Comparative Forecast Verification: Testing the Frequency of Better

Eric Gilleland and Domingo Muñoz-Esparza
Research Applications Laboratory
National Center for Atmospheric Research
David D. Turner
Global Systems Laboratory,
National Oceanographic and Atmospheric Association (NOAA)

January 11, 2023

NCAR | RESEARCH APPLICATIONS LABORATORY

DTC Developmental Testbed Center

NSF

This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977.

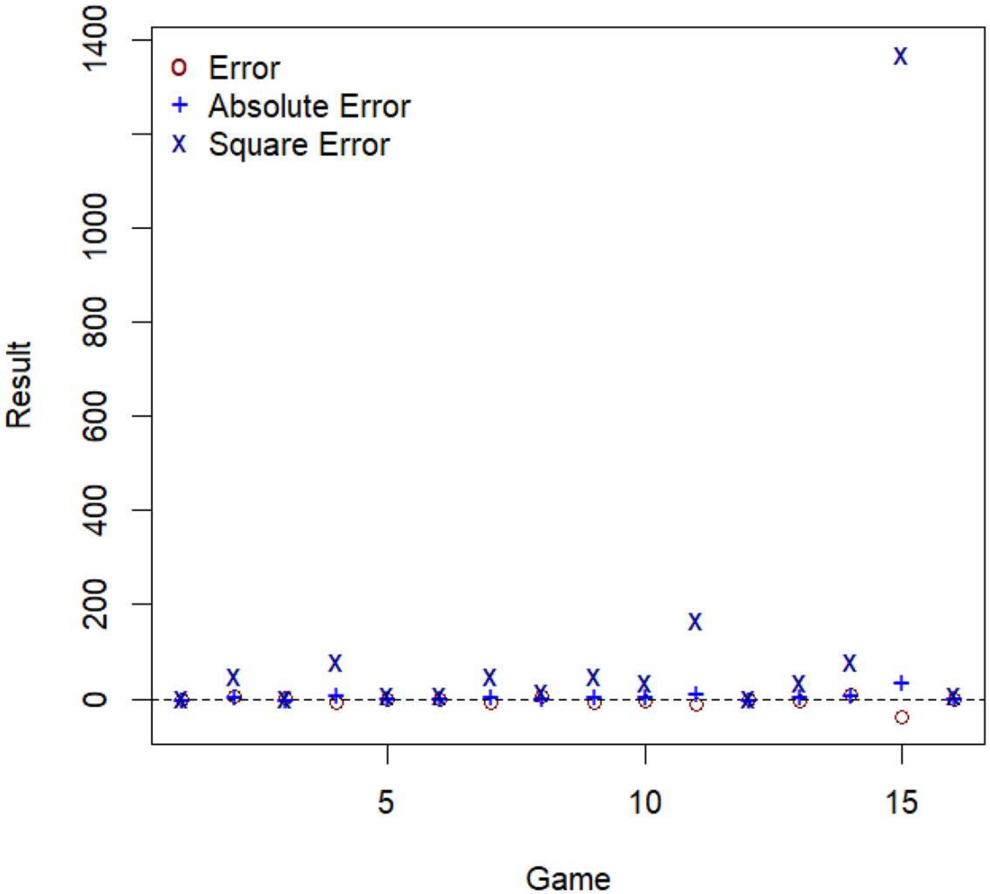
Brief Review of Statistical Hypothesis Testing

Competing Forecast Verification Setting

- Want to know if model A is better than model B.
- Assume neither is better than the other (null hypothesis, denoted $\mathcal{H}_0$).
- Calculate a test statistic (e.g., RMSE, MAE, etc.).
- Determine how likely it is to observe a test statistic as extreme as the one observed above (typically using assumptions like independence and identically distributed data, normality, etc.).
- Is it likely that model A is the same as model B based on the test statistic?
  - Yes! Fail to reject $\mathcal{H}_0$
  - No. Reject $\mathcal{H}_0$
- We could be wrong in two ways (uncertainty):
  - Type I error: Reject $\mathcal{H}_0$ when it is actually true (think convicting someone of murder when they didn't really do it!)
    - The ***size*** of a test is the probability of a type I error.
  - Type II error: Fail to reject $\mathcal{H}_0$ when it is not true (the murderer goes free)
    - The ***power*** of a test is the probability of detecting a true effect.
- A statistical test is only one piece of evidence!
- Cassie Kozyrkov has some very nice videos online that explain these concepts very well (e.g., using puppies). Just do a web search for her name and something like p-values.

Loss functions

2022 Denver Broncos



Record to date: 4 - 12

Score	Error	AE	SE
16-17	-1	1	1
16-9	3	3	9
11-10	1	1	1
23-32	-9	9	81
9-12	-3	3	9
16-19	-3	3	9
9-16	-7	7	49
21-17	4	4	16
10-17	-7	7	49
16-22	-6	6	36
10-23	-13	13	169
9-10	-1	1	1
28-34	-6	6	36
24-15	9	9	81
14-51	-37	37	1,369
24-27	-3	3	9
Mean	-4.6875	7.3125	11.08208

Root mean-square error (RMSE)

Power-divergence Statistic

Modeling discrete multivariate data

- Model A is better than model B or model B is better ($k = 2$ categories) according to some loss function
- Let X be the random variable where if model A is better, then $X = 1$ and if not, $X = 0$.
- Then $X \sim Binom(p)$, where p is the probability that $X = 1$, so $1 - p$ is the probability that $X = 0$.
- Want to test $\mathcal{H}_0: p = \frac{1}{2}$ meaning that model A and model B have the same frequency of being better than the other (i.e., neither model is better).
- More generally, the test is $\mathcal{H}_0: p = q$, where $q = \frac{1}{2}$ for our setting.

Power-divergence Statistic

$$I^{\lambda}(\widehat{\boldsymbol{p}}:\boldsymbol{q}) = \frac{1}{\lambda(\lambda+1)}\sum_{i=1}^{k}\hat{p}_i\left[\left(\frac{\hat{p}_i}{q_i}\right)^{\lambda} - 1\right]$$

where for our setting:

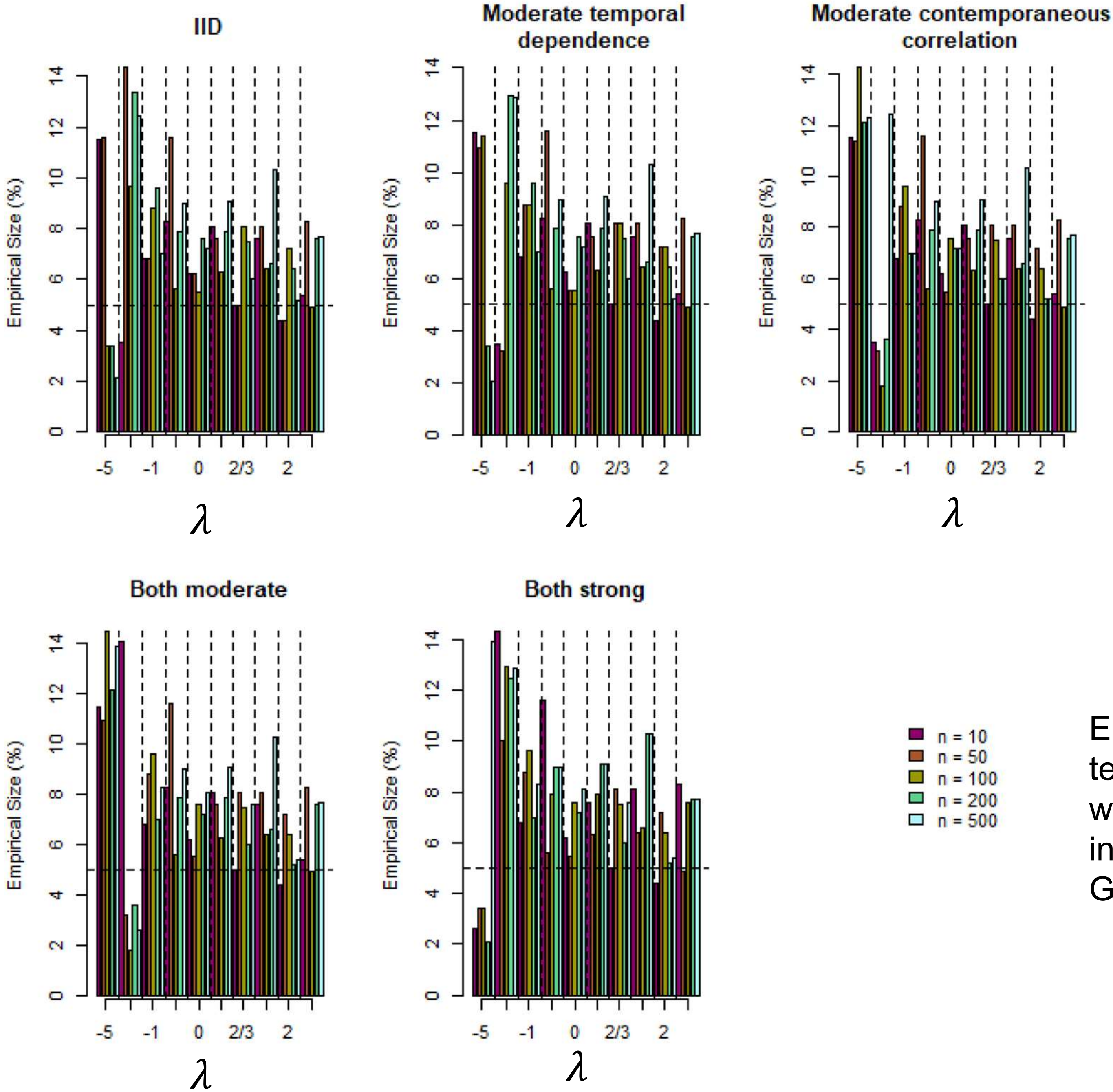
- $k = 2$
- $\widehat{\boldsymbol{p}} = (\hat{p}_1, \hat{p}_2) = (\hat{p}, 1 - \hat{p})$ is the estimate of p from the data
- $\boldsymbol{q} = (q_1, q_2) = (q_1, q_2\ ) = \left(\frac{1}{2}, \frac{1}{2}\right)$ is the vector of test parameters
- $\boldsymbol{\lambda}$ is a user-chosen value that yields different test statistics, but…
- asymptotically, they are all the same!
- Under certain assumptions that are not likely to be met with atmospheric data, $I^{\lambda}(\widehat{\boldsymbol{p}}:\boldsymbol{q}) \sim \chi^2_{k-1}$

Power-divergence Statistic

Statistic Name	λ	Definition	Notes
Neyman Modified X^2	$\lambda = -2$	$N^2 = \sum_{i=1}^{k} \frac{\hat{p}_i - q_i}{\hat{p}_i}$	Neyman (1949)
Kullback-Leibler	$\lambda = -1$	$KL = 2\sum_{i=1}^{k} q_i \log\left(\frac{q_i}{\hat{p}_i}\right)$	Kullback and Leibler (1951)
Freeman-Tukey	$\lambda = -\frac{1}{2}$	$F^2 = 4\sum_{i=1}^{k} \left(\sqrt{\hat{p}_i} - \sqrt{q_i}\right)^2$	Freeman and Tukey (1950)
Loglikelihood-ratio	$\lambda = 0$	$G^2 = 2\sum_{i=1}^{k} \hat{p}_i \log\left(\frac{\hat{p}_i}{q_i}\right)$	Optimal for testing against certain nonlocal alternatives with some near-zero probabilities. Neyman (1949)
Cressie-Read	$\lambda = \frac{2}{3}$	$CR = \frac{9}{5}\sum_{i=1}^{k} \hat{p}_i \left[\left(\frac{\hat{p}_i}{q_i}\right)^{2/3} - 1\right]$	A good choice when there is no knowledge of possible alternative models for both small and large sample sizes. Cressie and Read (1984)
Pearson's X^2	$\lambda = 1$	$X^2 = \sum_{i=1}^{k} \frac{(\hat{p}_i - q_i)^2}{q_i}$	Optimal for the equiprobable hypothesis against certain local alternatives in large sparse tables. Pearson (1900)

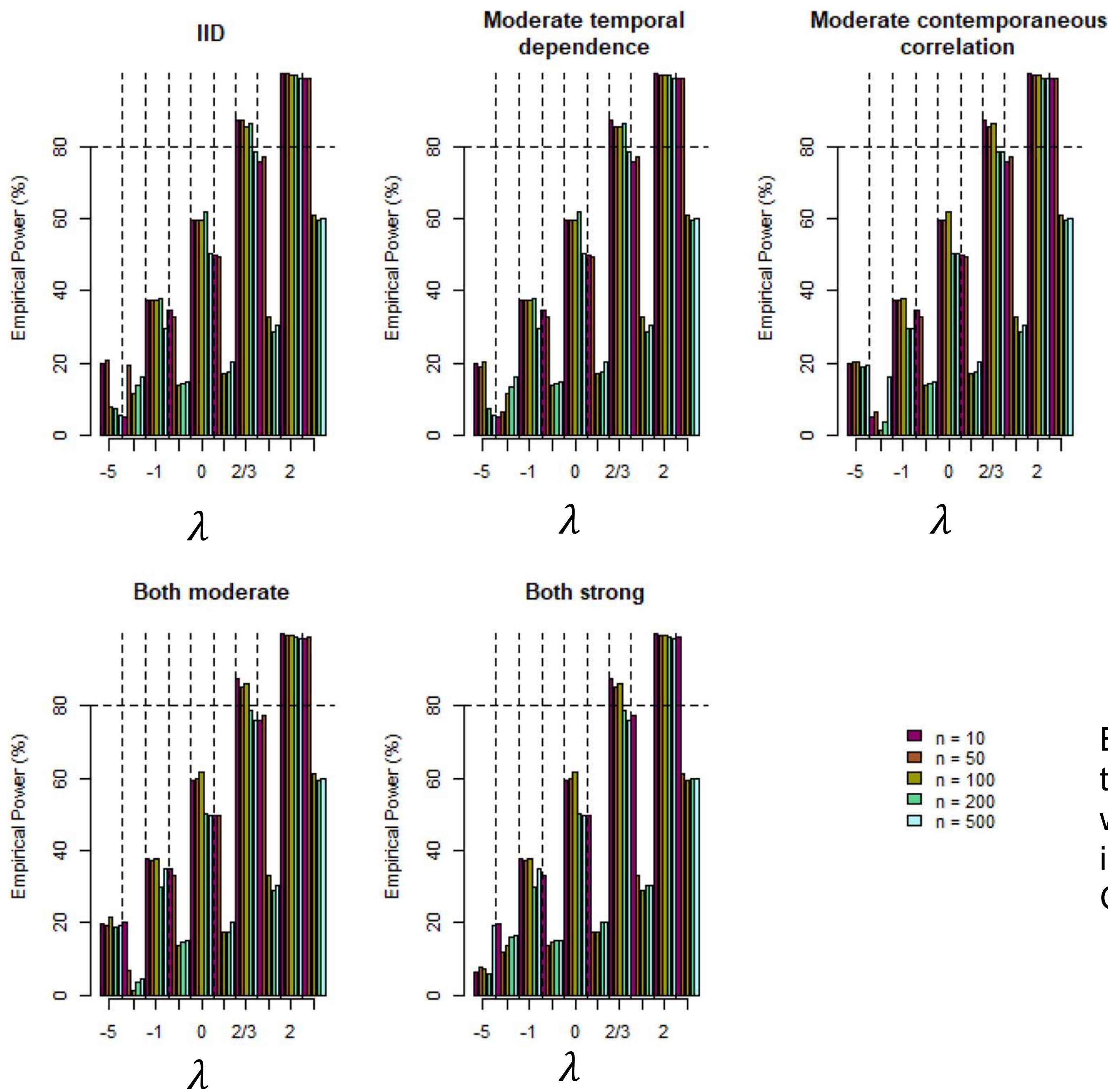
Above table is taken from Table 1 in Gilleland et al., (submitted). *And is a summary of some information taken from: Read and Cressie (1988).*

Power-divergence Statistic



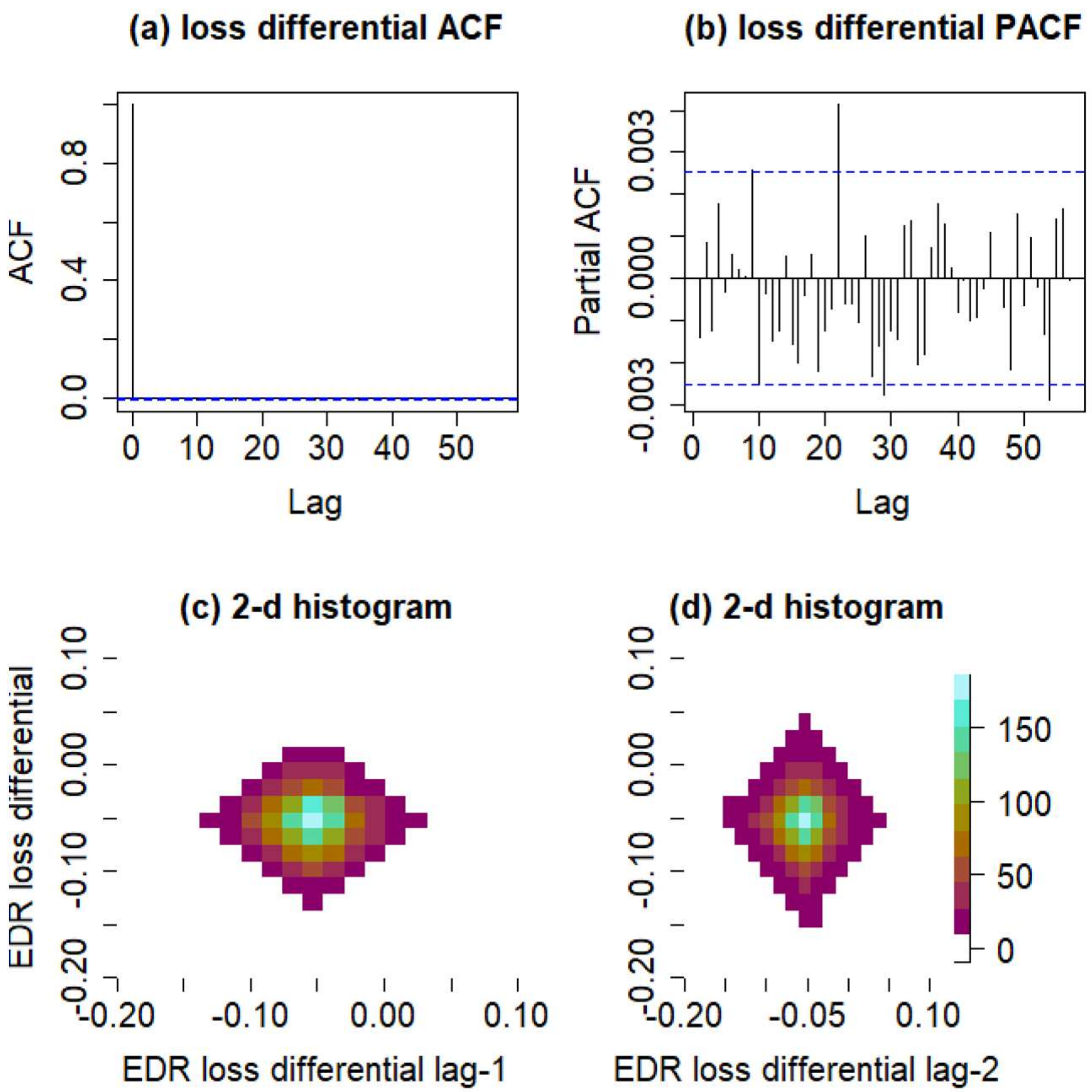
Empirical Size testing (using 5%) with simulations as in Hering and Genton (2011)

Power-divergence Statistic



Empirical Power testing (using 5%) with simulations as in Hering and Genton (2011)

Test Cases: Turbulence

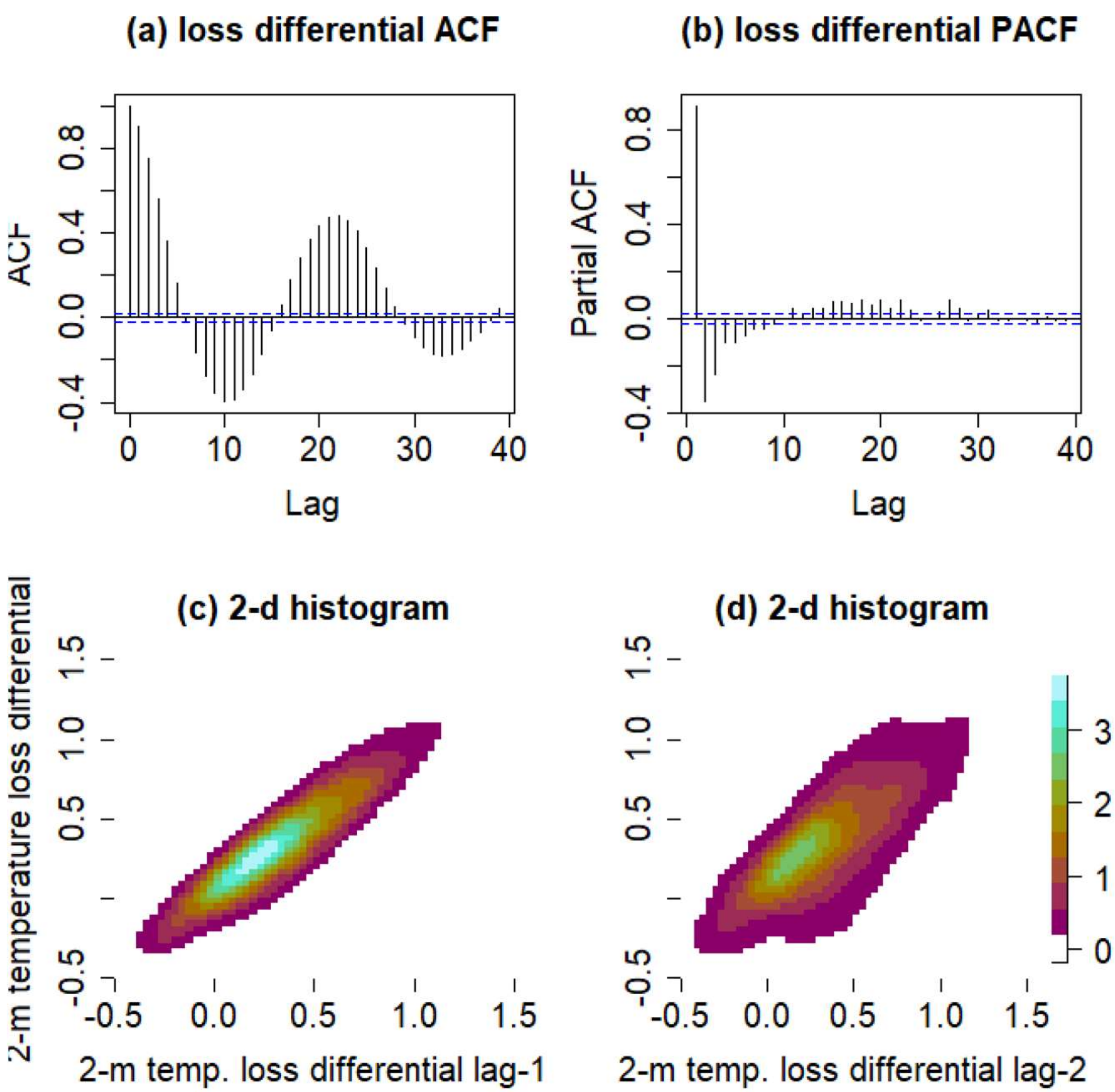


Two versions of 6-h turbulence forecasts called the Graphical Turbulence Guidance (GTG) algorithm for eddy dissipation rate (EDR, $\mathrm{m}^{2/3}\mathrm{s}^{-1}$, Sharman and Pearson 2017; Muñoz-Esparza and Sharman 2018; Muñoz-Esparza et al. 2020).

These turbulence forecasts use v. 3 of the High-Resolution Rapid Refresh (HRRR, Dowell et al. 2022; James et al. 2022) as the input NWP information for the 1 June 2018 to 30 September 2019 period.

Competing versions are: simple regression (HGTG, Sharman and Pearson 2017) and a machine-learning model based on regression trees (ML GTG, Muñoz-Esparza et al. 2020).

Test Cases: HRRR Temperature and Wind Speed



12-h forecasts of 2-m temperature (deg. C) extracted from the surface application of the Model Analysis Tool Suite (MATS, Turner et al. 2020). Comparing HRRR v. 3 and v. 4.

Matched observations are used with model forecast data from 1 August 2019 to 1 December 2020 when v. 3 of HRRR was operational at NCEP and v. 4 frozen as part of the evaluation phase.

Also looked at 10-m wind speed (m/s), which produces similar diagnostic plots as these, so not shown for brevity.

Test Cases: Turbulence

Moderate turbulence conditions: 0.1 $m^{2/3}s^{-1}$< EDR < 0.3$m^{2/3}s^{-1}$

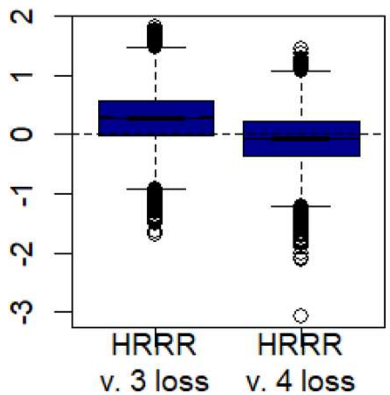
λ	-5	-2	-1	$-1/2$	0	$1/2$	$2/3$	1	2	5
ME										
Power div.	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34
p-value	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56

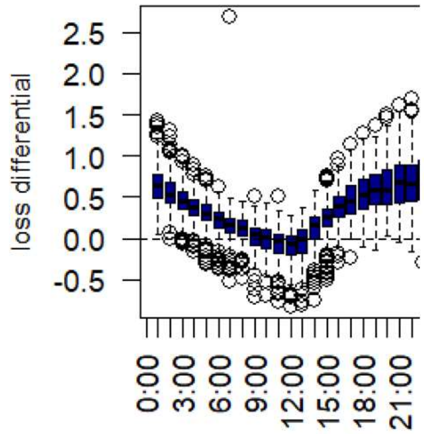
Severe turbulence conditions: EDR > 0.3$m^{2/3}s^{-1}$, which is about 0.1% of the total sample.

λ	-5	-2	-1	$-1/2$	0	$1/2$	$2/3$	1	2	5
ME										
Power div.	11.99	11.45	11.34	11.30	11.27	11.25	11.25	11.24	11.24	11.44
p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

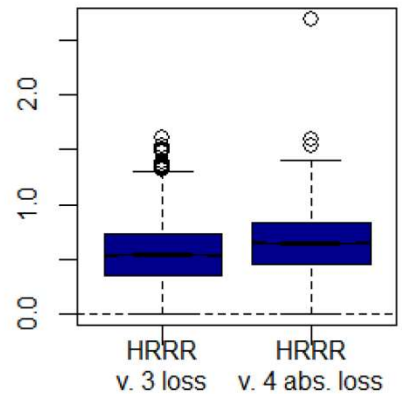
Test Cases: HRRR Temperature and Wind Speed

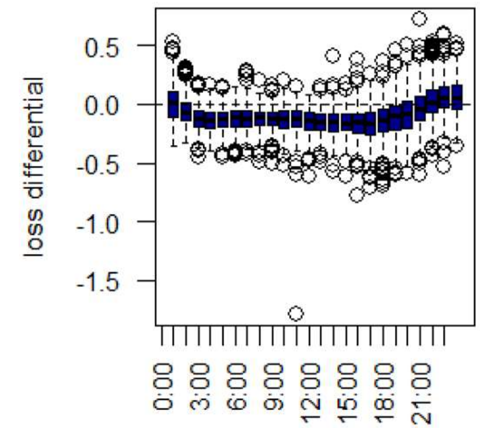
12-h forecasts of 2-m temperature (deg. C)

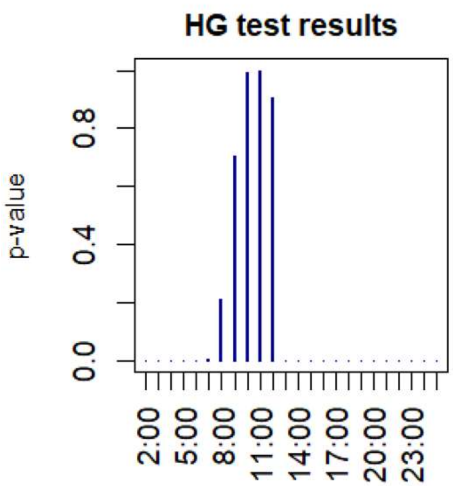


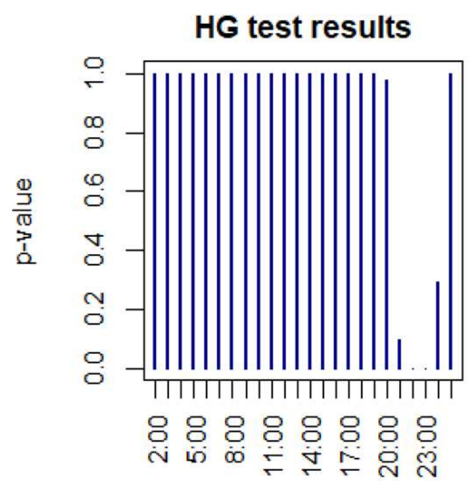


12-h forecasts of 10-m wind speed (m/s)





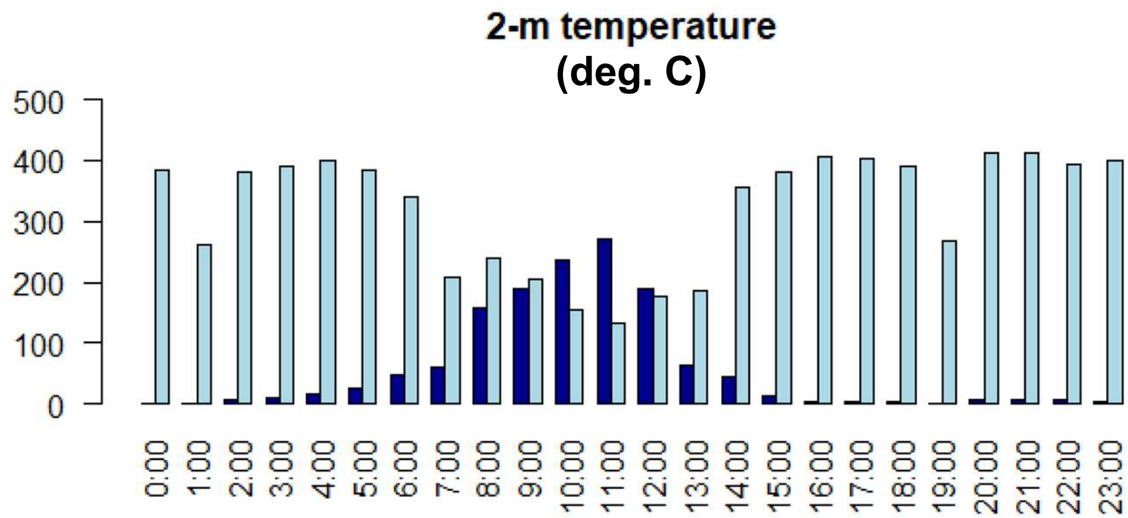




The Hering-Genton test (Hering and Genton 2011) is a t-test on the mean loss differential where the standard error is estimated in a way that accounts for temporal dependence, and the test is robust to contemporaneous correlation. It is a test on the intensity difference in error rather than the frequency of being better.

Test Cases: HRRR Temperature and Wind Speed

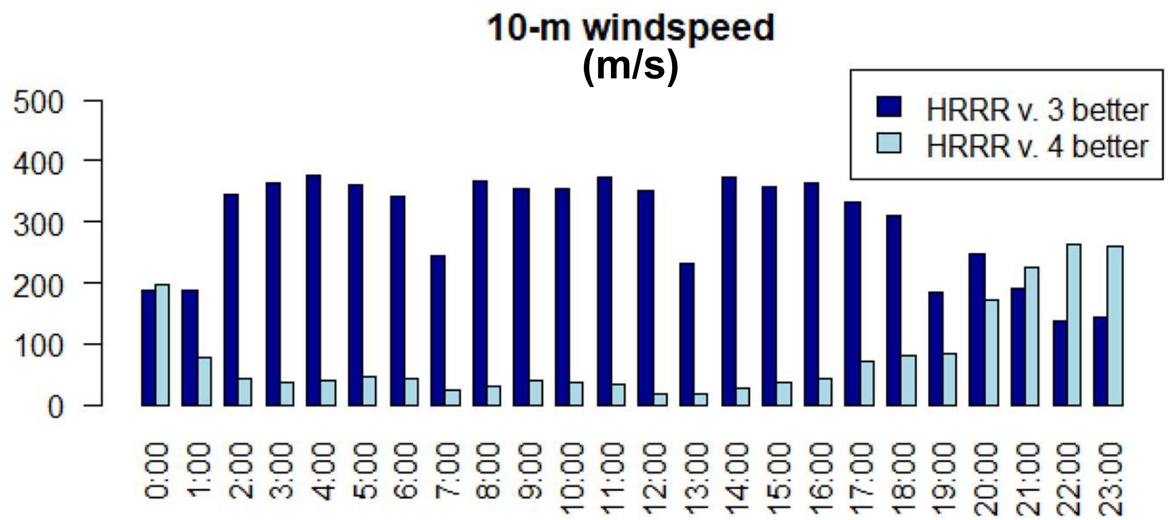
For all choices of λ applied previously, the power-divergence rejects $\mathcal{H}_0$ at all times except at 9 and 12 UTC



Using $\lambda = 2/3$, $\mathcal{H}_0$ is rejected at all time points.

For large negative λ the test fails to reject $\mathcal{H}_0$, where all of the choices of λ above -1, the test rejects $\mathcal{H}_0$.

Results based on a 5%-level test, but p-values estimated to be zero.



References

Cressie, N. A. C., and T. R. C. Read (1984) Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society Series B*, **46**, 440 – 464, doi: 10.1111/j.2517-6161.1984.tb01318.x.

Dowell et al. (2022) The High-Resolution Rapid Refresh (HRRR): An hourly updating convection-allowing forecast model. part 1: Motivation and system description. *Weather and Forecasting*, **37**, 1371 – 1396, doi: 10.1175/WAF-D-21-0151.1;

Freeman, M. F. and J. W. Tukey (1950) Transformations related to the angular and the square root. *Annals of Mathematical Statistics*, **21**, 607 – 611, doi: 10.1214/aoms/1177729756.

James et al. (2022) An hourly updating convection-allowing forecast model. Part 2: Forecast performance. *Weather and Forecasting*, **37**, 1397 – 1417, doi: 10.1175/WAF-D-21-0130.1

Gilleland, E. D. Muñoz-Esparza, and D. Turner (2023) "Competing forecast verification: Using the power-divergence statistic for testing the frequency of "better"." Submitted to *Weather and Forecasting* on 16 November 2022.

Hering and Genton (2011) Comparing spatial predictions. Technometrics, **53**, 414 – 425, doi:10.1198/TECH.2011.10136.

Kullback, S. and R. A. Leibler (1951) On information and sufficiency. *Annals of Mathematical Statistics*, **22** (1), 79 – 86, doi: 10.1214/aoms/1177729694.

Muñoz-Esparza and Sharman (2018) An improved algorithm for low-level turbulence forecasting. *Journal of Applied Meteorology and* Climatology, **57**, 1249 – 1263, doi: 10.1175/JAMC-D-17-0337.1.

Muñoz-Esparza, D., R. D. Sharman, and W. Deierling (2020) Aviation turbulence forecasting upper levels with machine learning techniques based on regression trees. *Journal of Applied Meteorology and* Climatology, **59**, 1883 – 1889, doi: 10.1175/JAMC-D-20-0116.1.

Neyman, J. (1949) Contribution to the theory of the χ^2 test. *Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability,* 239 – 273.

Pearson, K. (1900) On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophy Magazine*, **50**, 157–172, doi: 10.1007/978-1-4612-4380-9_2.

Read and Cressie, 1988. Goodness-of-Fit Statistics for Discrete Multivariate Data. Springer-Verlag, New York, NY, 211 pp.

Sharman, R. and J. Pearson (2017) Prediction of energy dissipation rates for aviation turbulence. Part I: Forecasting nonconvective turbulence. *Journal of Applied Meteorology and* Climatology*,* **56***, 317 – 337,* doi: 10.1175/JAMC-D-16-0205.1.

Turner et al. (2020) A verification approach used in developing the Rapid Refresh and other numerical weather prediction models. *J. Oper. Meteorol.*, **8**, 39 – 53, doi: 10.15191/nwajom.2020.0803.