Testing for the frequency of “better”

Eric Gilleland
Research Applications Laboratory
National Center for Atmospheric Research

38th Session of the Working Group on Numerical Experimentation
(WGNE-38; 27 November – 2 December 2023)

28 November 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.

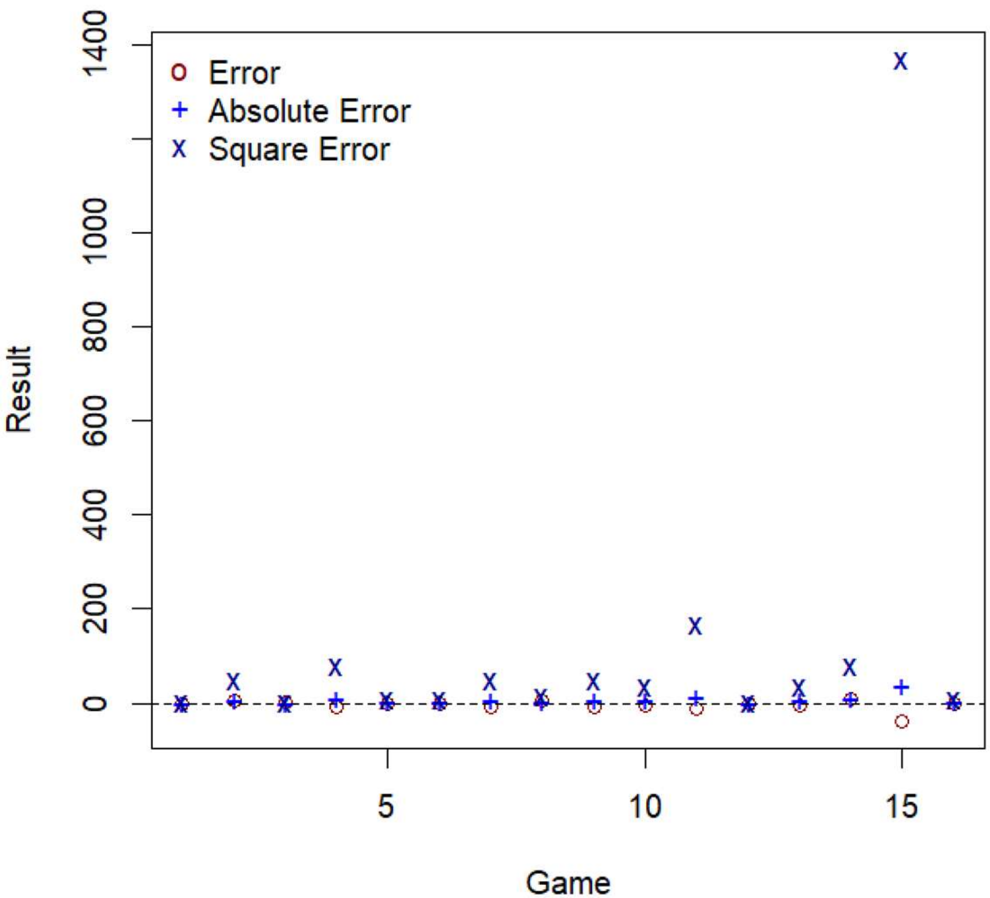
Statistical Hypothesis Testing

This talk covers the recently published paper:

Gilleland, E. D. Muñoz-Esparza, and D. Turner (2023) “Competing forecast verification: Using the power-divergence statistic for testing the frequency of “better”. *Weather and Forecasting*, **38** (9), 1539 – 1552, doi: 10.1175/WAF-D-22-0201.1.

Loss functions

2022 Denver Broncos



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

Record to date: 4 - 12

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 $q = \frac{1}{2}$ here.

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
- $\boldsymbol{q} = (q_1, q_2) = (q, 1 - q\ ) = \left(\frac{1}{2}, \frac{1}{2}\right)$ 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)

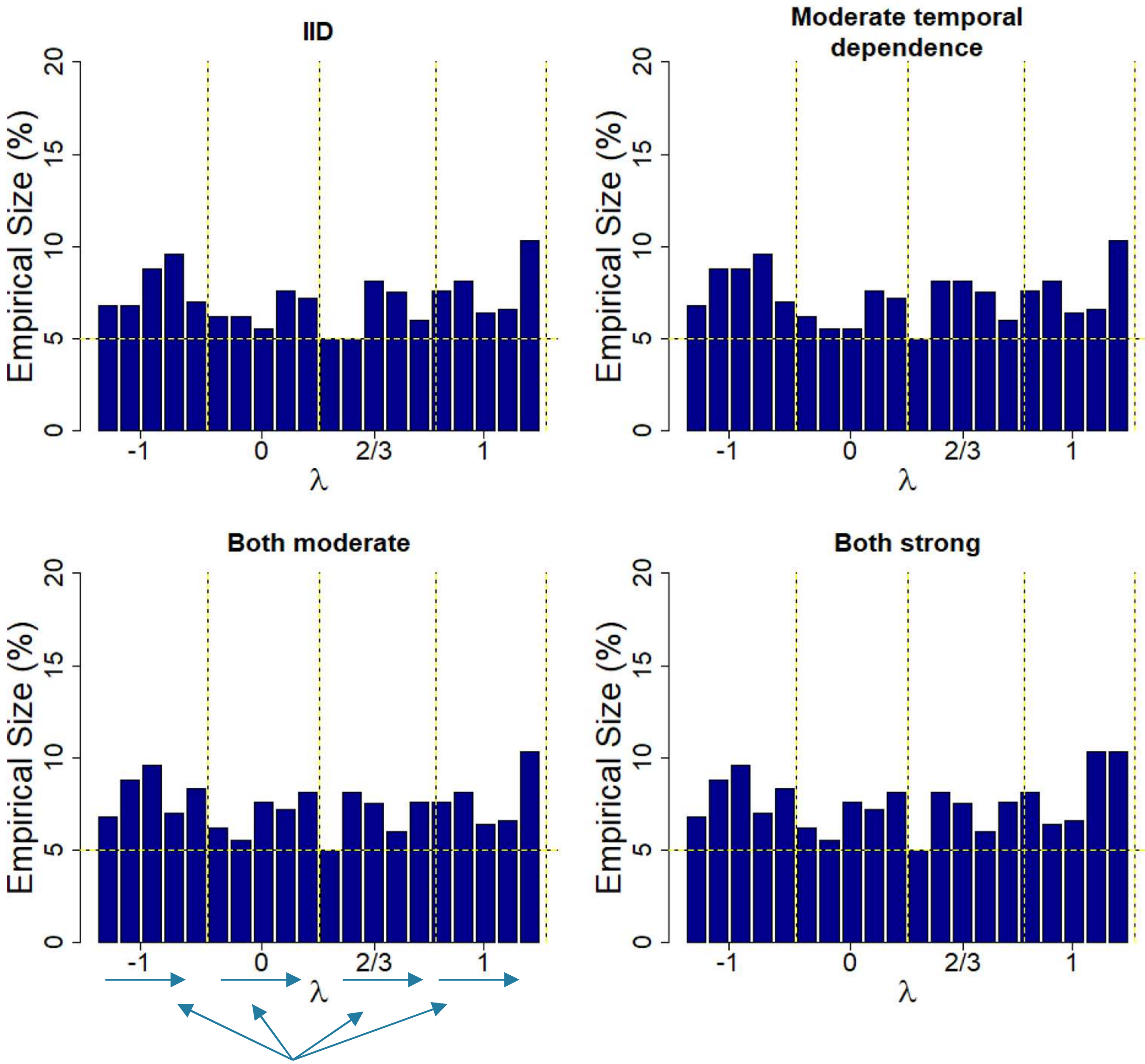
Above table is taken from Table 1 in Gilleland et al., (accepted to WAF). *And is a summary of some information taken from: Read and Cressie (1988).*

Simulation Experiment to test different hypothesis tests

Competing Forecast Verification Setting

- Simulate two time series of errors, $\varepsilon_A(t)$ and $\varepsilon_B(t)$, with
  - the same mean, $\mu_A = \mu_B = 0$, and with either
  - the same variances, $\sigma_A^2 = \sigma_B^2 = \sigma^2$ to empirically test for the size of various hypothesis tests, or
  - with $\sigma_B^2 > \sigma_A^2$ to empirically test for the power of the tests.
- Apply power-divergence test to test $\mathcal{H}_0\colon q_A = q_B = 1/2$ against $\mathcal{H}_1\colon q_A \neq q_B$.
  - Could test other alternative hypotheses, but here the focus is on the two-sided alternative.
- Repeat the above steps 1000 times.
  - For empirical size (when $\sigma_A = \sigma_B$), find the number of times $\mathcal{H}_0$ is (falsely) rejected and divide by 1000. The result is the empirical size of the test.
  - For empirical power, find the number of times $\mathcal{H}_0$ is (correctly) rejected and divide by 1000. The result is the empirical power of the test.

Power-divergence Statistic

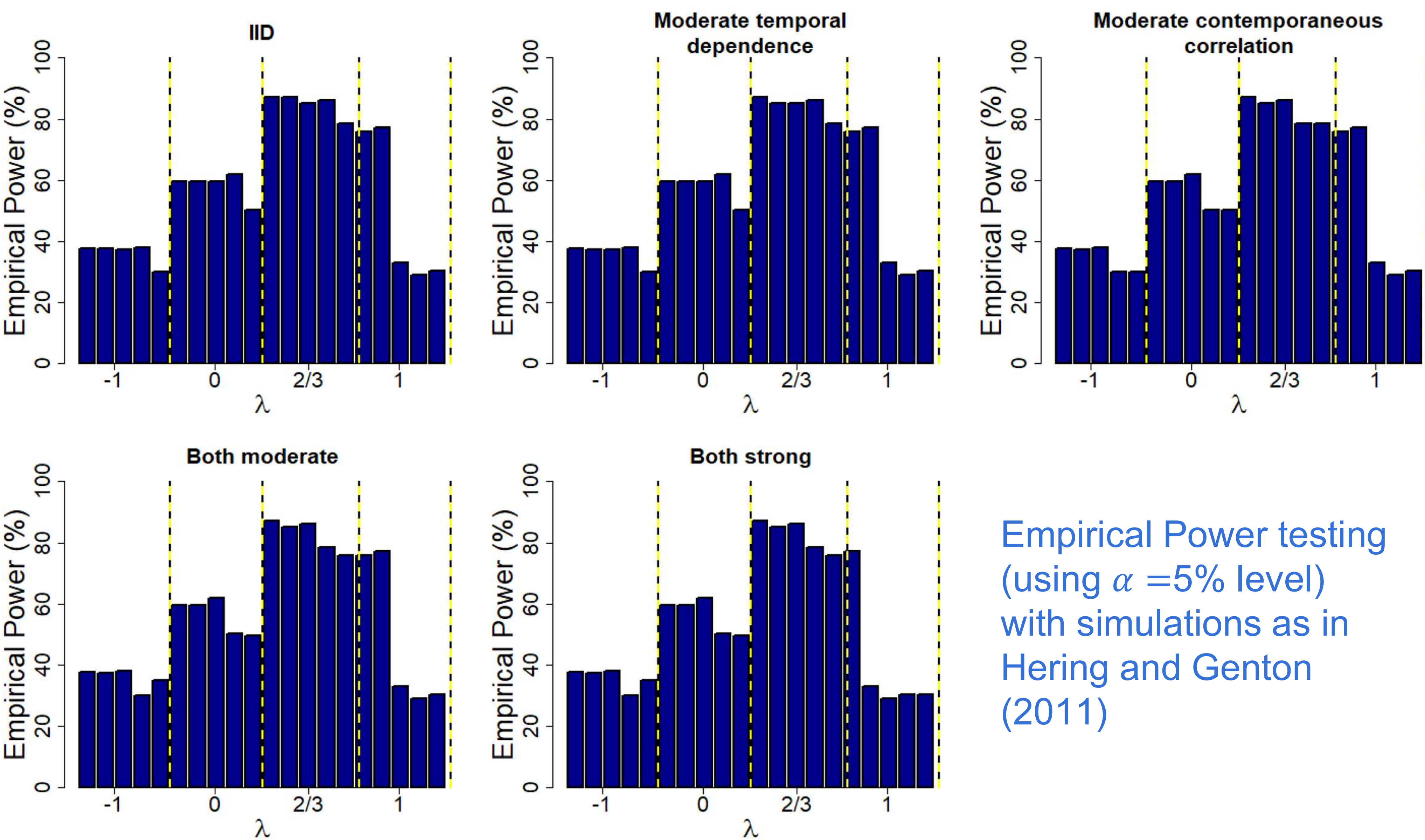


Increasing sample sizes

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

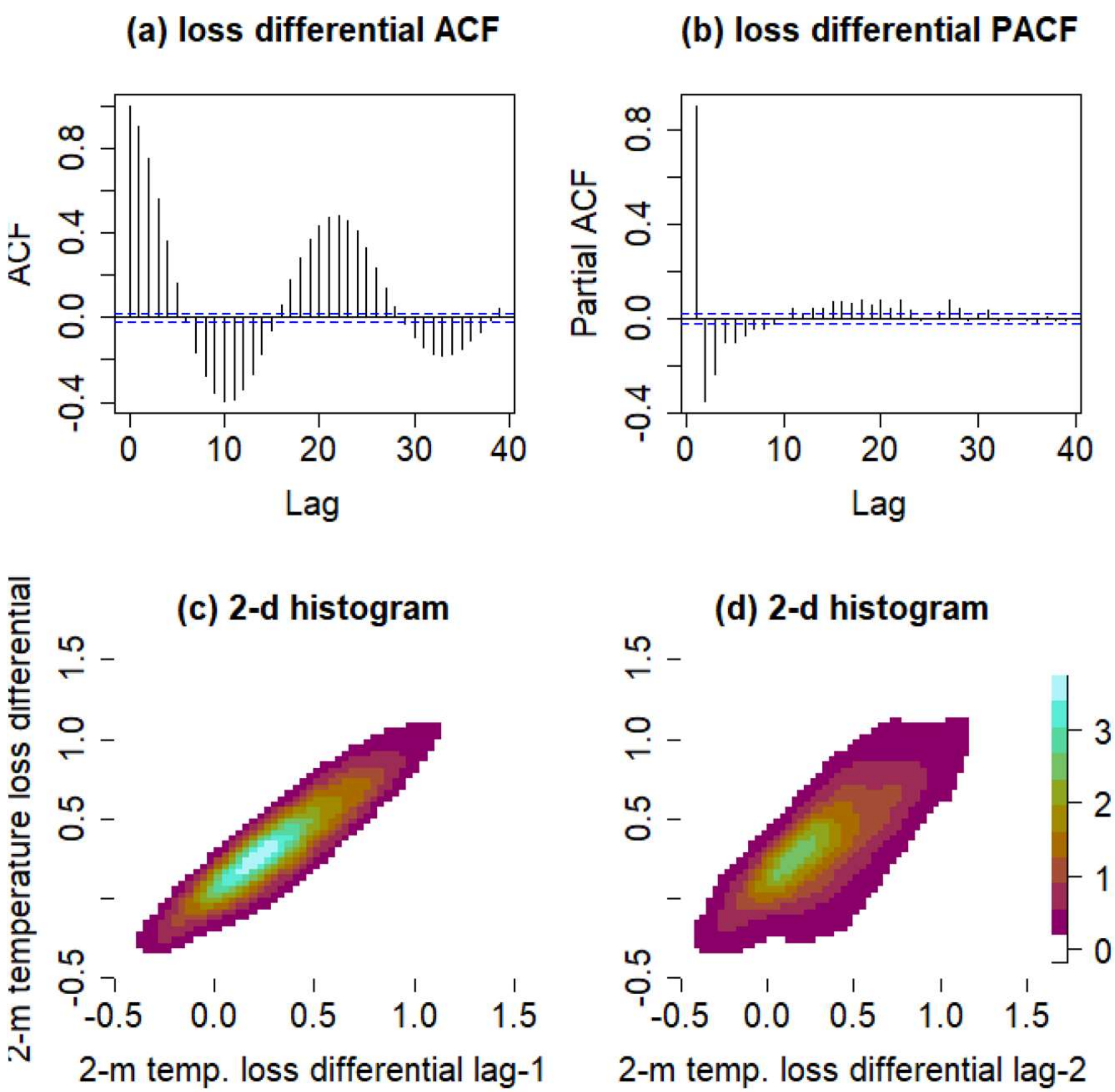
5% level

Power-divergence Statistic



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

Test Cases: HRRR Temperature and Wind Speed



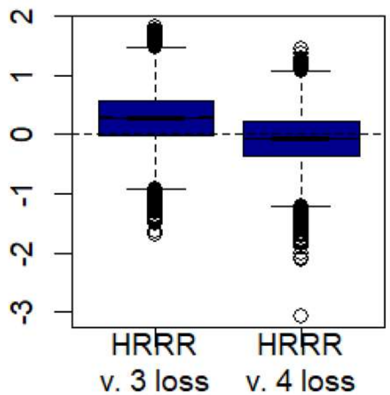
12-h forecasts of 2-m temperature (deg. C) extracted from the surface application of the Model Analysis Tool Suite (MATS, Turner et al. 2020). Comparing HRRR v. 3 and v. 4.

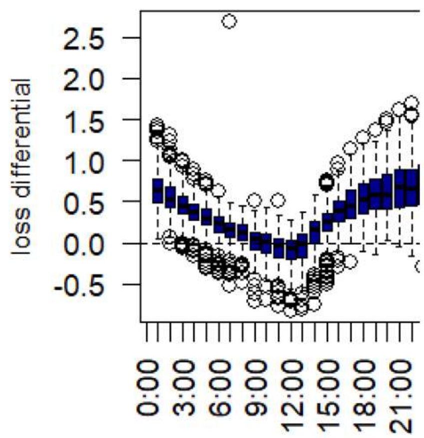
Matched observations are used with model forecast data from 1 August 2019 to 1 December 2020 when v. 3 of HRRR was operational at NCEP and v. 4 frozen as part of the evaluation phase.

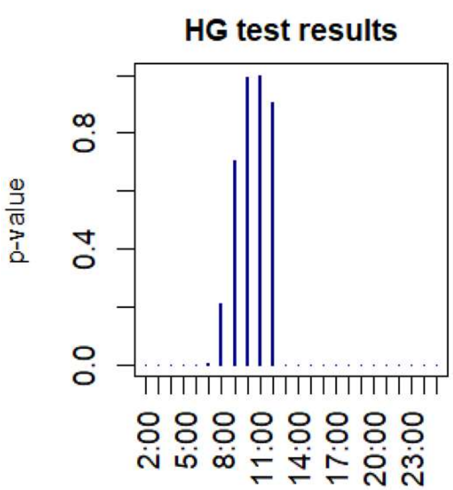
Also looked at 10-m wind speed (m/s), which produces similar diagnostic plots as these, so not shown for brevity.

Test Cases: HRRR Temperature and Wind Speed

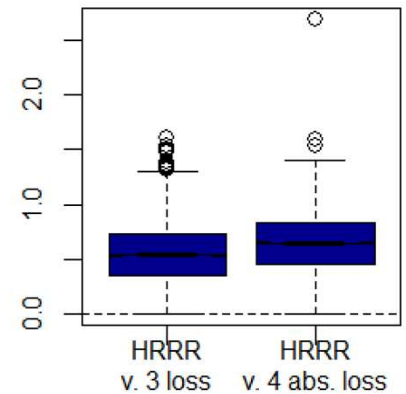
12-h forecasts of 2-m temperature (deg. C)

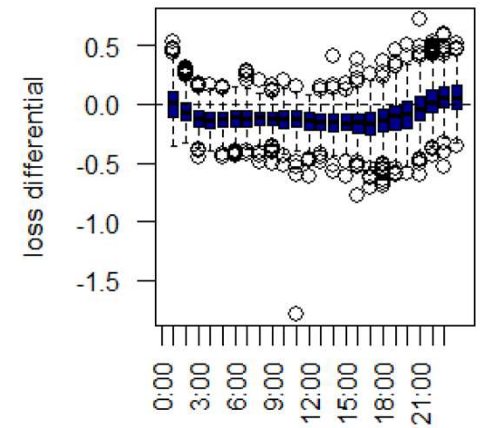


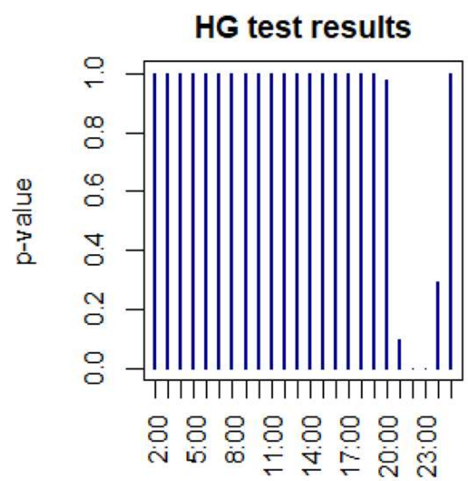




12-h forecasts of 10-m wind speed (m/s)







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: Turbulence

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

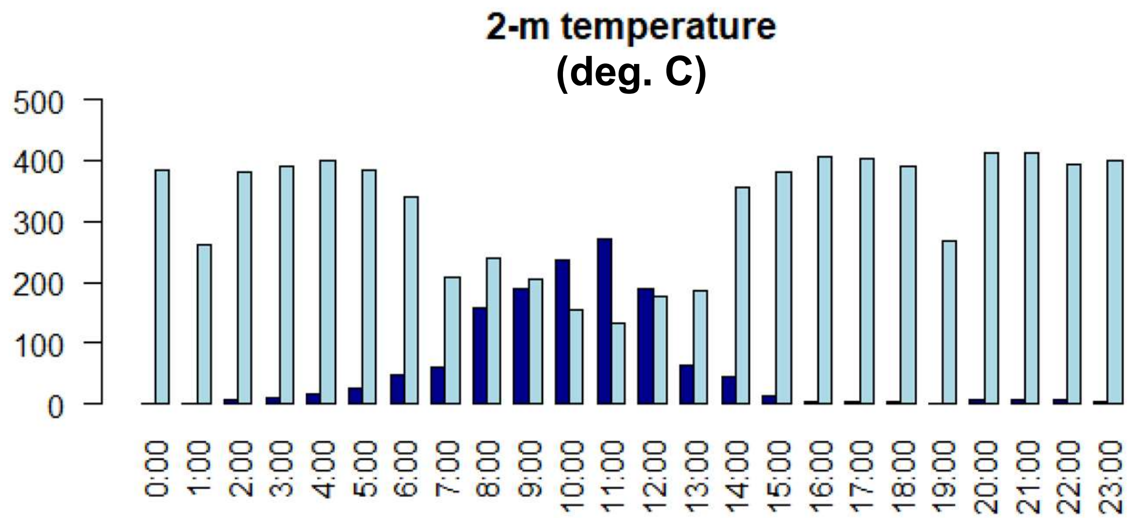
λ	-5	-2	-1	$-1/2$	0	$1/2$	$2/3$	1	2	5
ME										
Power div.	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34
p-value	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56

Severe turbulence conditions: EDR > 0.3$\mathrm{m}^{2/3}\mathrm{s}^{-1}$, 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

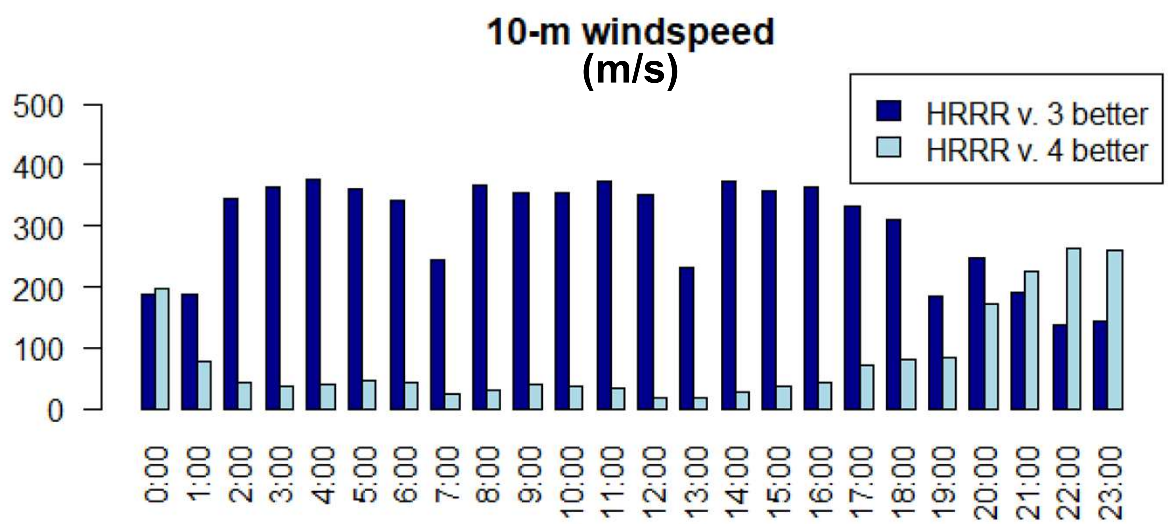
Test Cases: HRRR Temperature and Wind Speed

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.