HEFS workshop, 03/12/2015

Seminar C: ensemble verification concepts and requirements

James Brown

james.brown@hydrosolved.com

Contents

1. Why conduct verification?
2. What are the data requirements?
3. Attributes of forecast quality
4. Measures of forecast quality
5. Final thoughts and suggestions

1. Why conduct verification?

Why verify?

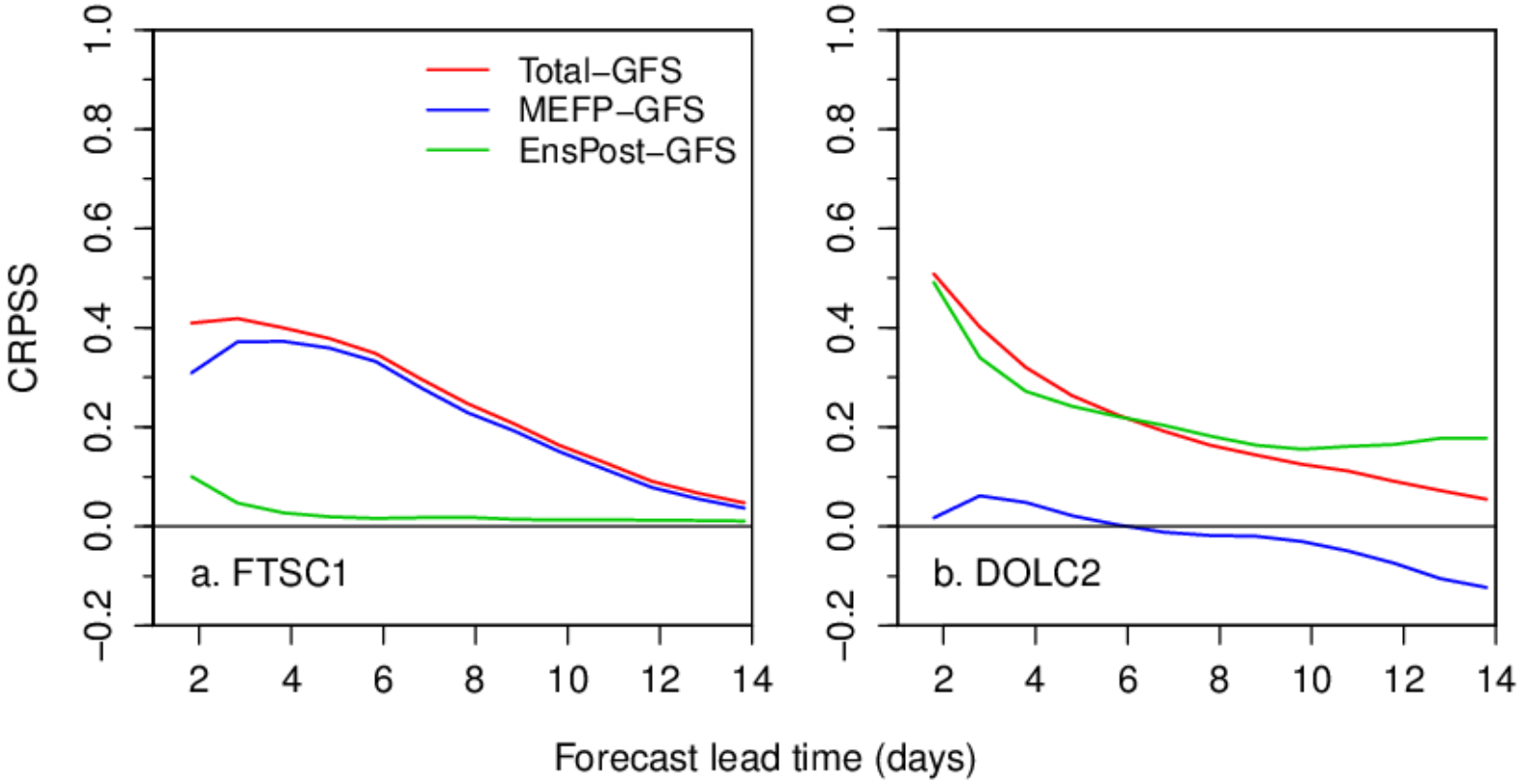
Forecasts incomplete if quality unknown

- Ensemble forecasts can be poor quality
- How much confidence to place in them?
- Are they unbiased and skillful? When/where/how?
- Where to focus improvements? Are they worth it?

An example: component error analysis

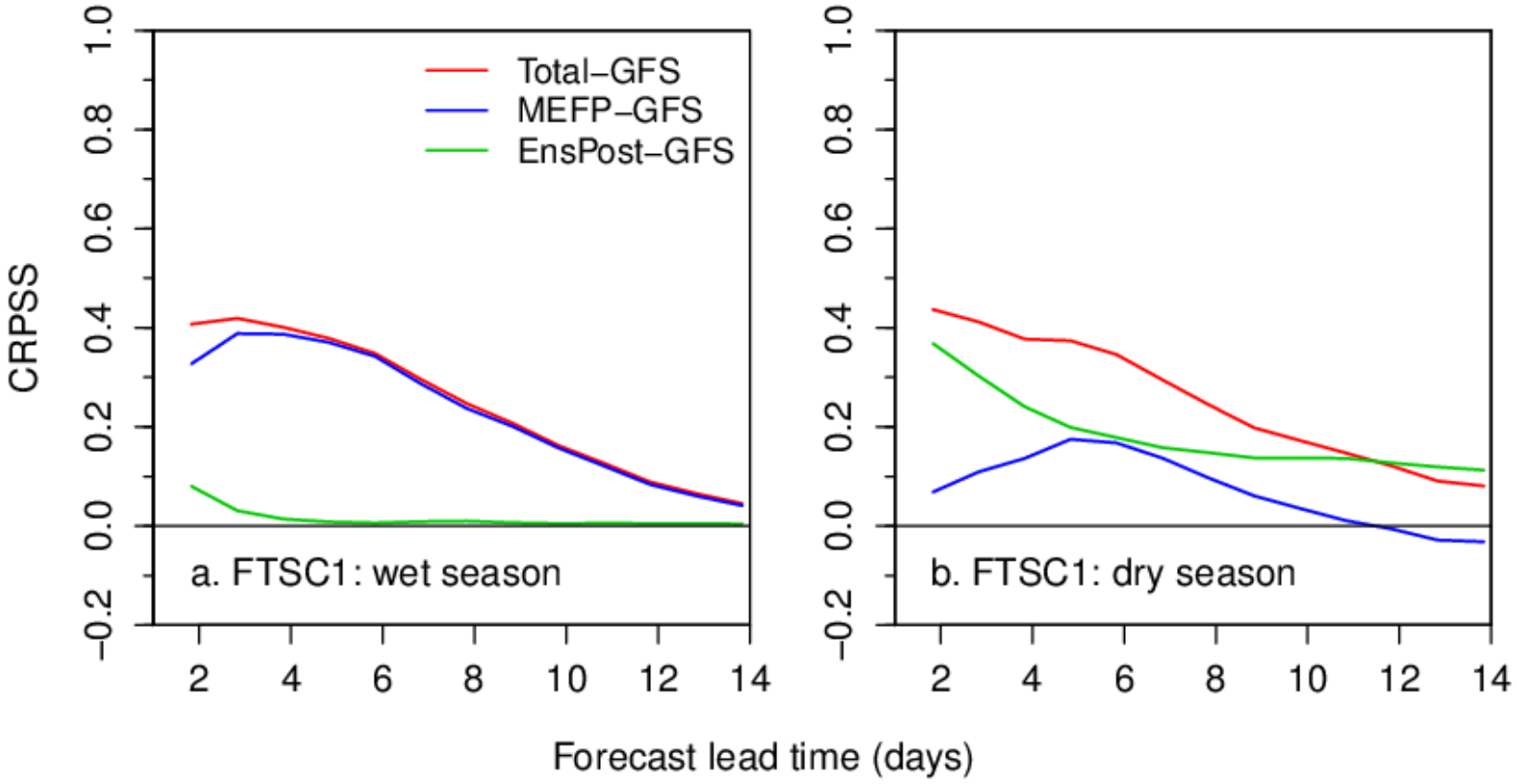
- Total uncertainty = meteorological + hydrologic
- In other words: HEFS = MEFP + EnsPost
- Component error analysis can separate the two

Example: two very different basins



- Fort Seward, CA (FTSC1) and Dolores, CO (DOLC2)
- Total skill in EnsPost-adjusted GFS streamflow forecasts is similar
- Origins are completely different (and understandable)

Example: two very different seasons



- However, in FTSC1, completely different picture in wet vs. dry season
- In wet season (which dominates overall results), mainly MEFP skill
- In dry season, skill mainly originates from EnsPost (persistence)

Different motivations/applications

Motivations and applications vary

1. National/routine verification (monitoring and reporting)
2. Forensic/diagnostic verification (to enhance/fix HEFS)
3. Screening HEFS before “go live” (selected locations)
4. Verification to support HEFS optimization locally
5. Verification to support local users (e.g. optimize DSS)

Centralized versus RFC efforts

- Details TBD, but (1)/(2) need a centralized/NWC effort
- RFCs will start with (3). Later on, (4) and (5)

2. What are the data requirements?

What data are required?

Datasets

1. Hindcasts or archived forecasts (forcing and flow)
2. Trustworthy observations (no major biases, gaps etc.)
3. Historical simulations for component error analysis

- Large sample and consistent record for (1)-(3)

Sampling uncertainty depends on

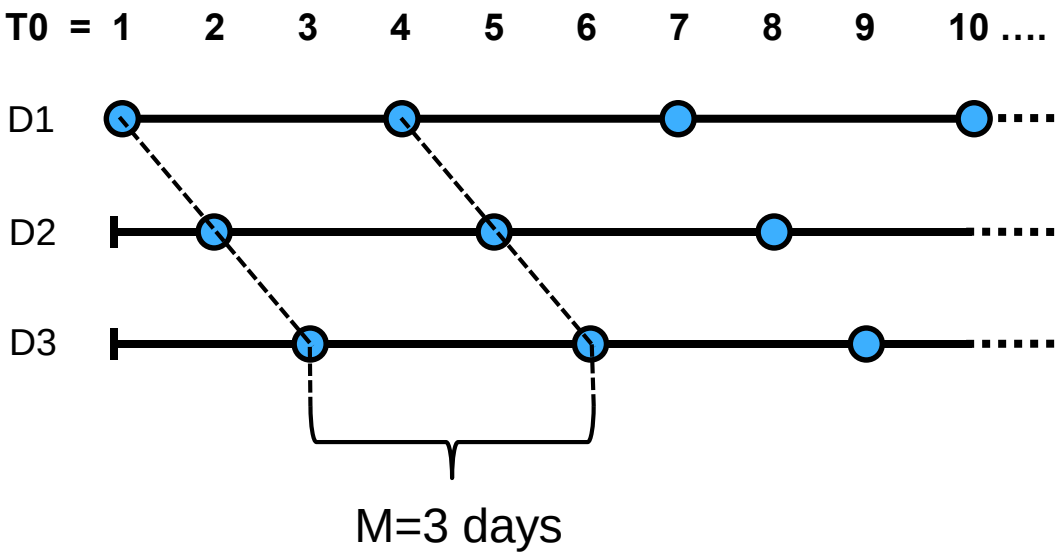
- Hindcasts: length, frequency, aggregation period
- Verification: sub-sampling or “conditional verification”
- Verification: choice of metric

Example: impacts of sample size

MEFP sensitivity study

- Explored sensitivity to both number of years (N) and interval between T0s (M)
- This diagram illustrates the approach for M where N is fixed (N=24 years)
- For M=3, there are three separate hindcast datasets {D1,D2,D3}, each separated by 1 day
- For M=3, compute verification for each D and plot the range of results
- Repeat for other values of M (next slide)

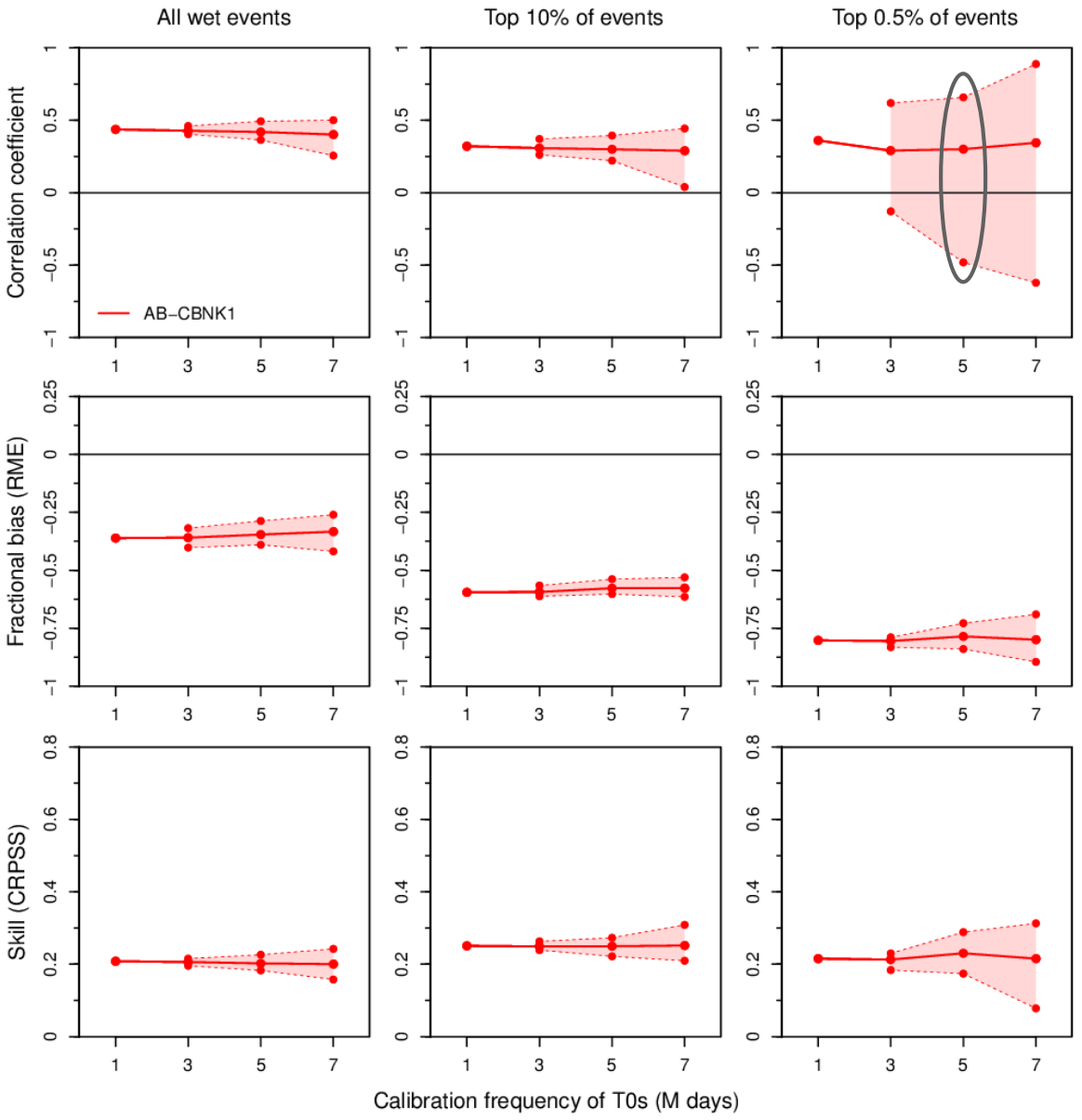
M=3 (3 days between T0s)



Example: impacts of sample size

MEFP precip. (1-3 days)

- Thinning by M is extremely aggressive, but varies with measure
- For example, at M=5, correlation for top 0.5% at AB-CBNK1 varies from -0.5 to +0.6 (circled)!
- Thus, need daily reforecasts to properly capture the most extreme precipitation
- Similar results at other locations and for N.
- Ideally need at least N=25 years of daily reforecasts (M=1) for extreme events



How to mitigate small sample?

Steps to reduce impacts

- Large and consistent (re)forecast sample (see earlier)
- Be careful with conditioning (i.e. avoid small subsets)
- Be mindful of aggregation impacts (e.g. A-J volumes)
- Take care with metric selection for small sample sizes
- Can set minimum sample size for EVS outputs

Steps to evaluate impacts

- Qualitative: check sample size plots in EVS
- Quantitative: compute confidence intervals in EVS

Data quality control (QC)

Before hindcasting: QC input data

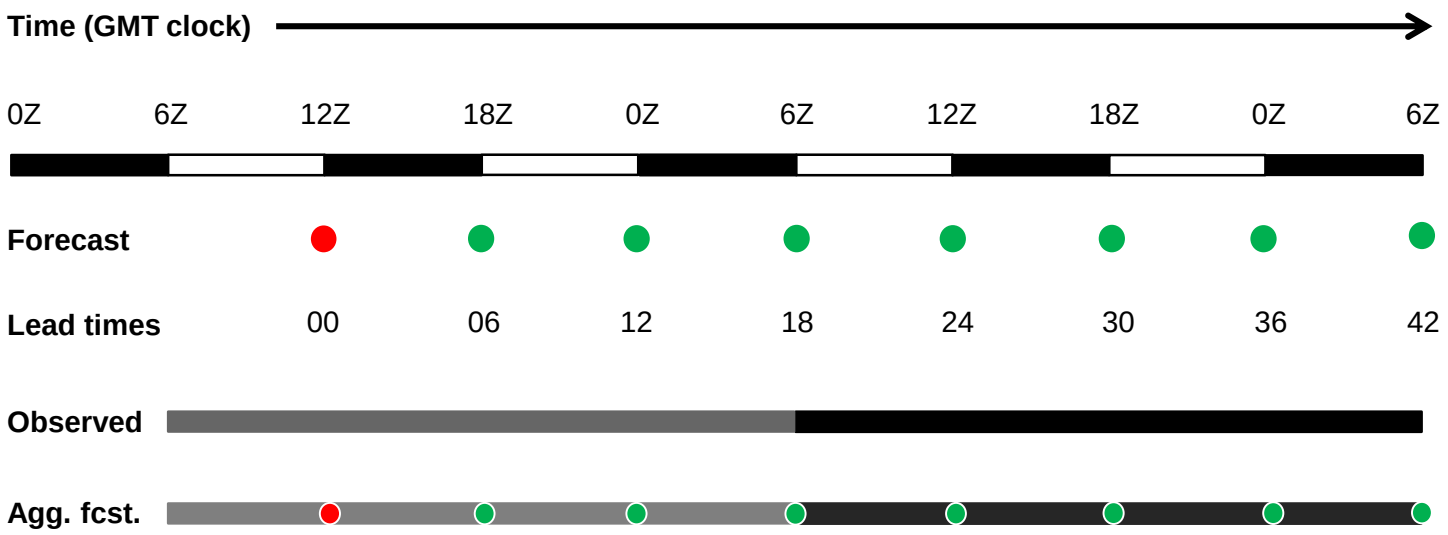
- Non-physical data and outliers (data diagnostics)
- Unrealistic parameter values (parameter diagnostics)

After hindcasting: QC output data

- Make test runs and visualize results for gross errors
- Check all expected forecasts/members present
- Check for non-physical values and outliers
- Outliers can have a large (obscured) impact on stats
- Ensure forecasts/observations are paired correctly…

Pairing mechanics and QC

- Pairing often requires assumptions/data manipulation
- For example, aggregation or re-timing of data
- Always QC the pairs (for selected locations)!
- Example: Forecast (6hr) vs. QME in ABRFC (GMT-6)



3. Attributes of forecast quality

First, the big picture

Three separate, but related, concepts

- **Quality:** concerned with forecast errors (verification)
- **Utility:** ability to serve a purpose (even with errors)
- **Consistency:** honest forecasts (no "gaming" quality)

Examples of quality vs. utility

- A flood forecasting system may be reliable (quality)…
- …but forecasts may not be timely (utility)
- Climatological ensembles are unskillfull (quality)…
- …but are useful for water resources planning (utility)

Focusing on quality (verification)

Decades of publications on quality!

- John Park Finley (1884): tornado verification
- Seminal paper: Murphy and Winkler (1987)
- Books: Jolliffe and Stephenson (2011), Wilks (2006)
- The Hydrologic Ensemble Prediction Experiment (HEPEX) is a great resource and community
  - www.hepex.org
  - http://hepex.irstea.fr/what-is-a-good-forecast/
- See resources and references slide

Two types of quality

Absolute quality vs. relative quality

- Absolute: properties of one system (vs. observed)
- Relative: comparison of two systems (vs. observed)
- Relative quality is also known as skill
- Skill is valuable, but choice of baseline needs care
  - Skill (% gain) is easy to communicate, but not always to interpret
  - Think about what you want the system to improve on (e.g. EnsPost should improve on raw streamflow forecasts)
  - Some baselines will show “naïve” skill

Attributes of quality

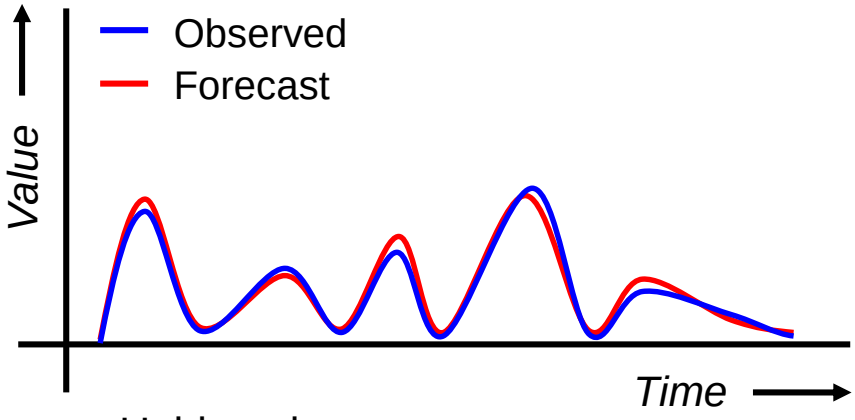
What is meant by attribute here?

- Single aspect or dimension of forecast quality
- A forecasting system has multiple quality attributes
- One attribute can have several statistical measures
- Familiar attributes from single-valued forecasting…

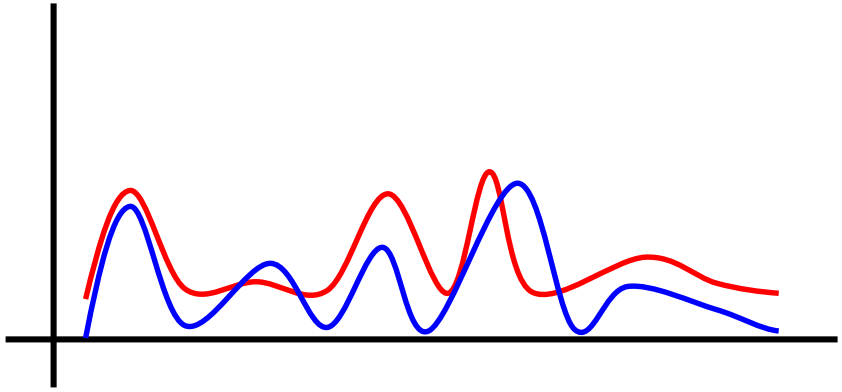
Accuracy, bias, and association

- Accuracy: concerned with total error (e.g. MSE)
  - Bias: concerned with directional error (e.g. ME)
  - Association: concerned with similarity (e.g. CORR)

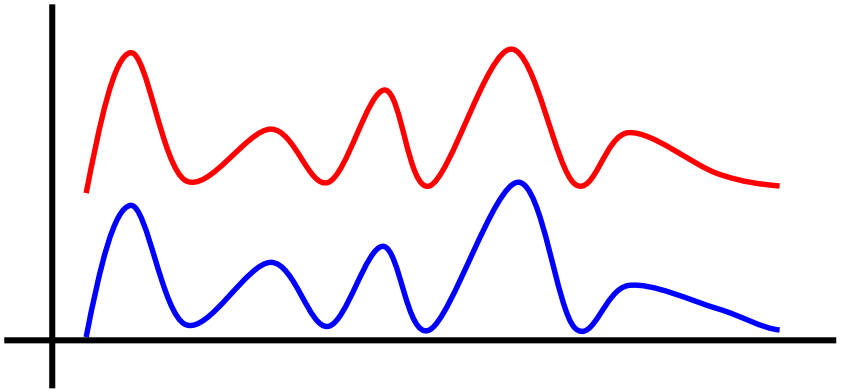
Attributes of quality: examples



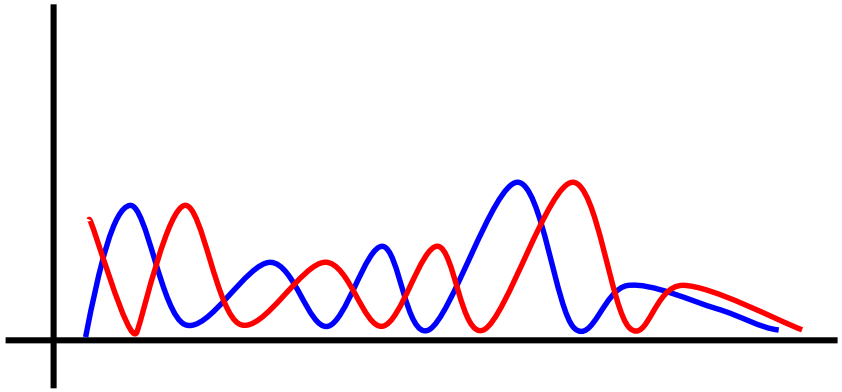
- Unbiased
- Strong association
- High accuracy (small total error)



- Some bias
- Moderate association
- Moderate accuracy (moderate total error)



- Large bias
- Strong association
- Low accuracy (high total error)



- Unbiased (but conditionally biased)
- Negative association
- Low accuracy (high total error)

Conditional attributes

Unconditional vs. conditional quality

- **Unconditional**
  - All data, no subsets (e.g. by season or amount)
  - Example: “ensemble mean has a consistent low bias”
- **Conditional**
  - Many possible conditions (season, amount etc.)
  - Example: “larger bias in ensemble mean for high flow”

Let’s move on to ensemble forecasts…

Ensemble forecasts: paired data

(X,Y)		(f(5.3),o(5.3))
({1.1,…,3.3}, 3.2)		(0.0, 0.0)
({2.6,…,21.5}, 20.2)		(0.9, 1.0)
({3.2,…,19.8}, 18.2)		(0.8, 1.0)
({4.5,…,12.5}, 13.4)		(0.7, 1.0)
({13.5,…,28.3}, 24.1)		(1.0, 1.0)
({0.2,…,7.8}, 2.1)		(0.3, 0.0)
({0.1,…,5.4}, 5.3)		(0.1, 0.0)
({7.3,…,16.5}, 12.4)		(1.0, 1.0)
({2.5,…,40.1}, 30.5)		(0.9, 1.0)
({4.9,…,57.3}, 47.2)		(0.9, 1.0)
…		…

Streamflow (Q) is both observed (Y) and forecast (X). Consider <u>one</u> discrete event: exceeding a flow threshold, <u>q=5.3 CFS</u>.

The forecast probability is f(q)=prob[X>q]. The observed probability is o(q)=prob[Y>q]. Their "joint probability distribution" is denoted g(f,o)

Example of unconditional bias

(f(5.3),o(5.3))		(f(5.3)-o(5.3))
(0.0, 0.0)		(0.0-0.0)=0.0
(0.9, 1.0)		(0.9-1.0)=-0.1
(0.8, 1.0)		(0.8-1.0)=-0.2
(0.7, 1.0)		(0.7-1.0)=-0.3
(1.0, 1.0)	→	(1.0-1.0)=0.0
(0.3, 0.0)		(0.3-0.0)=0.3
(0.1, 0.0)		(0.1-0.0)=0.1
(1.0, 1.0)		(1.0-1.0)=0.0
(0.9, 1.0)		(0.9-1.0)=-0.1
(0.9, 1.0)		(0.9-1.0)=-0.1
…		Bias=-0.04

The forecasts and observations should predict Q>q with the same probability, on average

In other words, bias ≈ 0:

$$1/n \sum_{i=1}^{n} [f_i(5.3) - o_i(5.3)]$$

Example of conditional bias

(f(5.3),o(5.3))	(f(5.3)-o(5.3))
(0.0, 0.0)	(0.0-0.0)=0.0
(0.9, 1.0)	(0.9-1.0)=-0.1
(0.8, 1.0)	(0.8-1.0)=-0.2
(0.7, 1.0)	(0.7-1.0)=-0.3
(1.0, 1.0)	(1.0-1.0)=0.0
(0.3, 0.0)	(0.3-0.0)=0.3
(0.1, 0.0)	(0.1-0.0)=0.1
(1.0, 1.0)	(1.0-1.0)=0.0
(0.9, 1.0)	(0.9-1.0)=-0.1
(0.9, 1.0)	(0.9-1.0)=-0.1
…	C. bias=-0.1

Given **f(5.3)** =0.9, the forecasts are "reliable" if the event is observed 90% of the time, on average

In other words, conditional bias ≈ 0:

$$\frac{1}{|f(5.3) = 0.9|} \sum_{f(5.3)=0.9} [0.9 - o(5.3)]$$

In practice, n>>3 is needed!

Attributes of quality: advanced

g(f,o)=r(o|f)s(f) "Calibration-refinement"

g(f,o)=v(f|o)u(o) "Likelihood-base-rate"

"**Sharpness**" is concerned with **s(f)**

"**Uncertainty**" is concerned with **u(o)**

"**Reliability**" is concerned with **r(o|f)** vs. **s(f)**

"**Resolution**" is concerned with **r(o|f)**

"**Discrimination**" is concerned with **v(f|o)**

"**Type-II bias**" is concerned with **v(f|o)** vs. **u(o)**

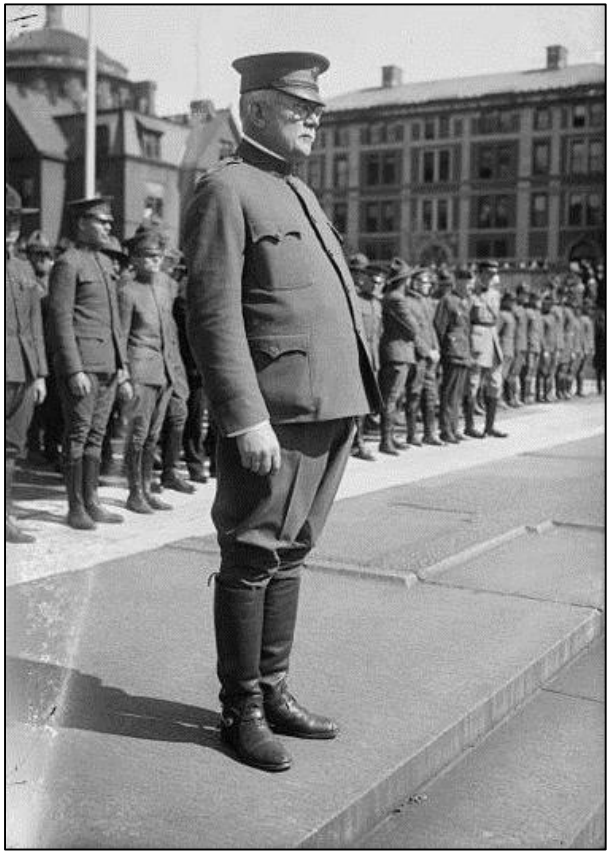
4. Measures of forecast quality

Tips on selecting measures

Things to consider

- Verification may address specific users/applications
- But, should not rely on a single attribute or measure
- Build a picture across several attributes/measures
  - Overall impression of accuracy (total error)
  - Unconditional and conditional biases (directional error)
  - Measures of association (e.g. correlation, discrimination)
  - Skill relative to a baseline
- Be mindful of sample size issues for some measures
- Statistics can be misleading (e.g. for extremes)…

Lies, damned lies and statistics!



John Park Finley: 1854-1943

N=2803		Forecast: Yes	Forecast: No
Observed	Yes	28	23
Observed	No	72	2680

Correct:
28+2680/(28+72+23+2680)=96.5%

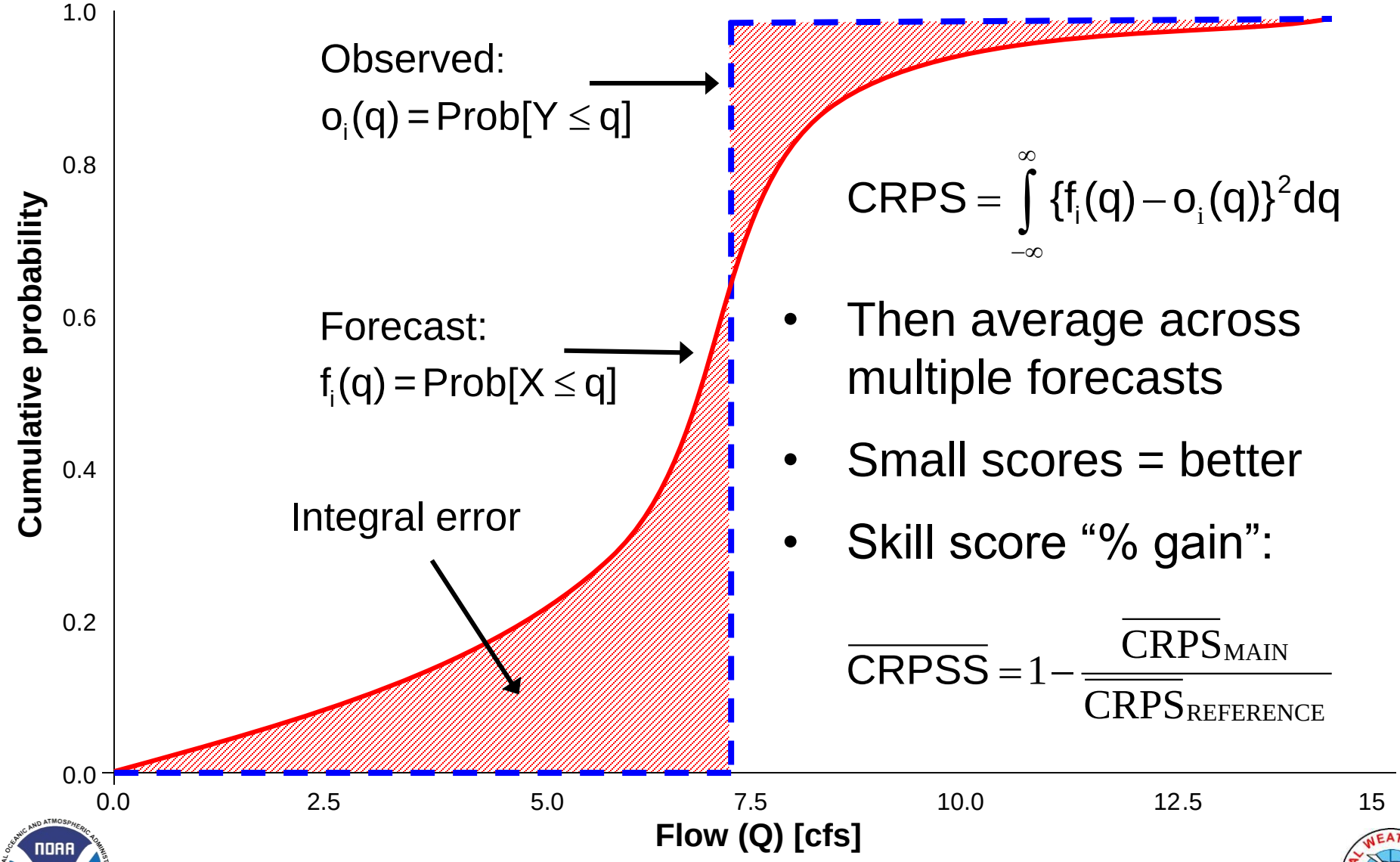
Correct if always forecasting “no tornado”:
72+2680/(28+72+23+2680)=98.1%!

Correct when tornado observed:
28/(28+23)=55%!

What measures in EVS?

Metric name	Feature tested	Discrete events?	Detail
Mean error	Ensemble average	No	Lowest
Relative mean error	Ensemble average	No	Lowest
RMSE	Ensemble average	No	Lowest
Mean absolute error	Ensemble average	No	Lowest
Correlation coefficient	Ensemble average	No	Lowest
Brier Score	Lumped error score	Yes	Low
Mean CRPS	Lumped error score	No	Low
Mean error in prob.	Reliability (unconditional bias)	No	Low
Brier Skill Score	Lumped error score vs. reference	Yes	Low
ROC score	Lumped discrimination score	Yes	Low
Mean CRPSS	Lumped error score vs. reference	No	Low
Spread-bias diagram	Reliability (conditional bias)	No	High
Rank histogram	Reliability (conditional bias)	No	High
Reliability diagram	Reliability (conditional bias)	Yes	High
ROC diagram	Discrimination	Yes	High
Modified box plots	Error visualization	No	Highest

Accuracy (total error): mean CRPS

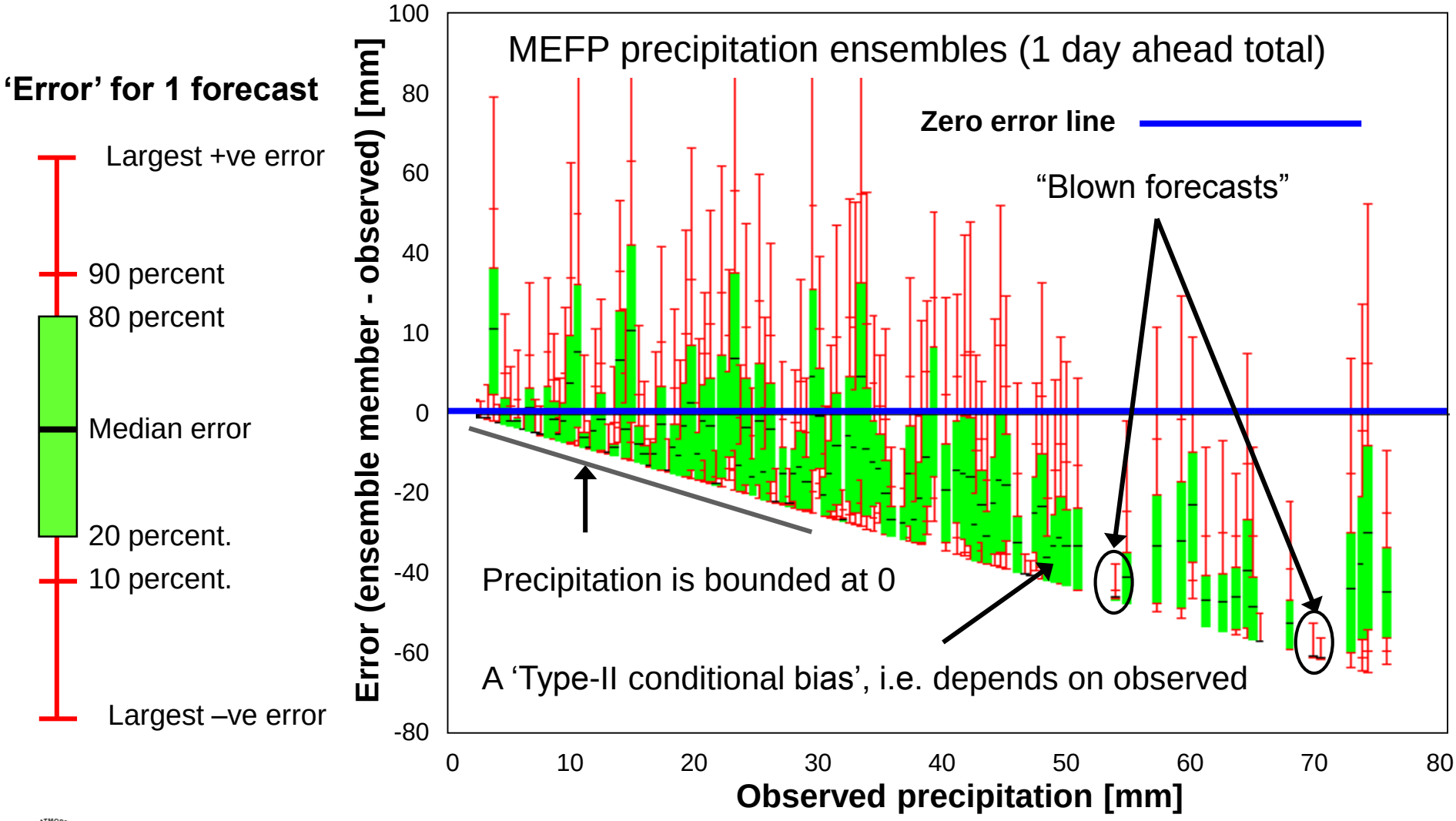


$$\text{CRPS} = \int_{-\infty}^{\infty} \{f_i(q) - o_i(q)\}^2 dq$$

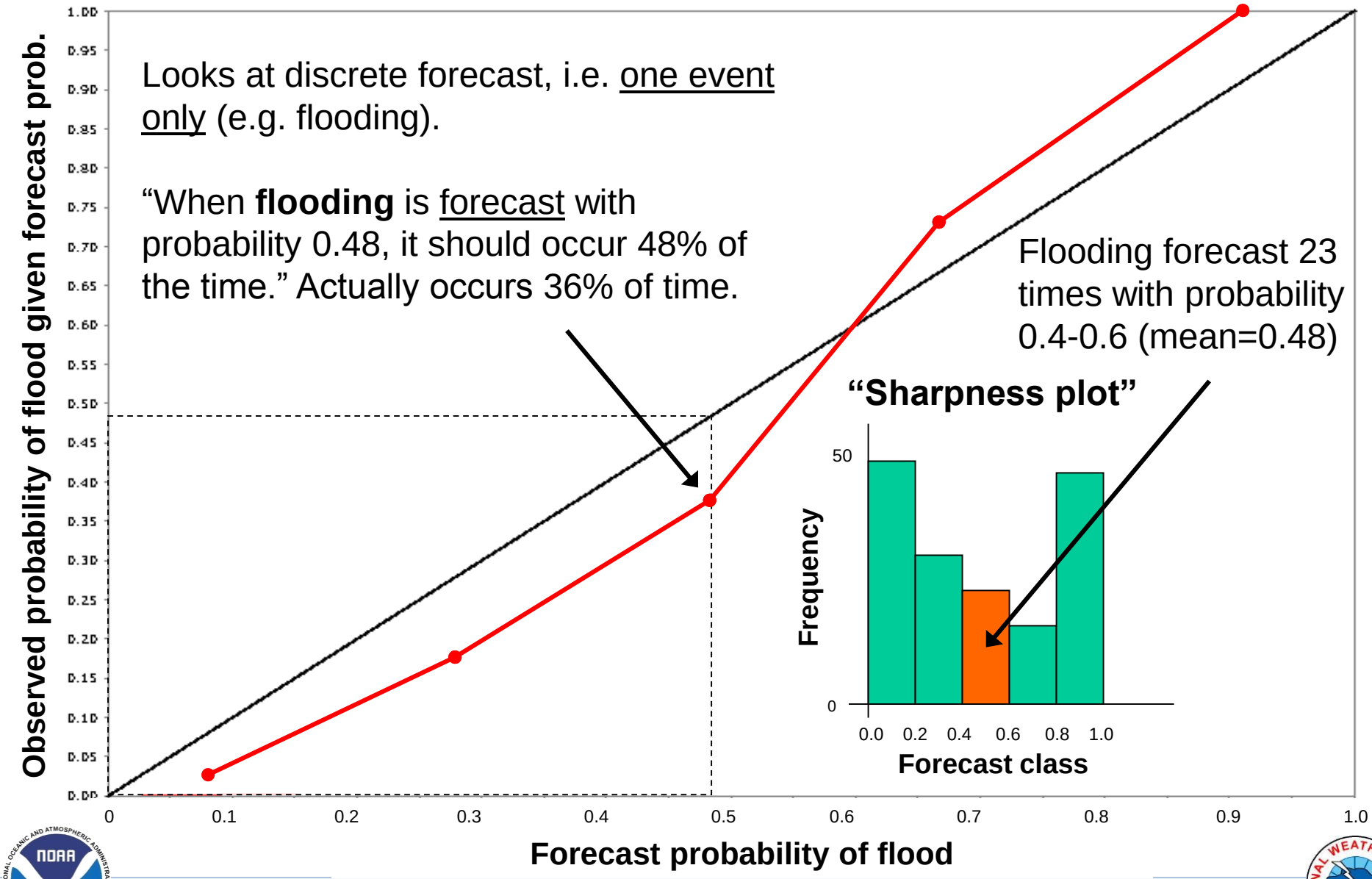
- Then average across multiple forecasts
- Small scores = better
- Skill score "% gain":

$$\overline{\text{CRPSS}} = 1 - \frac{\overline{\text{CRPS}}_{\text{MAIN}}}{\overline{\text{CRPS}}_{\text{REFERENCE}}}$$

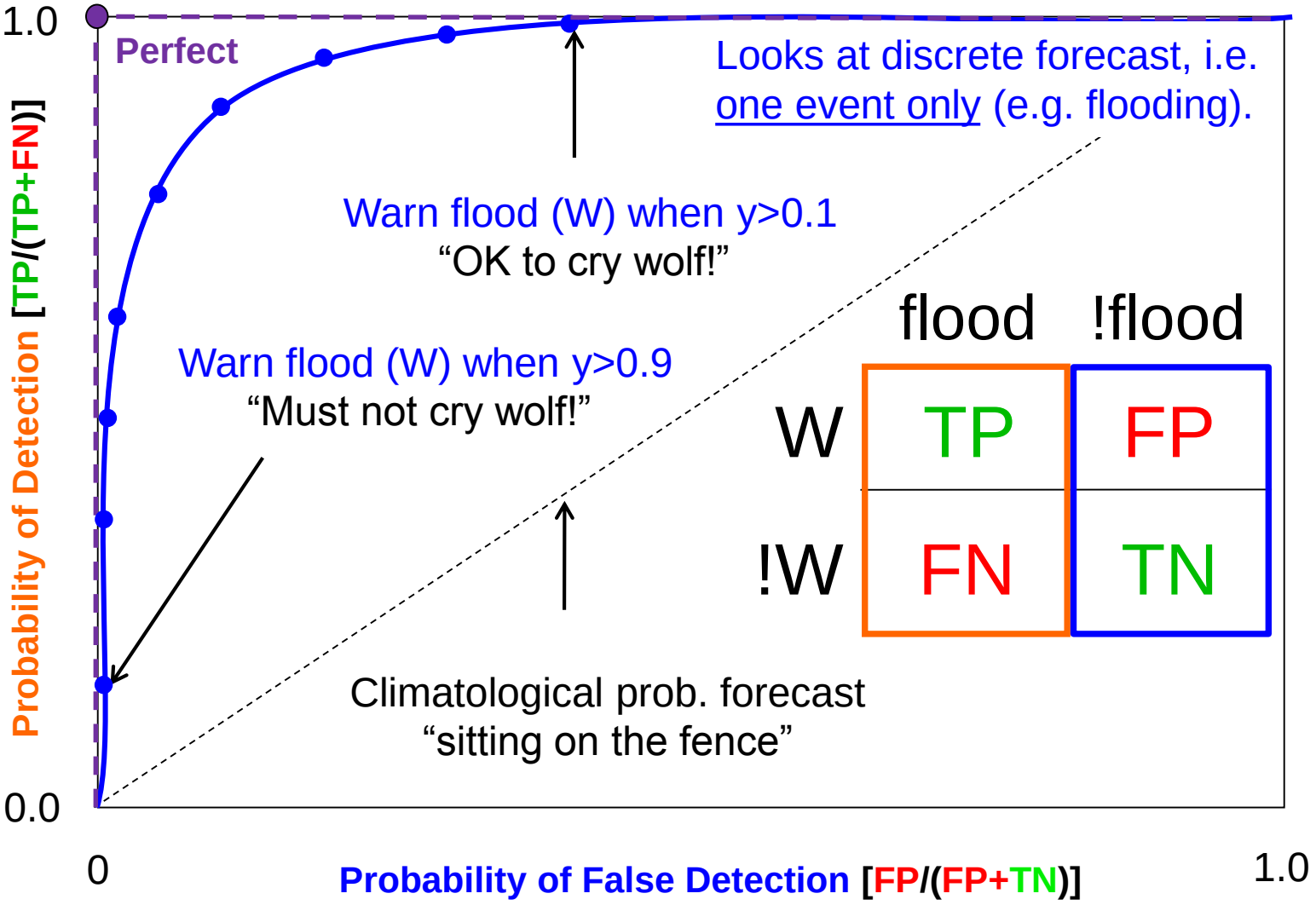
Conditional bias: box plots



Conditional bias: reliability diagram



Discrimination: ROC



5. Final thoughts and suggestions

Final thoughts

Things to consider

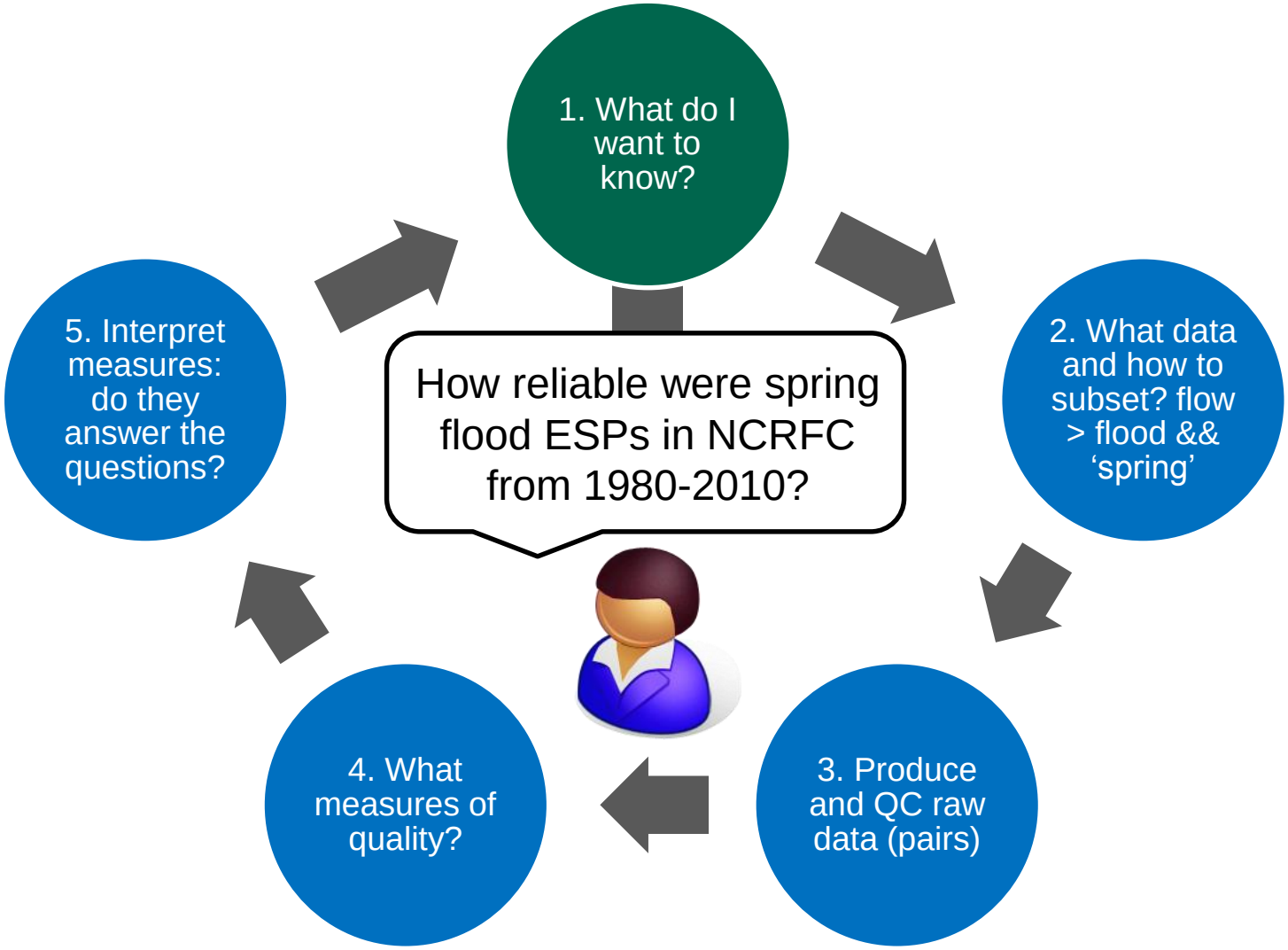
- Try to maximize period and consistency of record
- Ideally QC data/HEFS parameters before verification
- QC the pairs (for 1-2 locations): mistakes are easy
- Consider the scope/users of the verification results
- Consider several attributes and measures of quality
- Include contrasting attributes (e.g. bias/association)
- Be mindful of sample size issues
- Don’t be afraid to explore results iteratively!

Resources and references

- COMET module “Techniques in Hydrologic Forecast Verification”: https://www.meted.ucar.edu/training_module.php?id=453
- CACWR verification page: http://www.cawcr.gov.au/projects/verification/
- Brown, J.D., Demargne, J., Seo, D-J., and Liu, Y. (2010) The Ensemble Verification System (EVS): A software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. *Environmental Modelling and Software*, 25(7), 854-872.
- Demargne, J., Brown, J.D., Liu, Y., Seo, D-J., Wu, l., Toth, Z. and Zhu, Y. (2010) Diagnostic verification of hydrometeorological and hydrologic ensembles. *Atmospheric Science Letters*, 11(2), 114-122.
- Jolliffe, I.T., and Stephenson, D.B. (2011) *Forecast Verification: A Practitioner’s Guide in Atmospheric Science*. 2nd ed. John Wiley and Sons: Chichester.
- Wilks, D.S. (2006) *Statistical Methods in the Atmospheric Sciences.* 2nd ed. Elsevier: San Diego.
- Murphy, A.H. and Winkler, R.L. (1987) A general framework for forecast verification. *Monthly Weather Review*, 115, 1330-1338.

Extra slides

How to verify? The key steps.



EVS standalone (GUI mode)

Structured user interface

- Navigate through stages of verification study

1. Verification (per location)

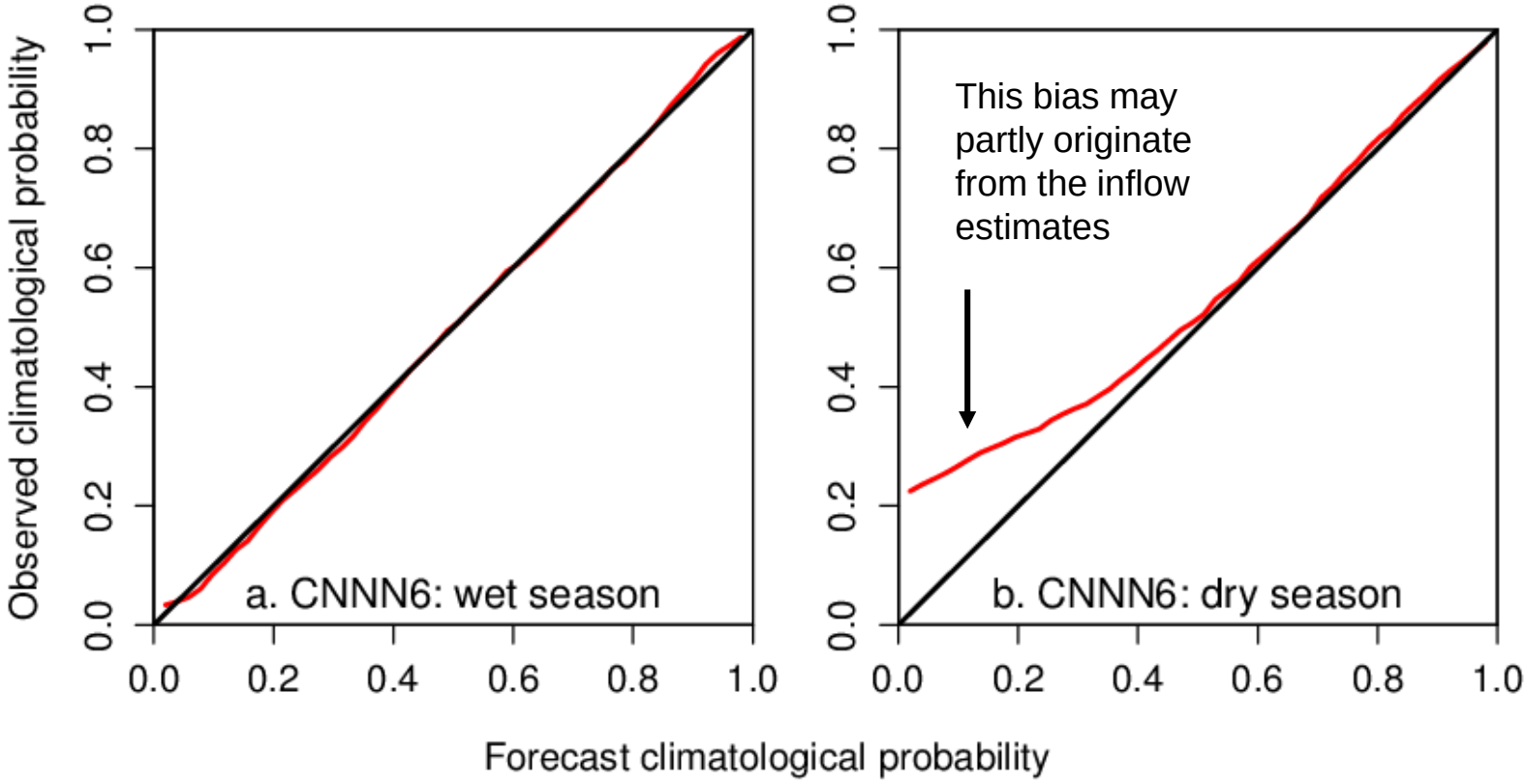
- Specify locations, data sources, metrics etc.

2. Aggregation (many locations): option

- Choose locations, aggregation method etc.

3. Output (graphical and numerical)

Data QC example



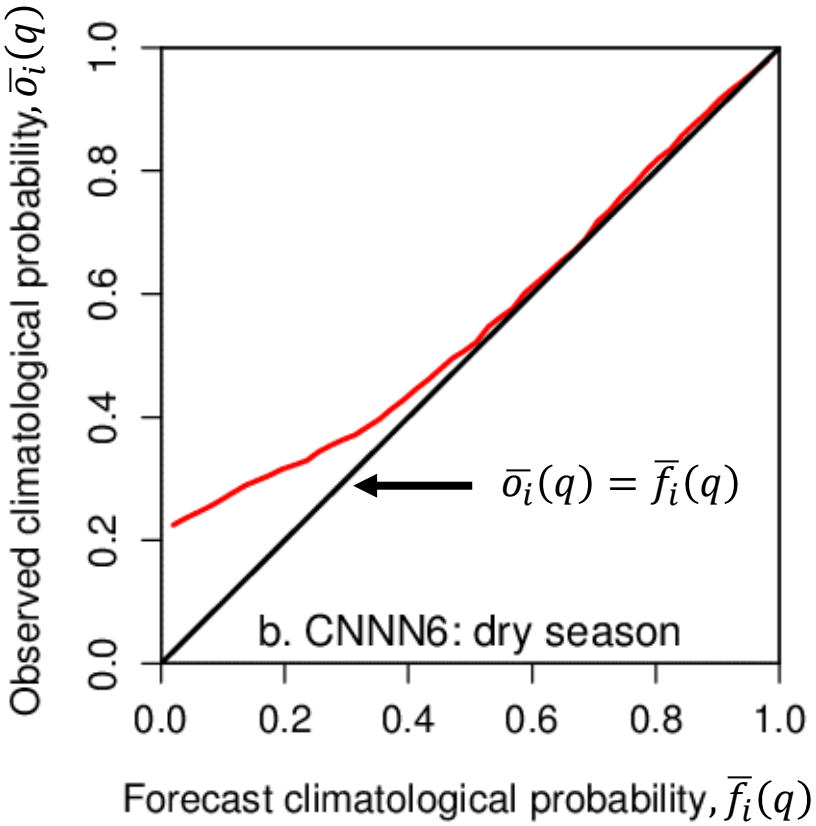
- Cannonsville, NY (CNNN6): reservoir inflows are estimated
- Inflow estimates do not include evaporation = biases in dry conditions
- Data QC problems can be insidious (e.g. masked by model errors)

Pairing tips

Things to remember when pairing

- Forecasts/simulations in UTC (12Z, Δt=1 or 6 hours)
- Observations in local time (e.g. 5Z, 11Z,.. in MARFC)
- Observations generally enforced as CST for pairing…
- …avoids interpolation, but adds error for non-CST
- …except where forecasts are hourly (then, no error)
- Remember, wrong pairs can be created quite easily…
- …especially when forecasts are hourly (CB, CN)
- So, always QC the pairs (see exercises)!

Unconditional bias: MEPD



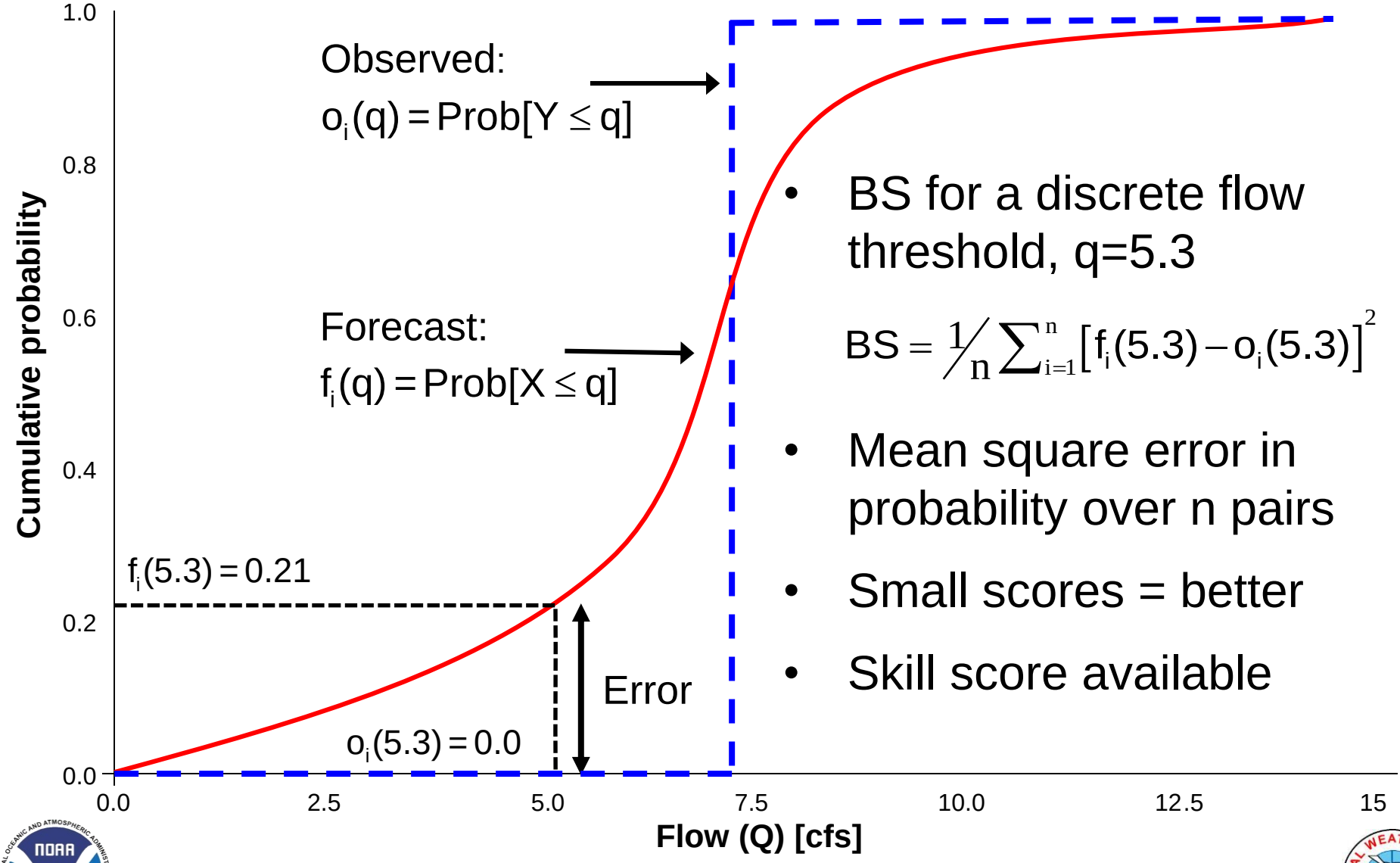
$$\overline{f}_i(q) = {}^1\!/_n \sum_{i=1}^{n} f_i(q) \quad \forall q$$

$$\overline{o}_i(q) = {}^1\!/_n \sum_{i=1}^{n} o_i(q) \quad \forall q$$

Unbiased: $E[f(q) - o(q)] = 0$

- Recall example of Cannonsville, NY (CNNN6) with dry bias
- Mean Error of Probability Diagram: average forecast CDF vs. observed
- Shows climatological bias in the forecasts, i.e. mean probability error

Accuracy (total error): Brier Score



- BS for a discrete flow threshold, q=5.3

$$BS = \frac{1}{n}\sum_{i=1}^{n}\left[f_i(5.3) - o_i(5.3)\right]^2$$

- Mean square error in probability over n pairs
- Small scores = better
- Skill score available