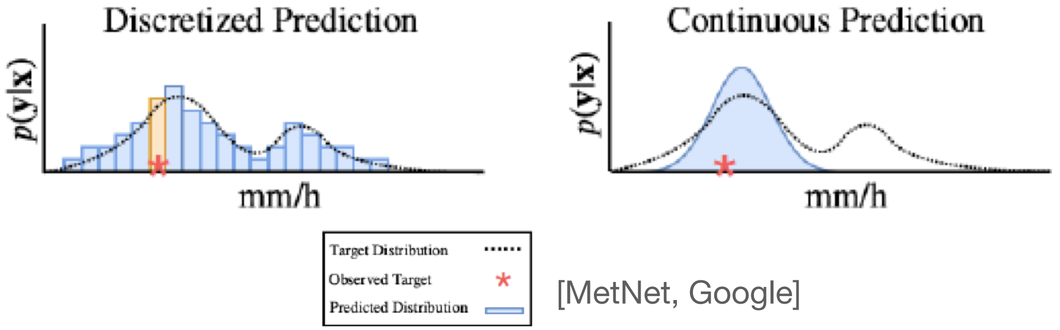




* didn't implement

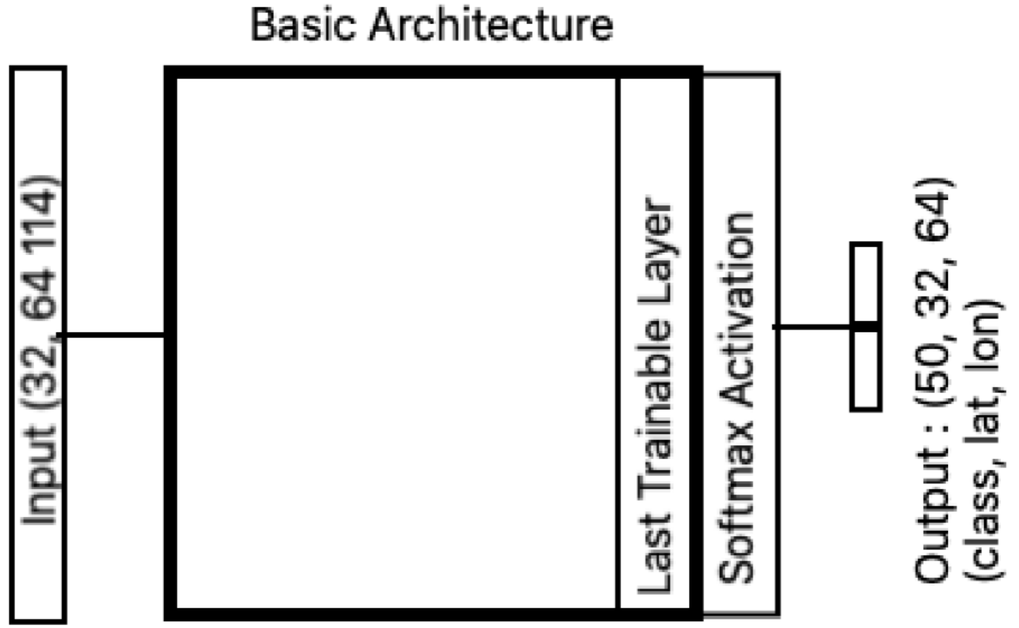
Exp. 3: Categorical Forecast

- Divide range into 50 discrete bins
- Turns into multi-class classification problem



[MetNet, Google]

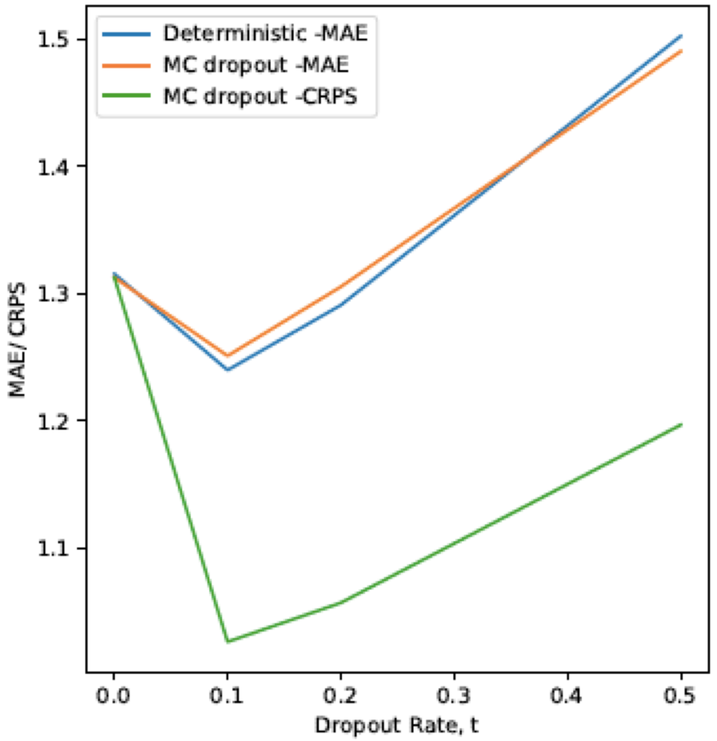
- Highly flexible, expressive, tractable, scalable for all kinds of distribution
- Performance dependent on bin size.
  - Too few- coarse approximation
  - Too many- low #samples/bin. Non-smooth.

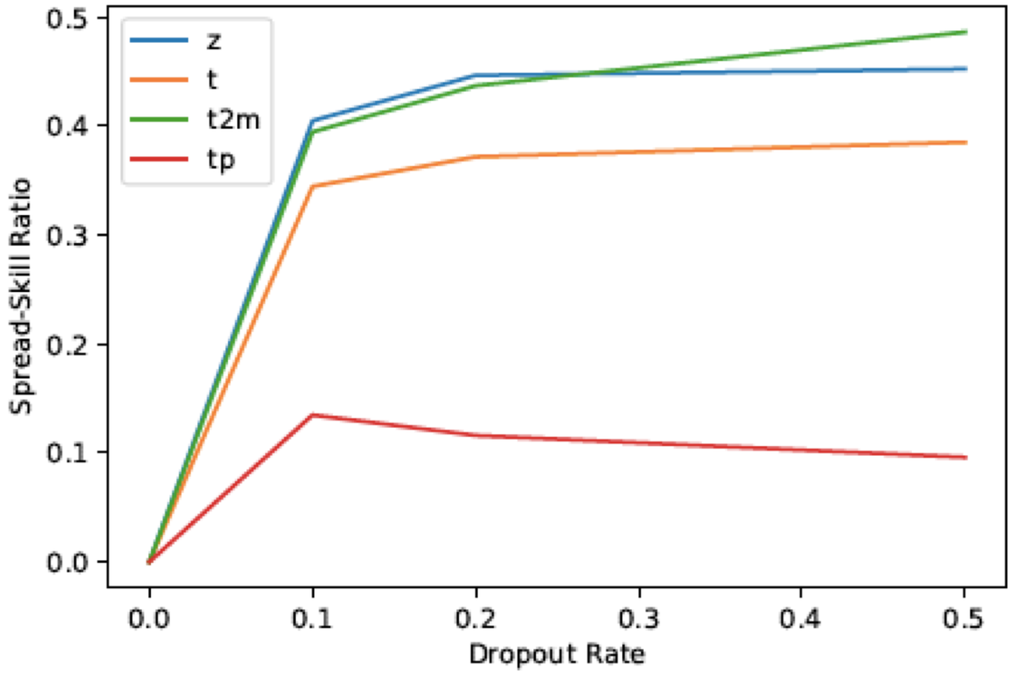


- **Note: Separate networks for : Z, T, T2M, TP**

Results

MC Dropout vs Deterministic



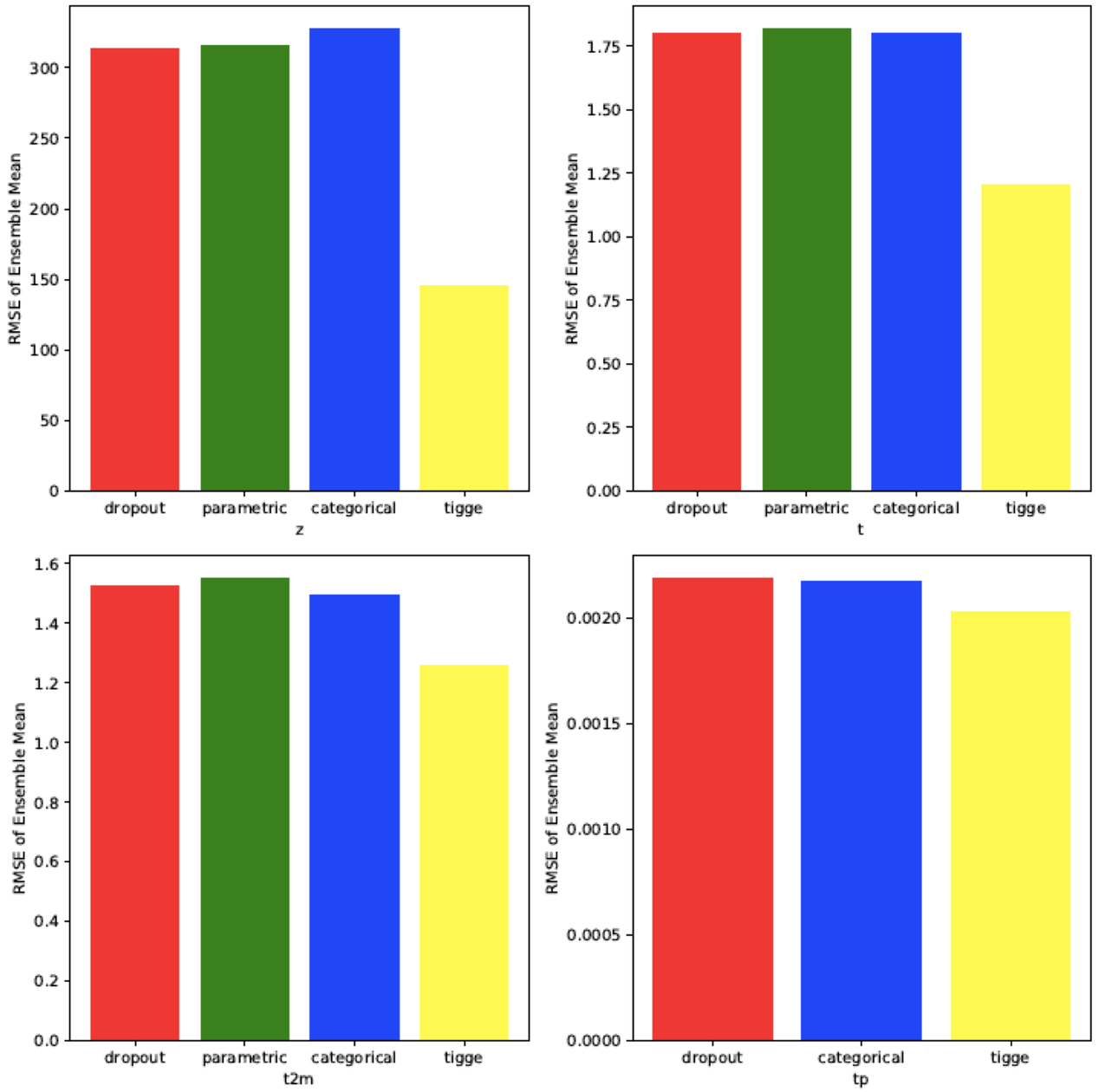


- MC Dropout performs similar on MAE, better on CRPS
  By drawing multiple realisations (50) with dropout, more probabilistic information is learned.

- Ideal Spread-Skill ratio=1 (For ‘perfect’ ensemble, Avg. RMSE= Avg. Standard Deviation)
  MC Dropout has low spread-skill ratio —> Largely overconfident predictions

RMSE of Ensemble Mean

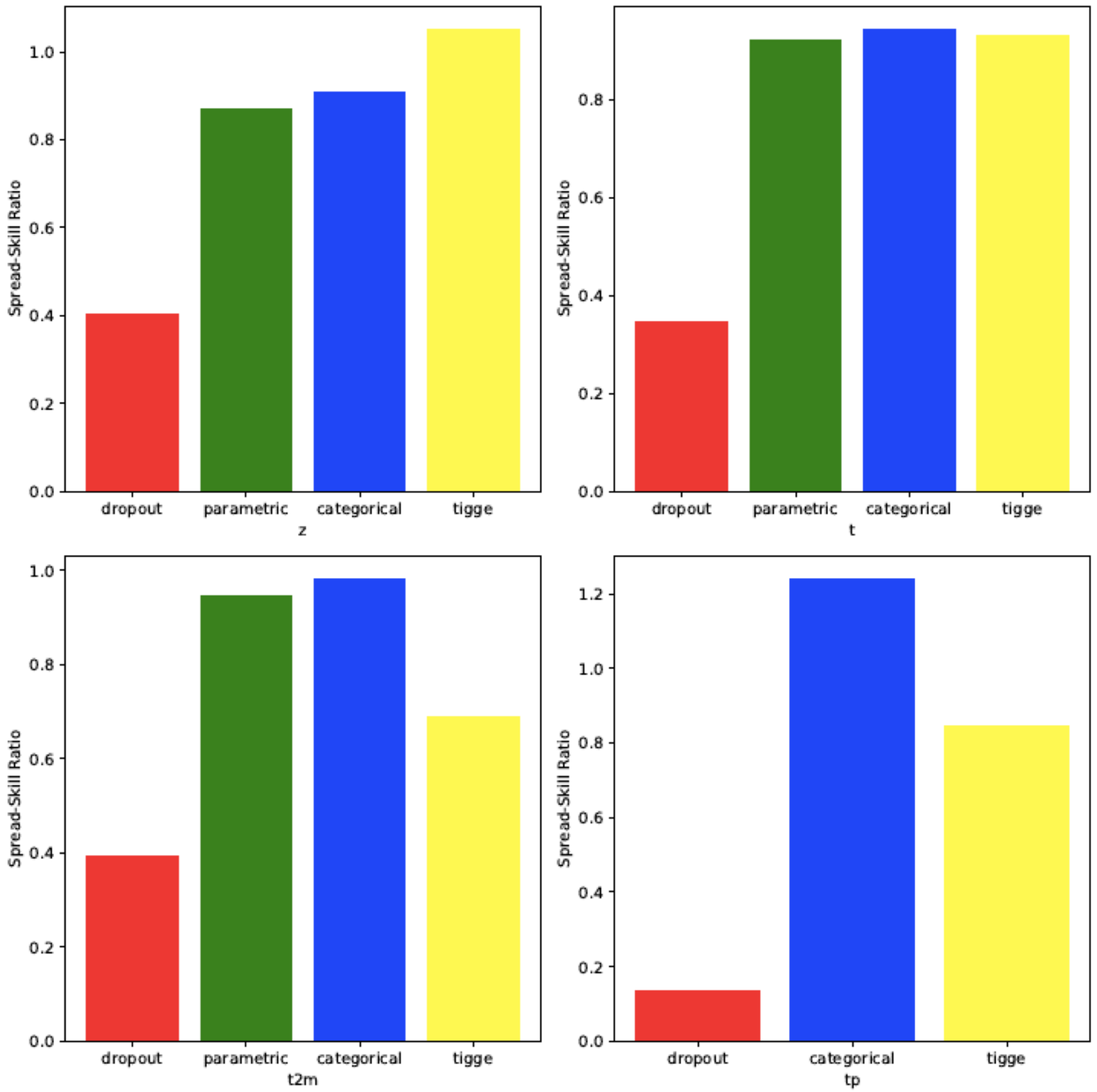
- TIGGE: State-of-the-art NWP model with 50 members, not post-processed.
- MC Dropout performs better on deterministic score since it has directly been optimised for MSE.
- However lacks variability in predictions, i.e., is overconfident
- Parametric and Categorical perform reasonably well even though optimised on probabilistic scores



Spread-Skill Ratio

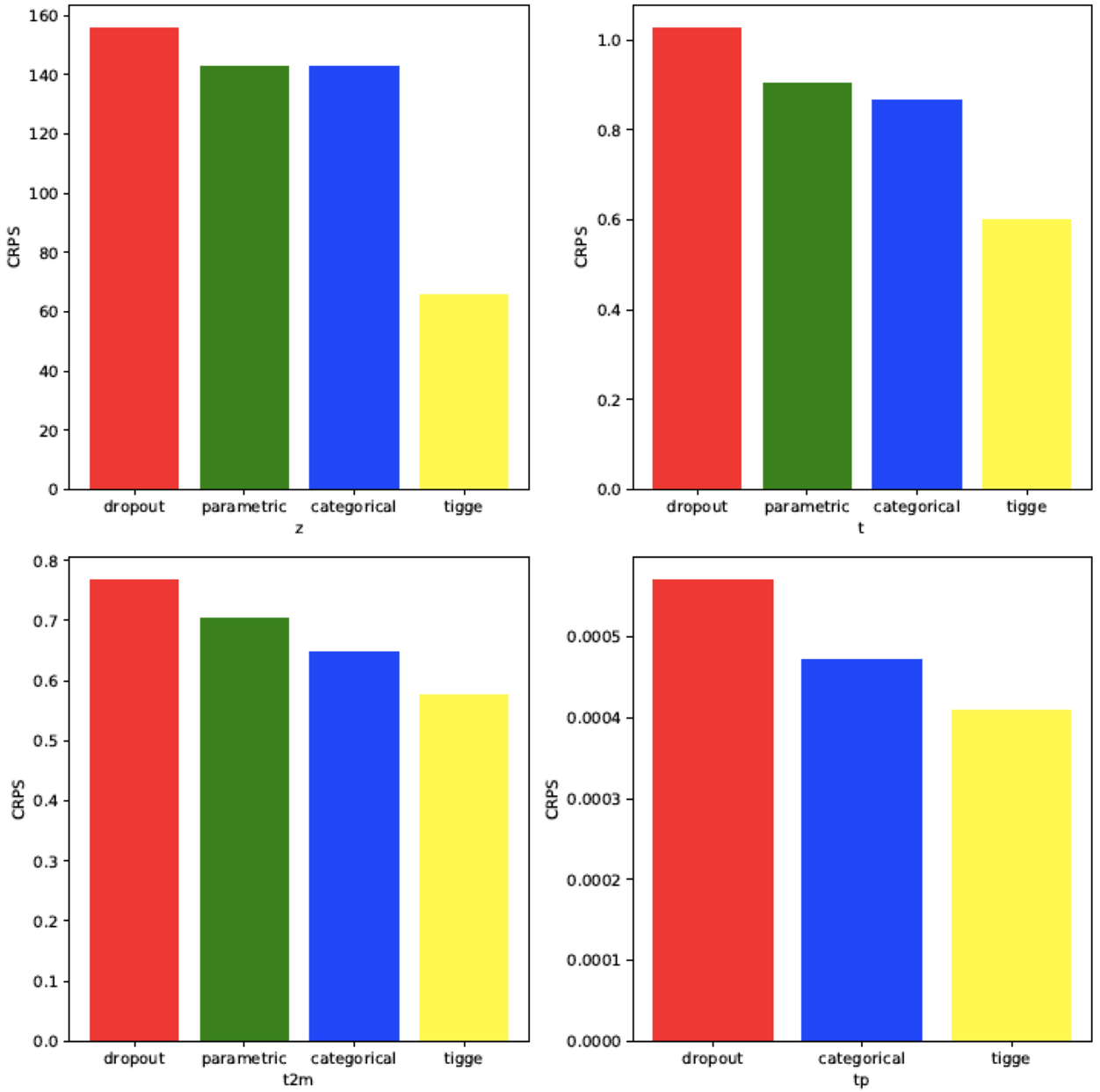
- MC Dropout has low spread-skill ratio, seems to evolve independent of RMSE

- Parametric and Categorical perform well, reaching >0.9 in some cases

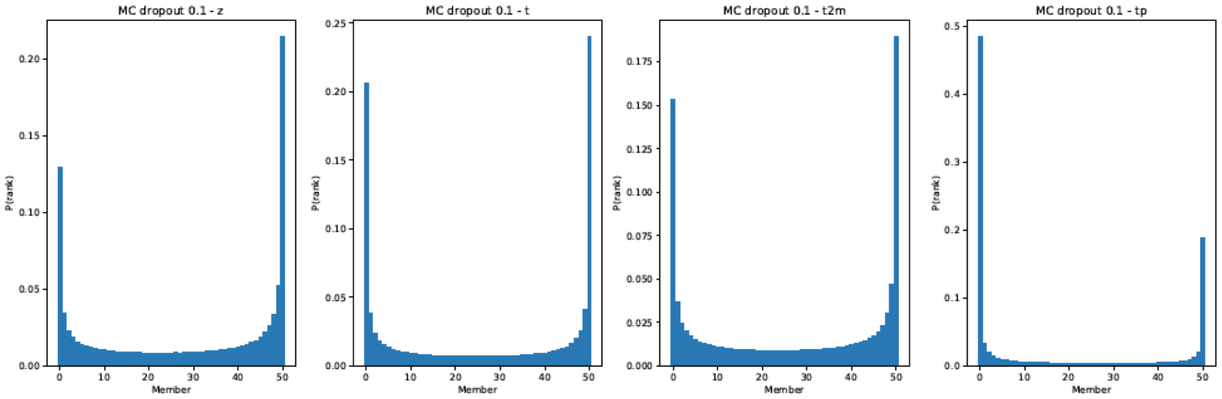


Continuous Ranked Probability Score (CRPS)

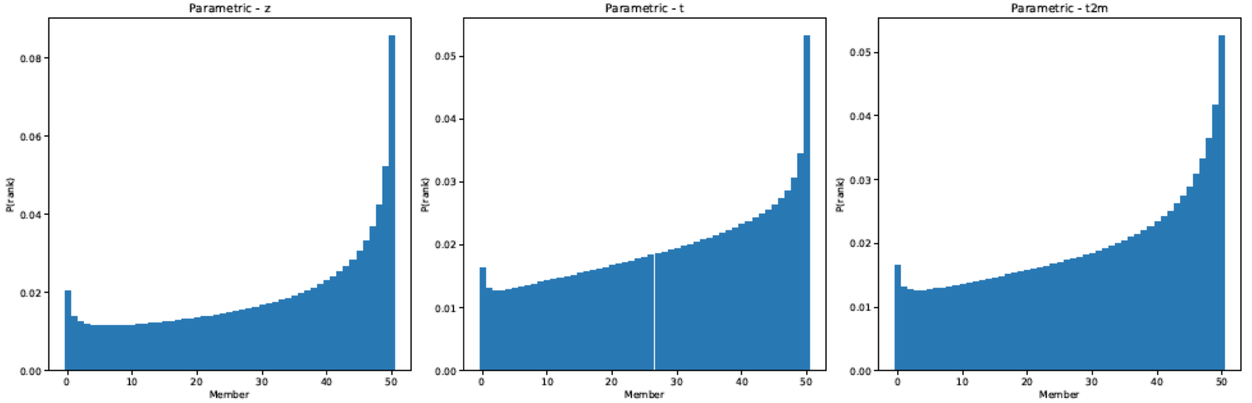
- Parametric forecast shows advantage of using a Probabilistic Score (CRPS) as Loss function
- Performance depends on how well the function represents data (T, T2M better than Z)
- Categorical has surprisingly good performance, and is much easier to implement for all kinds of distribution



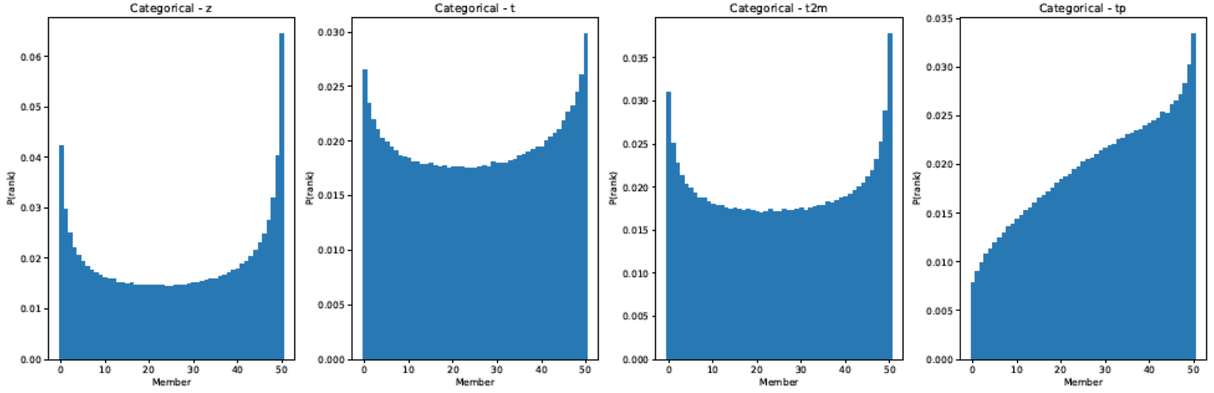
Rank Histograms



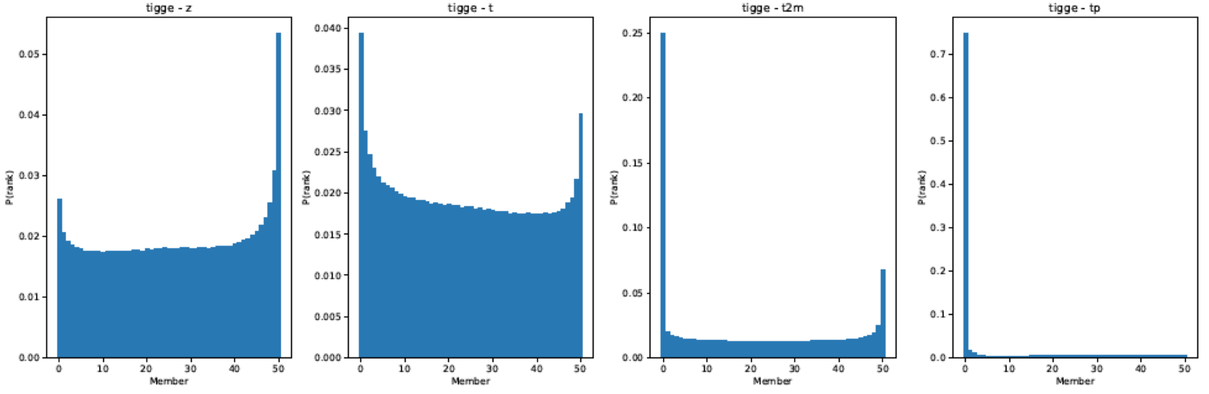
MC Dropout- Largely overconfident



Parametric- Mostly uniform, although simple bias

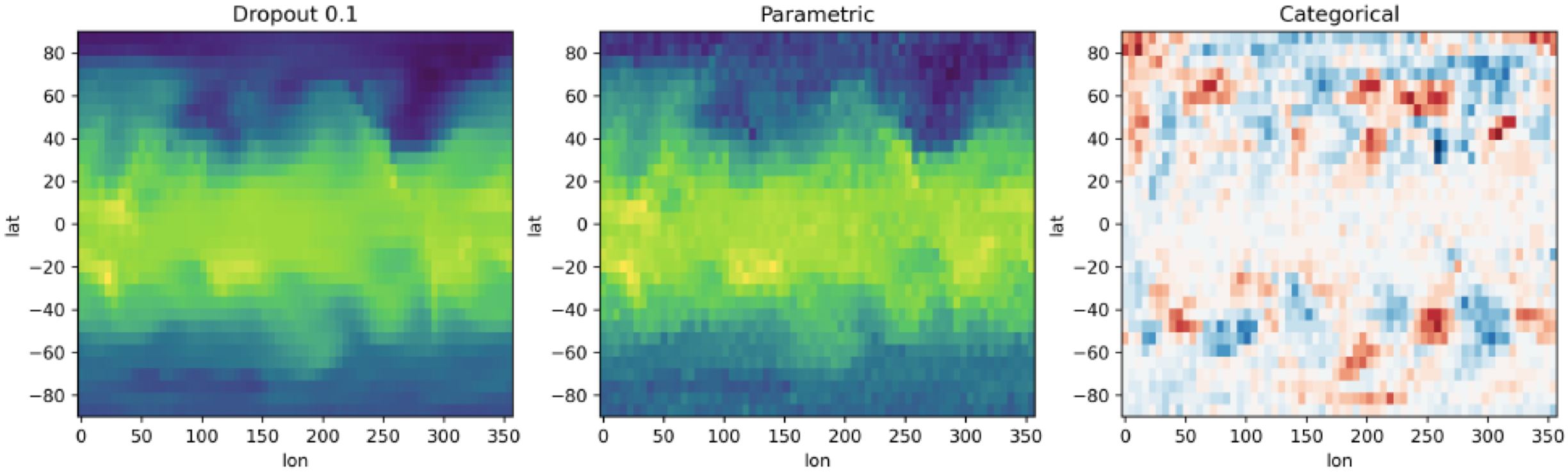


Categorical- Somewhat uniform, worse near extra-tropics & polar regions. Negative bias in TP



Tigge - Mostly uniform. ‘Drizzle bias’ in TP

Global Weather Maps



- MC Dropout- Smooth realisation with no extremes. Predicts ‘mean’ behaviour

- Parametric, Categorical - Non-smooth. Each grid-point sampled individually with no covariance learned. Good for local-scale predictions only

Conclusions

- MC Dropout is an improvement over deterministic but remains overconfident. Better implementations, specifically for ResNet, now available.

- Parametric forecast shows advantages of using a Probabilistic score as loss. Performance and ease of implementation dependent upon how well data distribution is represented by a continuous function.

- Categorical forecast is easy to implement, highly flexible, expressive for all data distributions. Performance dependent upon proper bin-size choice.

Purely data-driven models show reasonable skill and can prove useful for weather research. They should integrate knowledge from Meteorology in order to build better uncertainty-aware predictions.

Future Outlook

- Purely data-driven models won't have enough training data to compete with NWP models for global medium-range weather forecasting (~10^{30} years) Useful in nowcasting, subseasonal forecasts, local forecasts, and many more.
- As complimentary tool, useful for specific NWP stages: initial conditions generations, parametrization schemes, post-processing.
- Possible Improvements:
  - Pre-training on coarse simulation models, fine-tuning on observations.
  - Better boundary conditions: cube-sphered geometry, spherical geometry
  - Architectures: Using spatial and temporal indicators such as with U-Nets, RNNs. For realistic realizations, conditional GANs might perform well

References

- P. Bauer, A. Thorpe, and G. Brunet. “The quiet revolution of numerical weather prediction.” In: Nature 525.7567 (2015), pp. 47–55.
- S. Rasp, P. D. Dueben, S. Scher, J. A. Weyn, S. Mouatadid, and N. Thuerey. “WeatherBench: A benchmark dataset for data-driven weather forecasting.”
- Y. Gal and Z. Ghahramani. “Bayesian convolutional neural networks with Bernoulli approximate variational inference.” In: arXiv preprint arXiv:1506.02158 (2015).
- C. K. Sønderby, L. Espeholt, J. Heek, M. Dehghani, A. Oliver, T. Salimans, S. Agrawal, J. Hickey, and N. Kalchbrenner. “MetNet: A Neural Weather Model for Precipitation Forecasting.” In: arXiv preprint arXiv:2003.12140 (2020).
- T. M. Hamill. “Interpretation of rank histograms for verifying ensemble forecasts.” In: Monthly Weather Review 129.3 (2001), pp. 550–560 • T. Palmer. “A Vision for Numerical Weather Prediction in 2030.” In: arXiv preprint arXiv:2007.04830 (2020).