Statistical Hypothesis Testing In Atmospheric Science Applications

Eric Gilleland
Research Applications Laboratory
National Center for Atmospheric Research

28 June 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 mainly covers the following papers:

- Gilleland, E., 2020. Bootstrap methods for statistical inference. Part I: Comparative forecast verification for continuous variables. *Journal of Atmospheric and Oceanic Technology*, **37** (11), 2117 - 2134, doi: 10.1175/JTECH-D-20-0069.1.
- Gilleland, E., 2020. Bootstrap methods for statistical inference. Part II: Extreme-value analysis. *Journal of Atmospheric and Oceanic Technology*, **37** (11), 2135 - 2144, doi: 10.1175/JTECH-D-20-0070.1.
- Gilleland, E., A. S. Hering, T. L. Fowler, and B. G. Brown, 2018. Testing the tests: What are the impacts of incorrect assumptions when applying confidence intervals or hypothesis tests to compare competing forecasts? *Mon. Wea. Rev.*, **146** (6), 1685 - 1703, doi: 10.1175/MWR-D-17-0295.1.
- Gilleland, E. D. Muñoz-Esparza, and D. Turner (2023) "Competing forecast verification: Using the power-divergence statistic for testing the frequency of "better"." Accepted to Weather and Forecasting, doi: 10.1175/WAF-D-22-0201.1.

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., I will call these ***loss functions***).
- 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.

T-test

- Two-sample test statistic
- Denote the sample means of the loss function for models A and B $\varepsilon_A(t)$ and $\varepsilon_B(t)$ by $\bar{\varepsilon}_A$ and $\bar{\varepsilon}_B$, respectively.

$$\hat{T}_n = \frac{\bar{\varepsilon}_A - \bar{\varepsilon}_B - \mu_A - \mu_B}{\widehat{\text{se}}(\bar{\varepsilon}_A - \bar{\varepsilon}_B)}$$

- Paired test statistic
- Let $d(t) = \varepsilon_A(t) - \varepsilon_B(t)$, called the loss differential series, and denote its population mean by μ_d and its sample mean by $\bar{d}$.

$$\hat{T}_n = \frac{\bar{d} - \mu_d}{\widehat{\text{se}}(\bar{d})}$$

Need to be estimated from the data and each involves division by a term involving $\frac{1}{\sqrt{n}}$, and each estimate involves an assumption of temporally independent series.

Variance Inflation Factor

- Variance Inflation Factor (VIF)
- Multiply the estimated standard error by a factor, $\mathcal{V}$, to increase its value, where

$$\mathcal{V} = 1 + 2\sum_{i=1}^{n-1}\left(1 - \frac{i}{n}\right)\ \hat{\varphi}_i,$$

where $\hat{\varphi}_i$ is the estimated correlation of the time series between all points in time separated by a lag of i.

- This approach works well if the underlying time series follows an AR(p) process, but usually the simplifying assumption that $p = 1$ is used in practice.

Wilks, D. S. (1997) doi: 10.1175/1520-0442(1997)010<0065:RHTFAF>2.0.CO;2

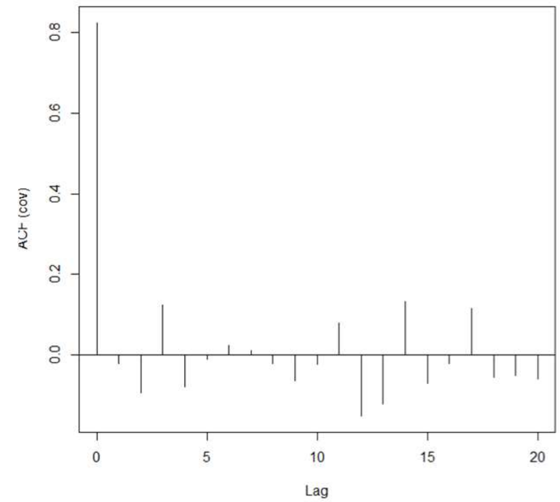
Zwiers and von Storch (1995) doi: 10.1175/1520-0442(1995)008<0336:TSCIAI>2.0.CO;2

Hering-Genton (HG) Test

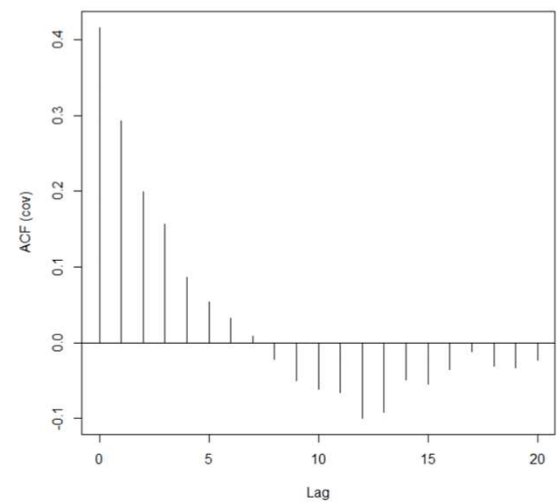
- Instead of inflating the variance, try to estimate it while accounting for the dependence directly
- Follows Diebold-Mariani approach in using a weighted average of the the auto-covariance function (ACF) over several lags, but instead fits a parametric model to the ACF.

A. S. Hering and M. G. Genton (2011) doi: 10.1198/TECH.2011.10136

ACF for an iid N(0,1) series



ACF for a dependent series



Need to have a notion of likelihood

- Interested in the ***mean*** loss differential.
- Most common estimate for the mean is the sample mean given by

$$\bar{d} = \frac{1}{n}\sum_{t=1}^{n} d_t$$

- Suppose the true mean is μ_d and the standard deviation of the sample is σ. Then…

$$E[\bar{d}] = \mu_d$$

$$\mathrm{Var}[\bar{d}] = \frac{\sigma^2}{n}$$

So, the standard error is given by

$$\mathrm{se}[\bar{d}] = \frac{\sigma}{\sqrt{n}}$$

Need to have a notion of likelihood

- Need to know the shape of $\bar{d}$'s probability distribution.
- Central limit theorem (CLT) applies to ***independent and identically distributed random variables*** (iid).
- That is, if $d_1, d_2, \ldots, d_n$ are independent with probability distribution F, then

$$\frac{d_1 - \mu_d}{\sigma/\sqrt{n}} + \frac{d_2 - \mu_d}{\sigma/\sqrt{n}} + \cdots + \frac{d_n - \mu_d}{\sigma/\sqrt{n}} = \frac{\bar{d} - \mu_d}{\sigma/\sqrt{n}} = Z \sim N(0,1)$$

Two problems:

1. We know that $d_1, d_2, \ldots, d_n$ are not independent!
2. We do not know σ, and therefore, $\mathrm{se}[\bar{d}]$, so it has to be estimated.

For 1, the CLT still applies but the estimate for $\mathrm{se}[\bar{d}]$ needs to be adjusted as the effective sample size is smaller than we think it is (that is where the HG estimate comes in, and the VIF, etc.).

For 2, the t-distribtution with $n-1$ degrees of freedom can be used instead, which is approximately standard normal for large enough n.

Bootstrapping: when you don't have a notion of likelihood

- Can be used to estimate the standard error directly, or
- To obtain a confidence interval with or without directly estimating the standard error
  - Many such methods available
  - See doi: 10.5065/D6WD3XJM, and references therein for a review.
- Most methods do not require any distributional assumption (though there are still other assumptions).

- IID bootstrap is most common
  1. Let the true but unknown value of the test statistic or parameter of interest be denoted by θ. For example, θ is μ_d in our setting.
  2. Denote $\hat{\theta}$ as the estimated parameter value for θ from the original data set (call it the bootstrap estimate of the test statistic or parameter of interest)
  3. Take a sample with replacement from the data and estimate $\hat{\theta}$. Denote this estimate by $\hat{\theta}^*$.
  4. Repeat step 3 many times, say B times, to obtain a sample $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ of $\hat{\theta}$.
  5. Estimate $\widehat{\text{se}}(\hat{\theta})$ from the sample in step 4 or estimate CI's.

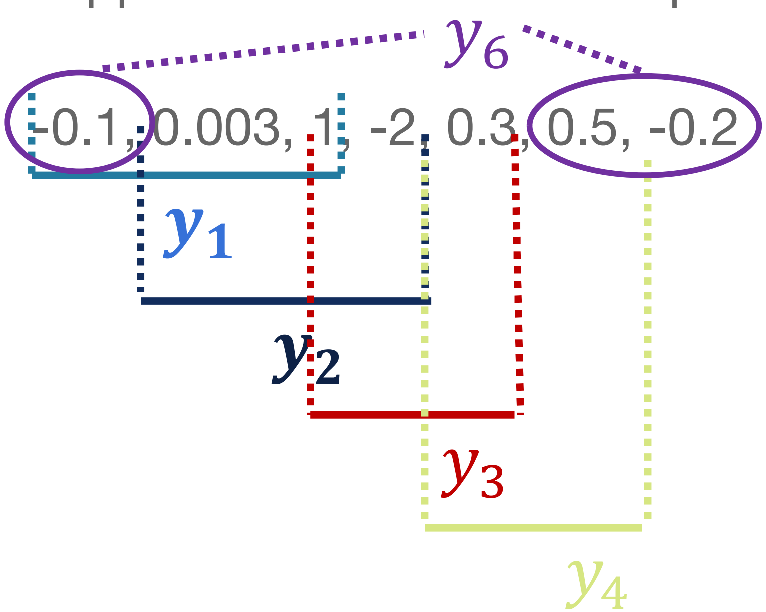
Bootstrapping

IID bootstrap is not appropriate when data are temporally (or spatially) correlated. The circular-block (CB) bootstrap can be used in its stead.

Instead of sampling with replacement from the original data, say $d_1 = \varepsilon_{A1} - \varepsilon_{B1}, \dots, d_n = \varepsilon_{An} - \varepsilon_{Bn}$, sample with replacement from $y_1, \dots, y_n$, where $y_1 = \{d_1, d_2, \dots, d_k\}, y_2 = \{d_2, d_3, \dots, d_{k+1}\}, \dots, y_\ell = \{d_\ell, \dots, d_{\ell+k-1}\}, \dots, y_n = \{d_n, d_1, \dots, d_{k-1}\}$.

Bootstrapping

For example, suppose we observe: -0.1, 0.003, 1, -2, 0.3, 0.5, -0.2, and suppose we choose to sample blocks of length $k = 3$, then we would sample:



$$y_1 = -0.1, 0.003, 1$$
$$y_2 = 0.003, 1, -2$$
$$y_3 = 1, -2, 0.3$$
$$y_4 = -2, 0.3, 0.5$$
$$y_5 = 0.3, 0.5, -0.2$$
$$y_6 = 0.5, -0.2, -0.1$$
$$y_7 = -0.2, -0.1, 0.003$$

Bootstrapping

If we are interested in a statistic, say T_n (e.g., $T_n = \bar{d}_n$), then we start with the paradigm that T_n is a random variable that follows some distribution function, say F.

Note that T_n is based on data. For example, $\bar{d}_n = \frac{1}{n}\sum_{t=1}^{n} d_t$ is based on $d_1, d_2, \ldots, d_n$.

When we resample from the data (or however we sample) the resulting statistic, T_m^*, is based on the resampled data, which in our example would be $d_1^*, d_2^*, \ldots, d_m^*$. Note that m may or may not be the same as n.

T_m^* is a random variable that follows a distribution, say F_n.

Bootstrapping

Bootstrapping works when…

If the law of T_n tends *weakly* to a limit as $n \to \infty$, and the law of T_n^* tends weakly to the same limit law with probability one as $m, n \to \infty$ (Bickel and Freedman 1981).

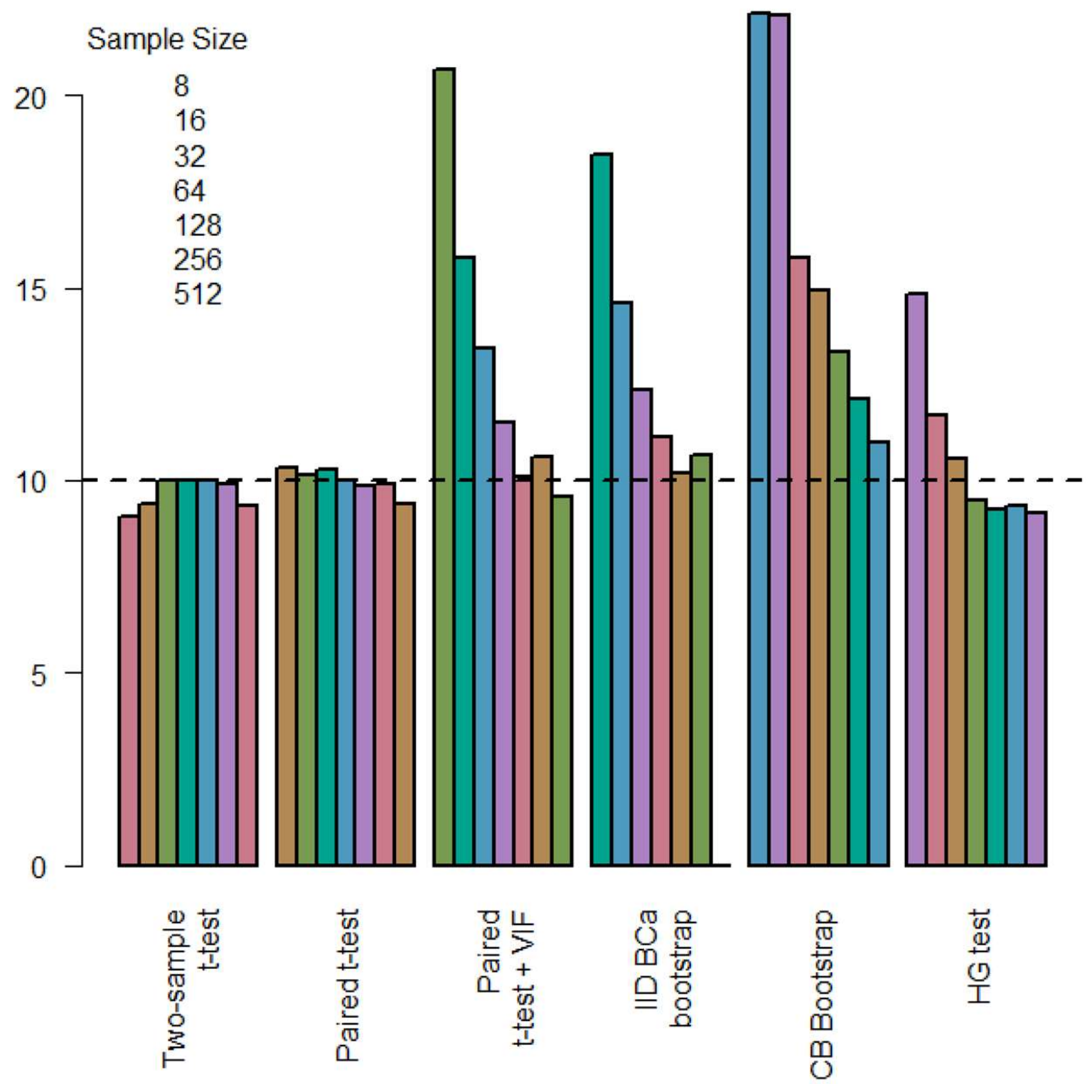
Simulation Experiment to test different hypothesis tests

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 different test procedures to test $\mathcal{H}_0\colon \mu_A = \mu_B$ against $\mathcal{H}_1\colon \mu_A \neq \mu_B$ for various loss functions, such as AE or SE.
  - Note that although the raw error series are simulated to have mean zero, when testing for AE or SE loss, $|\varepsilon(t)| > 0, \varepsilon^2(t) > 0$ for all t so that the MAE or RMSE will be positive valued.
  - 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.

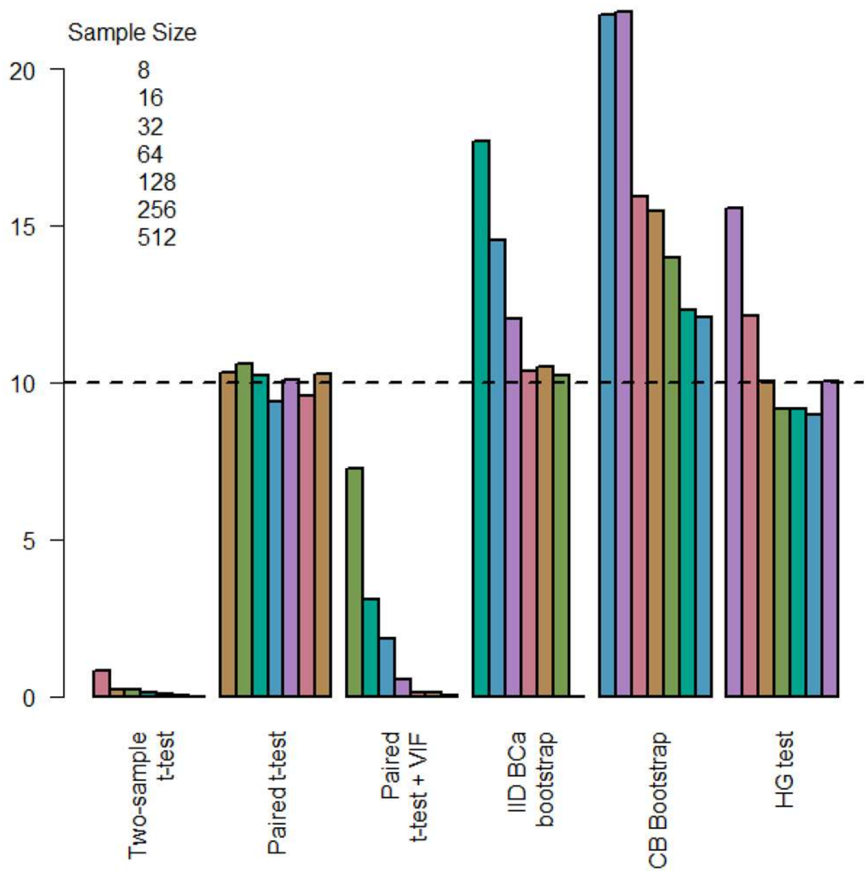
Testing the tests

Independence Case
$(\rho = 0, \theta = 0)$

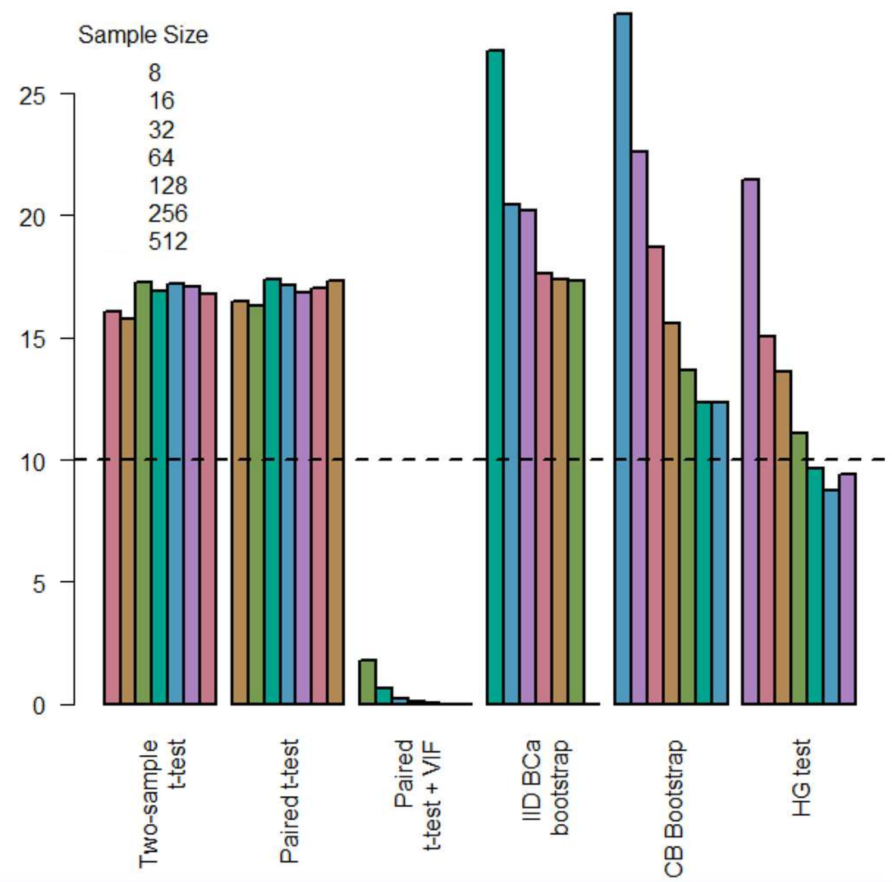


Testing the tests

Strong contemporaneous correlation, temporal independence
$(\rho = 0.9, \theta = 0)$



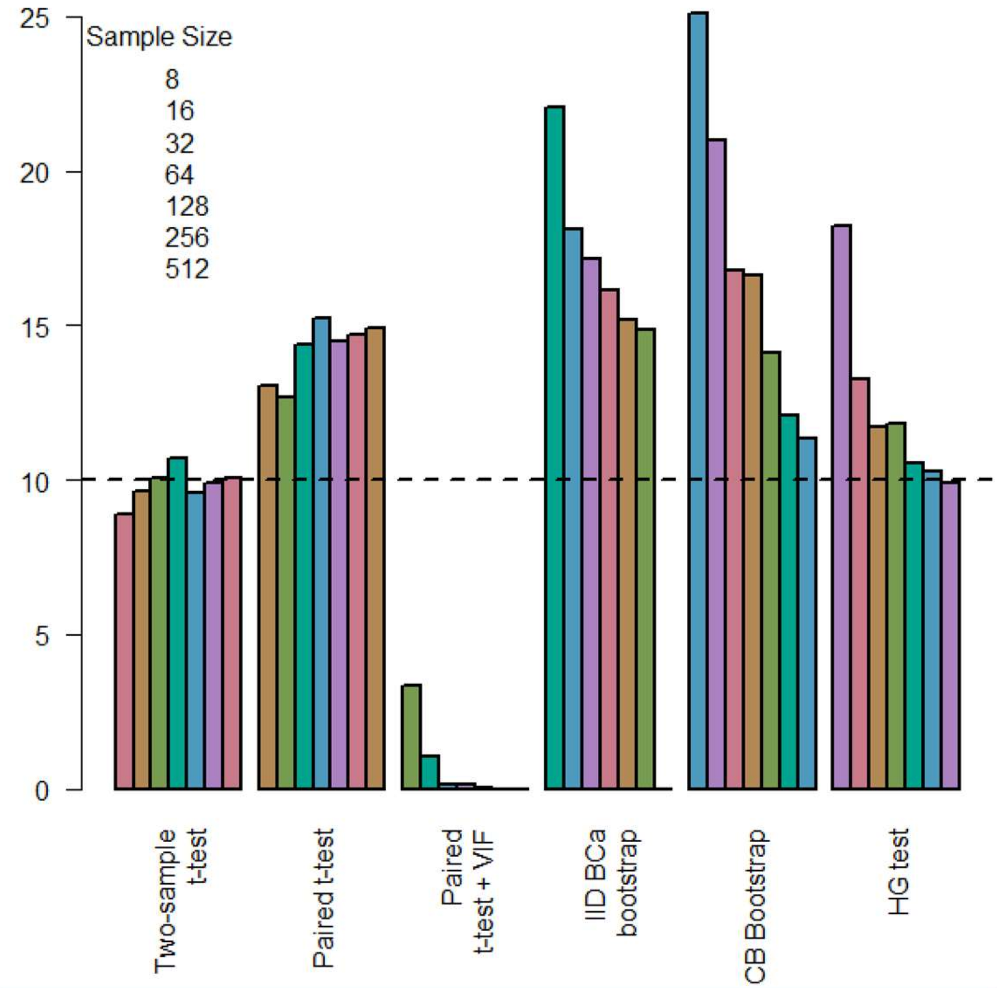
No contemporaneous correlation, temporal dependence
$(\rho = 0, \theta = 0.9)$



Testing the tests

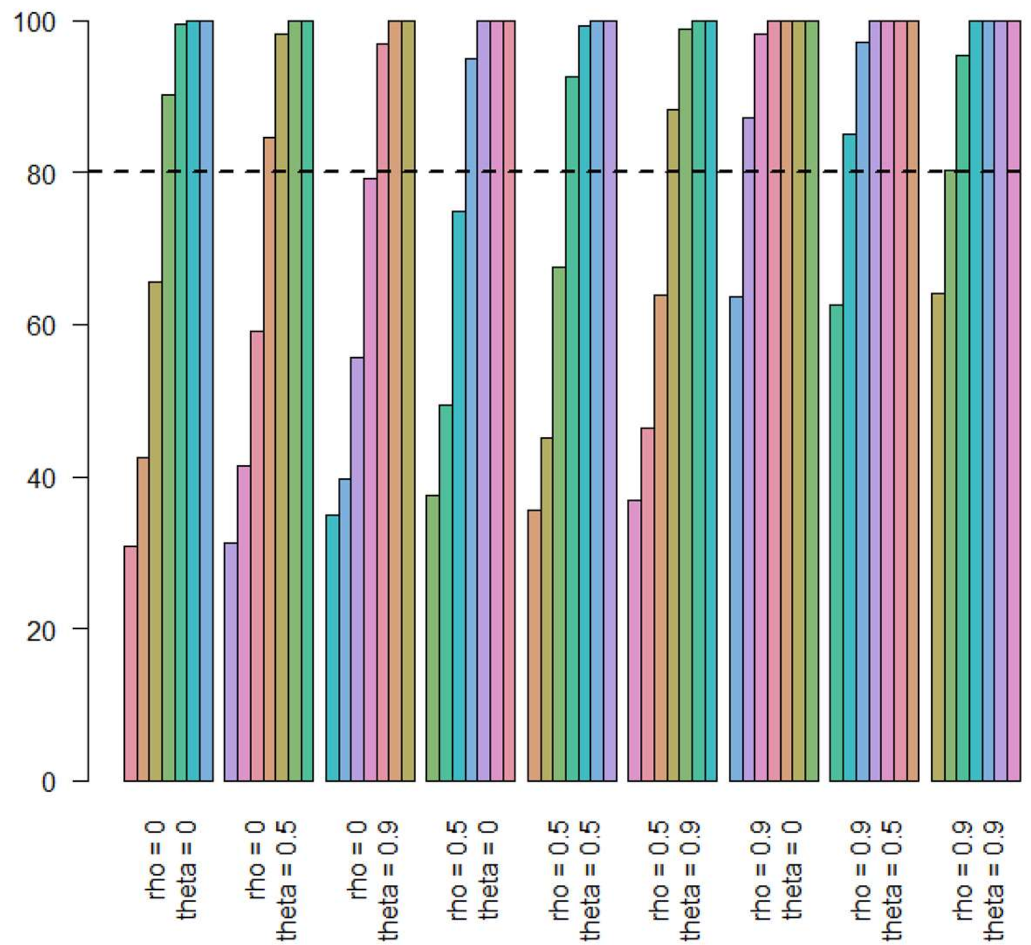
Moderate contemporaneous correlation and temporal dependence case

$$\left(\rho = \frac{1}{2}, \theta = \frac{1}{2}\right)$$

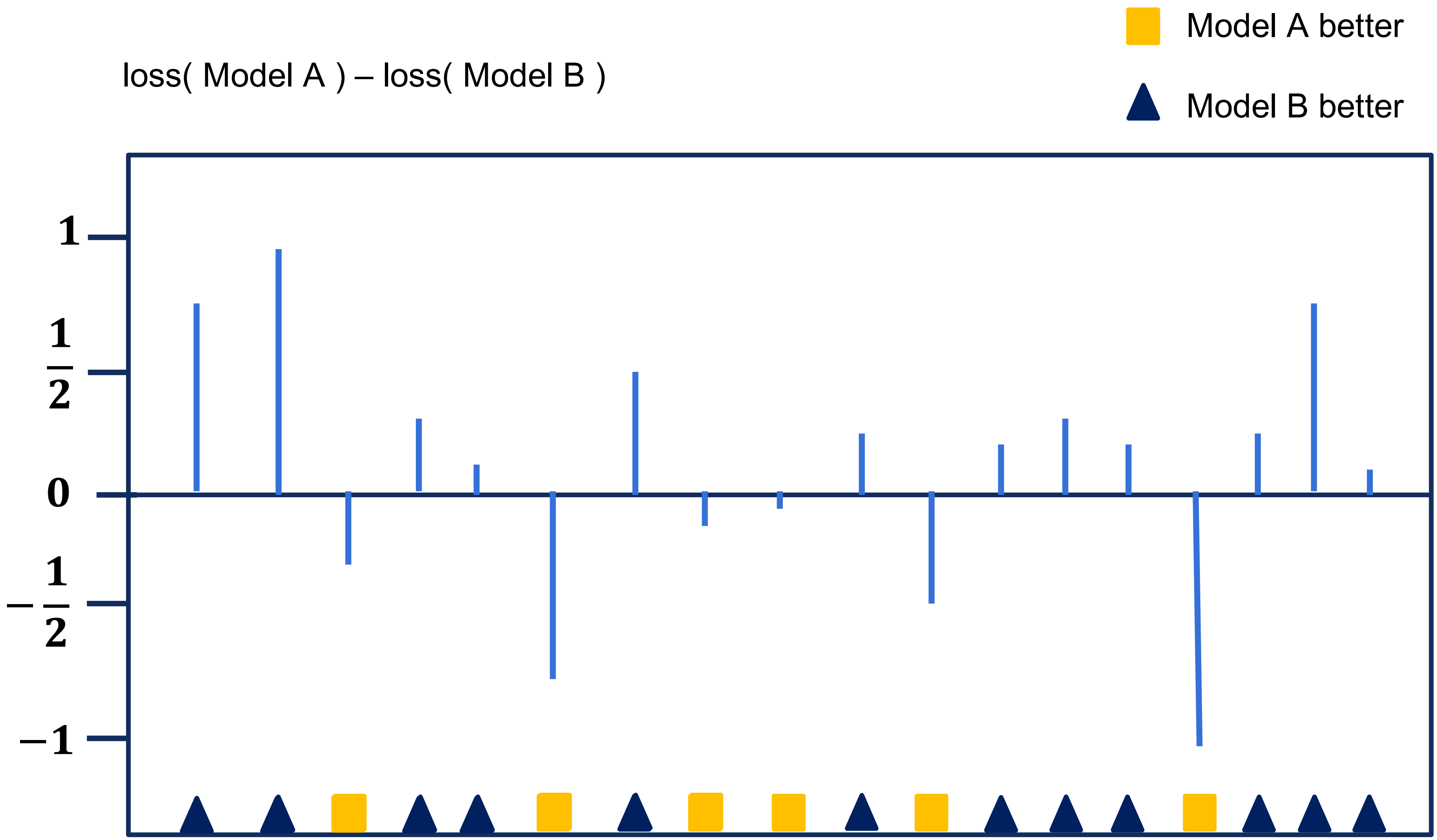


Power for HG test

Sample Sizes
8 16 32 64 128 256 512

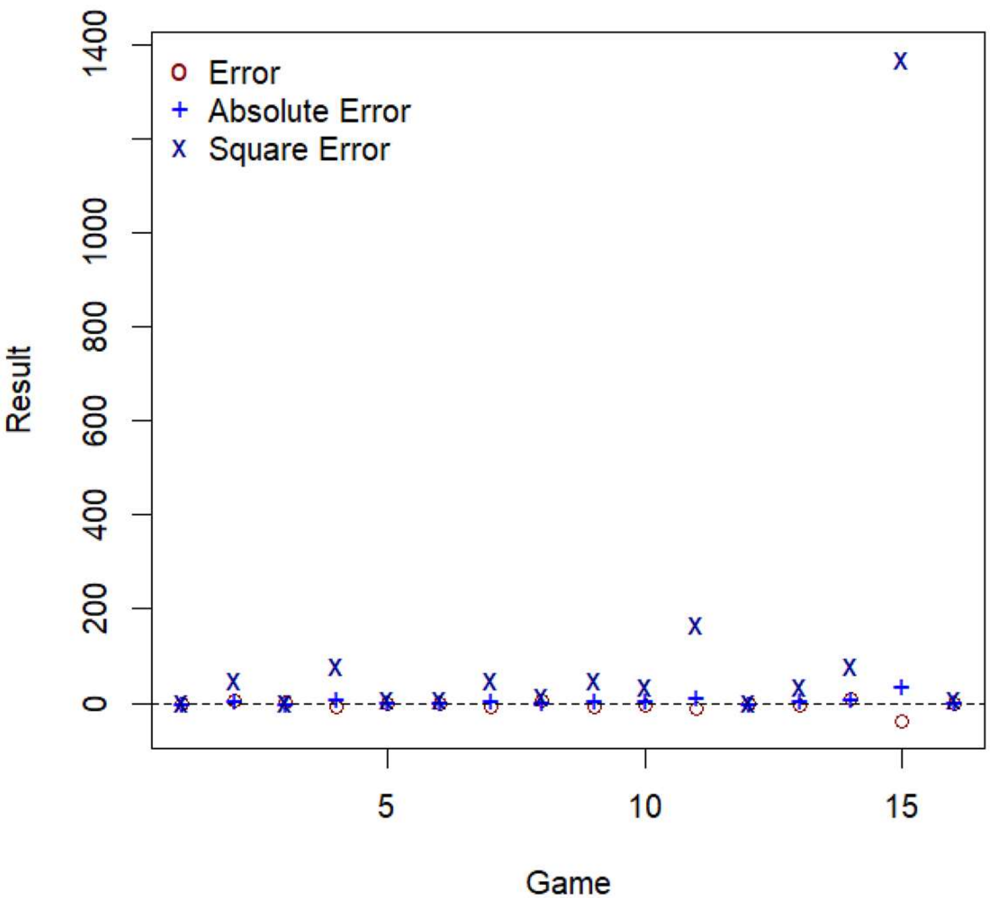


Testing the Frequency of “Better”



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 $q = \frac{1}{2}$ for our setting.

Power-divergence Statistic

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

where for our setting:

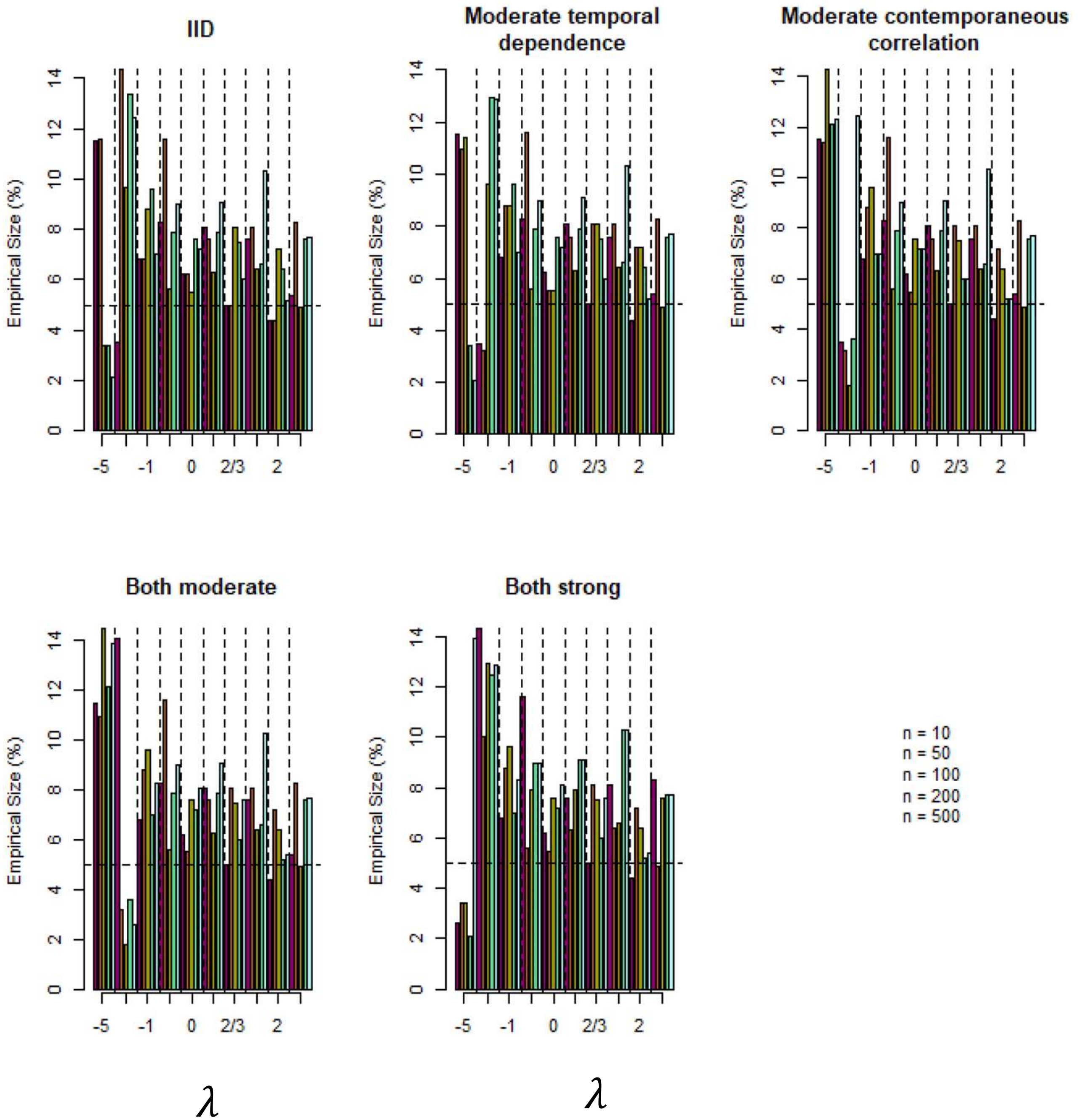
- $k = 2$
- $\hat{\boldsymbol{p}} = (\hat{p}_1, \hat{p}_2) = (\hat{p}, 1 - \hat{p})$ is the estimate of p from the data
- $\boldsymbol{q} = (q_1, q_2) = (q, 1 - q\ ) = \left(\frac{1}{2}, \frac{1}{2}\right)$ is the vector of test parameters
- $\boldsymbol{\lambda}$ is a user-chosen value that yields different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

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)

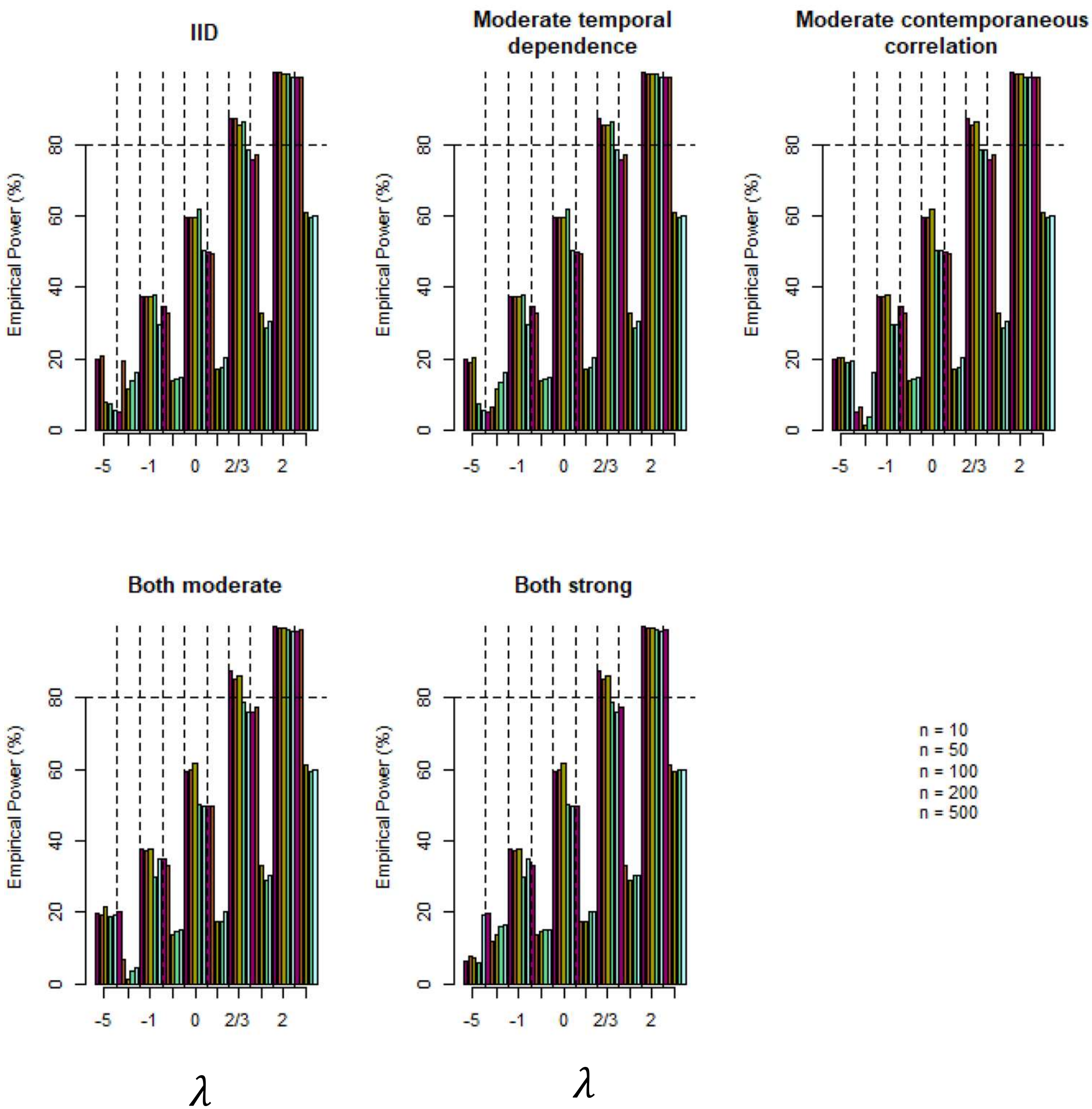
Above table is taken from Table 1 in Gilleland et al., (accepted to WAF). *And is a summary of some information taken from: Read and Cressie (1988).*

Power-divergence Statistic



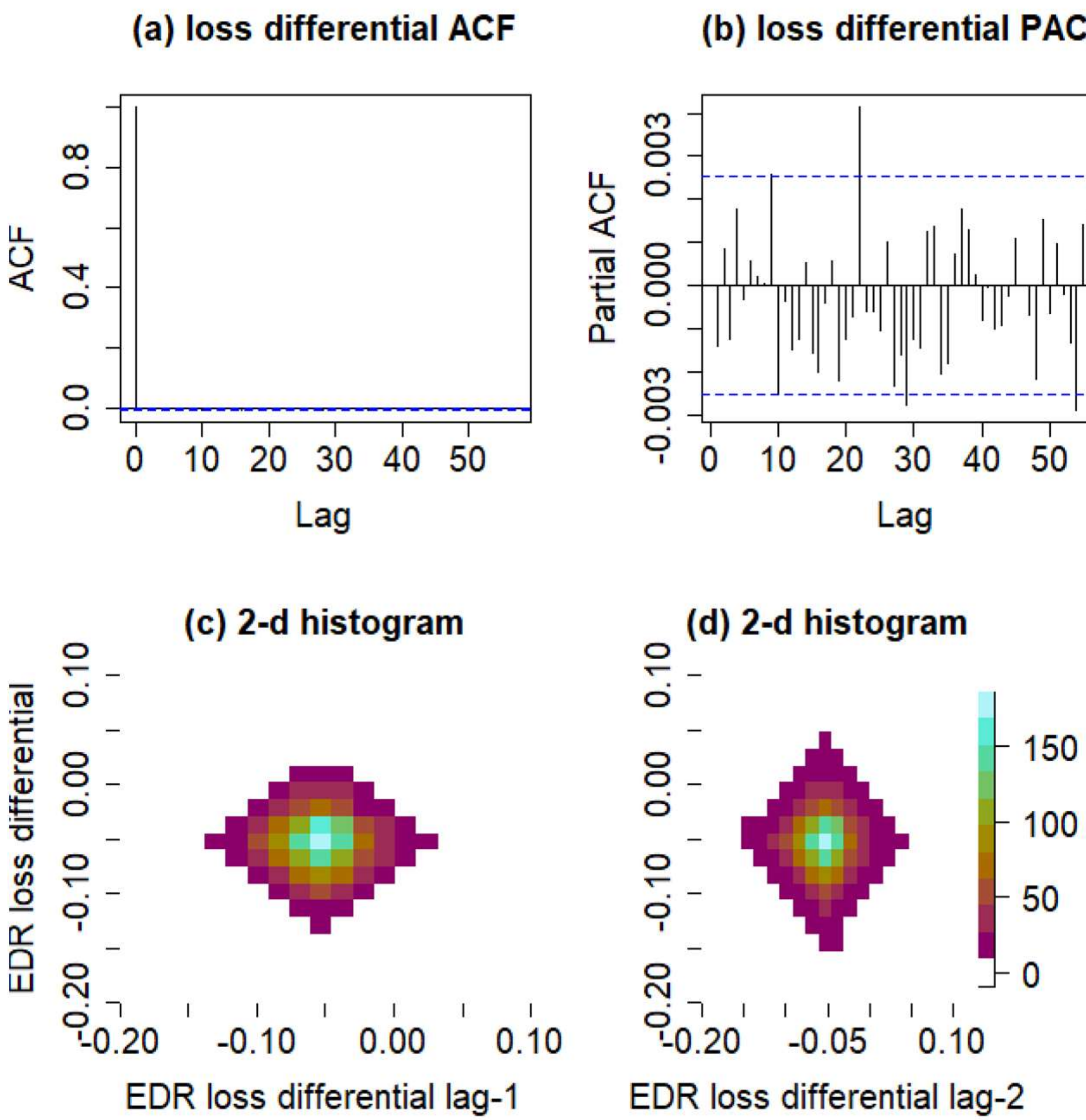
Empirical Size testing (using 5%) with simulations as in Hering and Genton (2011)

Power-divergence Statistic



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

Test Cases: Turbulence

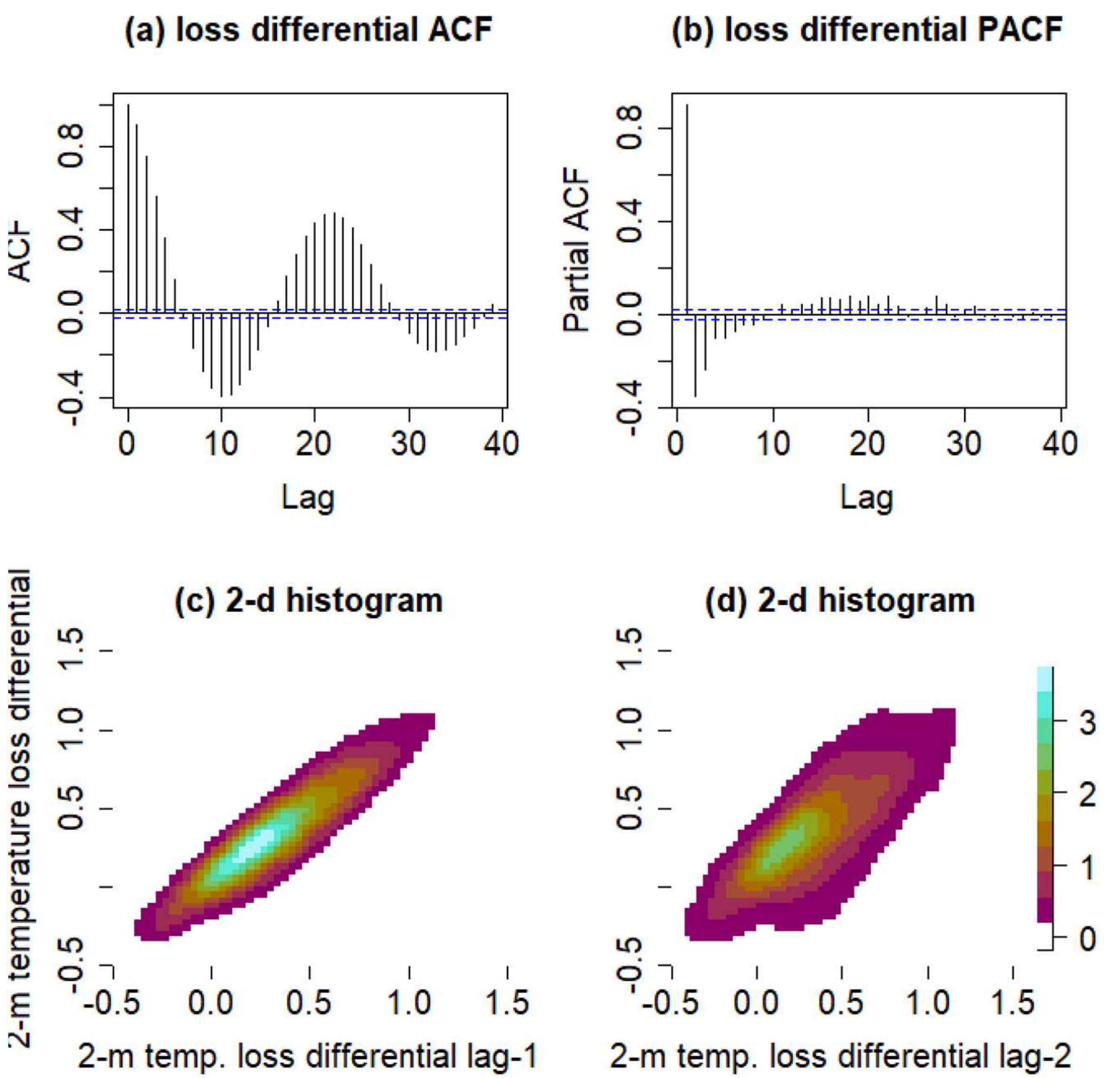


Two versions of 6-h turbulence forecasts called the Graphical Turbulence Guidance (GTG) algorithm for eddy dissipation rate (EDR, $\mathrm{m}^{2/3}\mathrm{s}^{-1}$, Sharman and Pearson 2017; Muñoz-Esparza and Sharman 2018; Muñoz-Esparza et al. 2020).

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$m^{2/3}s^{-1}$

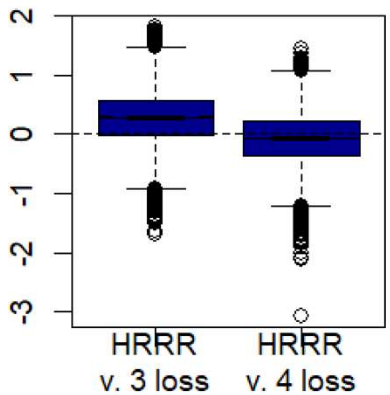
λ	-5	-2	-1	$-1/2$	0	$1/2$	$2/3$	1	2	5
ME										
Power div.	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34
p-value	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56

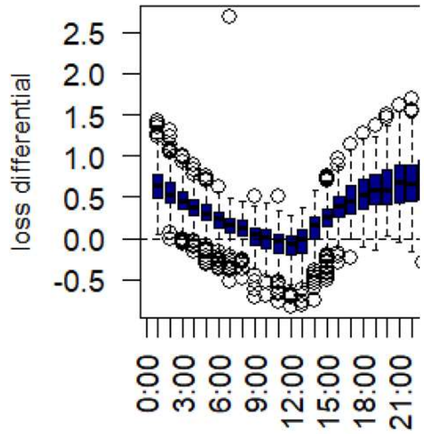
Severe turbulence conditions: EDR > 0.3$m^{2/3}s^{-1}$, which is about 0.1% of the total sample.

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

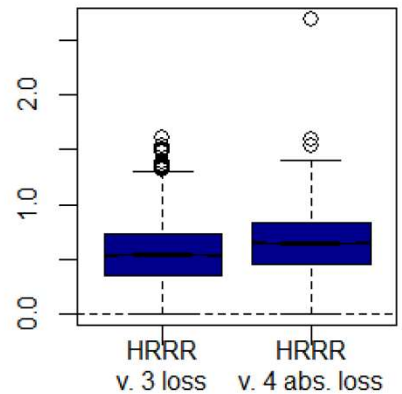
Test Cases: HRRR Temperature and Wind Speed

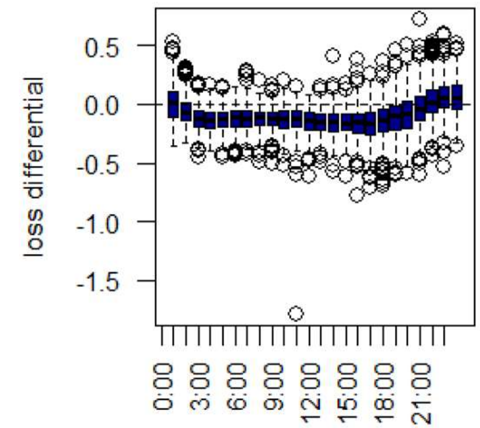
12-h forecasts of 2-m temperature (deg. C)

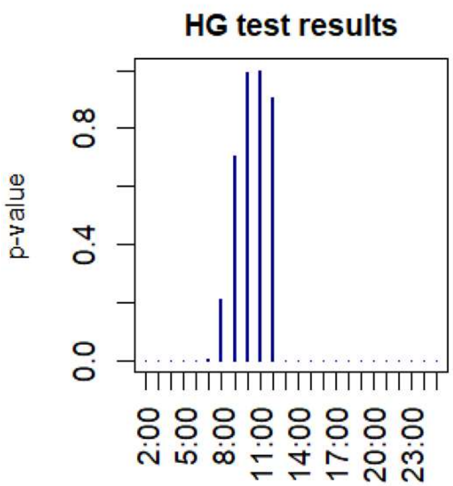


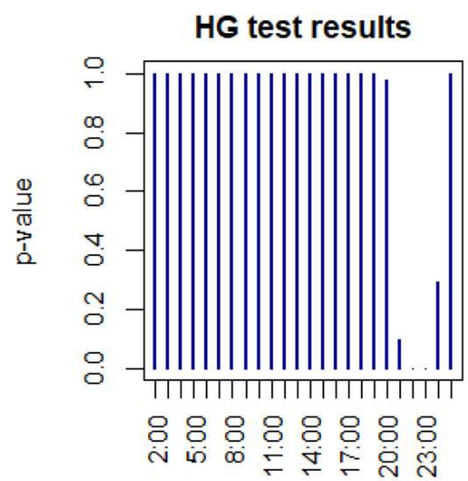


12-h forecasts of 10-m wind speed (m/s)





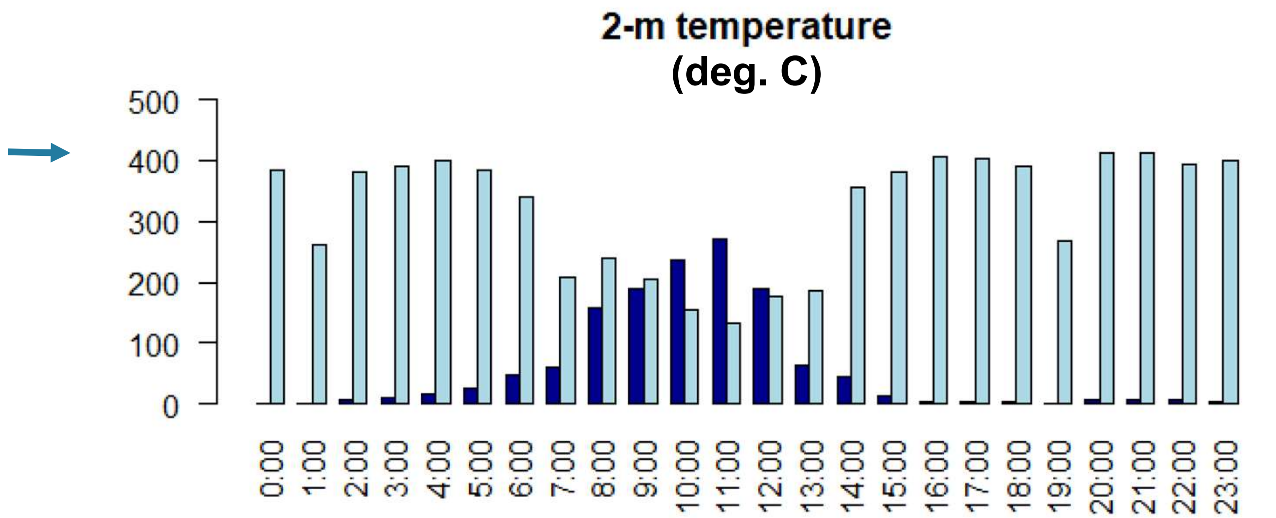




The Hering-Genton test (Hering and Genton 2011) is a t-test on the mean loss differential where the standard error is estimated in a way that accounts for temporal dependence, and the test is robust to contemporaneous correlation. It is a test on the intensity difference in error rather than the frequency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

Test Cases: HRRR Temperature and Wind Speed

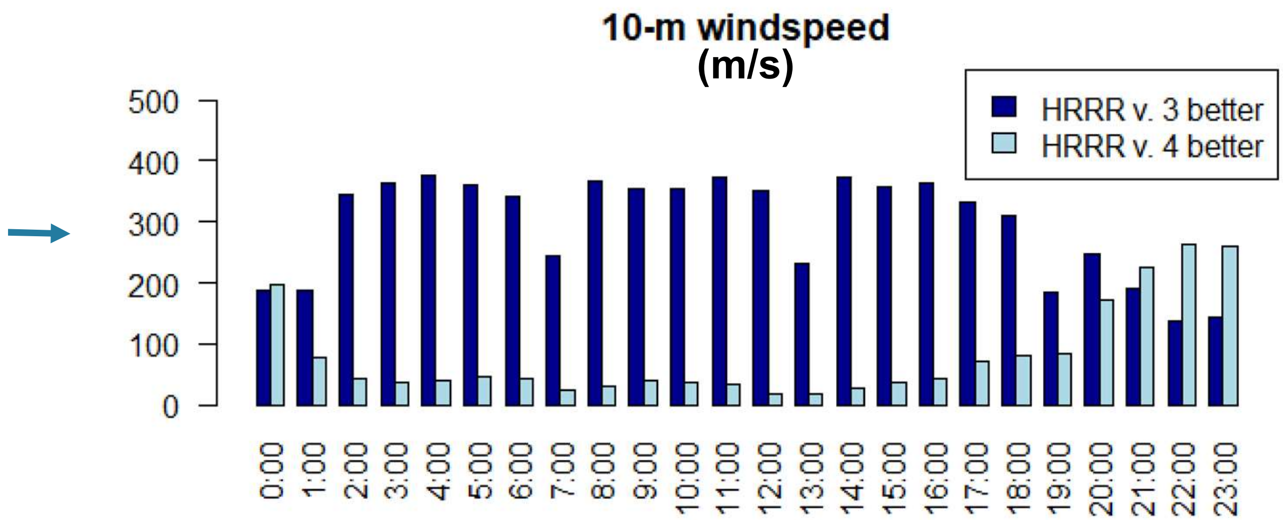
For all choices of λ applied previously, the power-divergence rejects $\mathcal{H}_0$ at all times except at 9 and 12 UTC



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 $\mathcal{H}_0$.

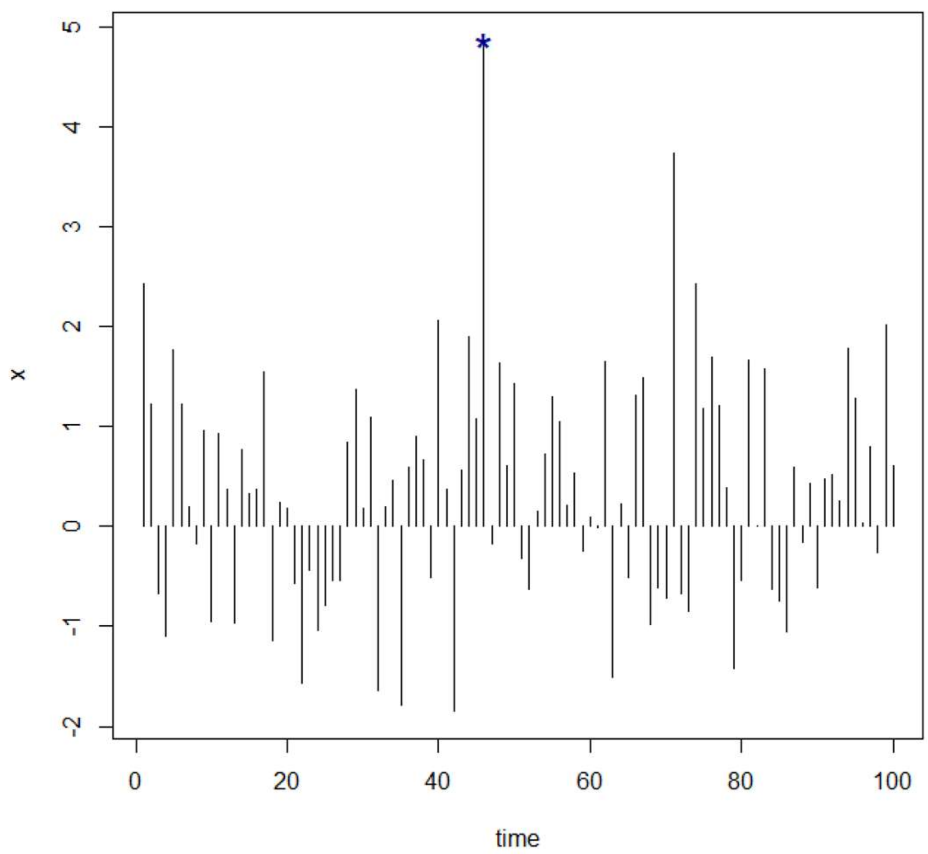
Results based on a 5%-level test, but p-values estimated to be zero.



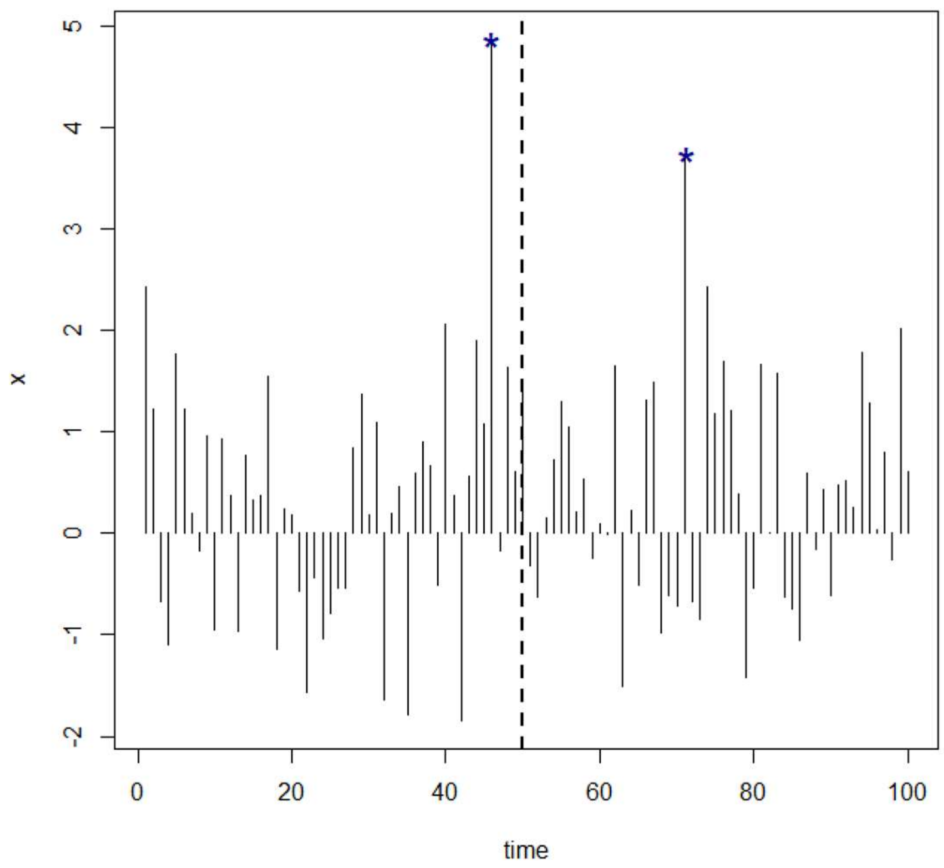
Extreme Values

Image citation: http://n2t.net/ark:/85065/d72v2d5b

Maximum



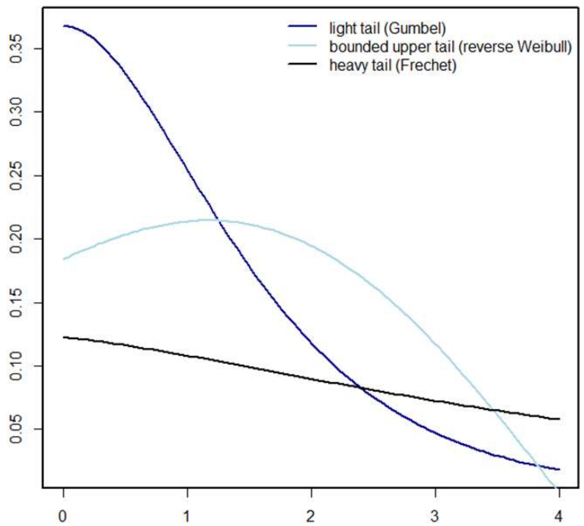
x
time
-2
-1
0
1
2
3
4
5
0
20
40
60
80
100



x
time
-2
-1
0
1
2
3
4
5
0
20
40
60
80
100

Maximum

Theoretical justification for the GEV(μ, σ, ξ) as the limiting distribution for maxima over long blocks of time (think annual). Analogous results for excesses over a high threshold. Combine them all



$$G(z) = \exp\left\{-\left[1 + \frac{\xi}{\sigma}(z - \mu)\right]_+^{-\frac{1}{\xi}}\right\}$$

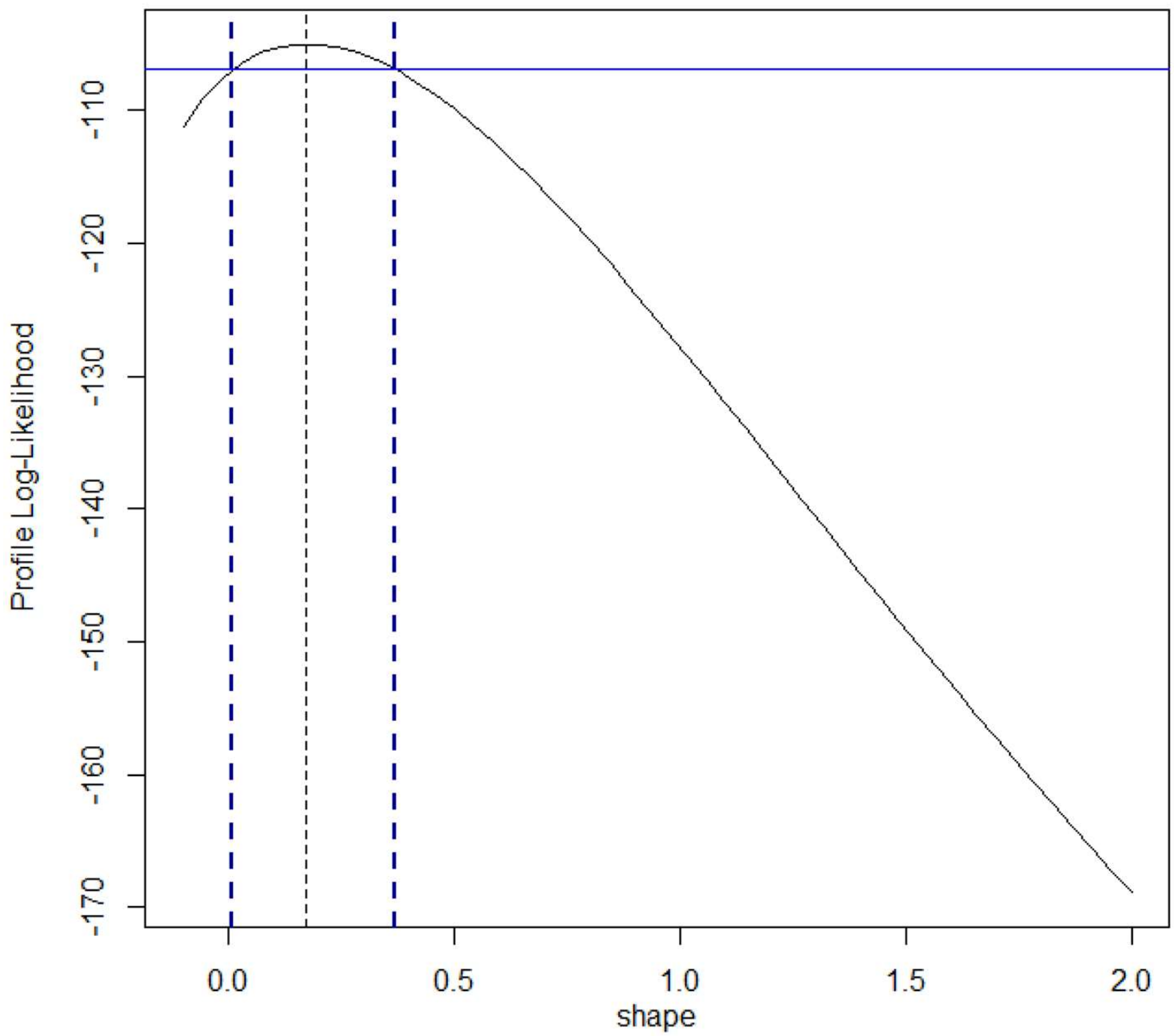
Estimation

Can use maximum-likelihood (ML), L-moments (and other moment-type methods), Bayesian and various non-parametric methods. MLE is perhaps the most commonly used.

MLE issues with EVD's:

- Regularity assumptions for the MLE to follow a normal distribution are not always met so that the assumptions for using parametric CI's for parameters and/or return levels may not be valid.
- Bootstrapping is more complicated because of the slow convergence to the error distribution ($m < n$ bootstrap is appropriate for heavy-tail case).
- Profile-likelihood and test-inversion bootstrap are best choices, but both can be very difficult to implement and automate.

Estimation



Profile Log-Likelihood
-110
-120
-130
-140
-150
-160
-170
0.0
0.5
1.0
1.5
2.0
shape

References

Bickel, P. J. and D. A. Freedman (1981) Some asymptotic theory for the bootstrap. *The Annals of Statistics*, **9** (6), 1196 – 1217, doi: 10.1214/aos/1176345637.

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.

References

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.