5th HEFS workshop, 02/25/2014

Seminar D: ensemble verification concepts and requirements

James Brown

james.brown@hydrosolved.com

Contents

1. Motivations for verification

Why verify?

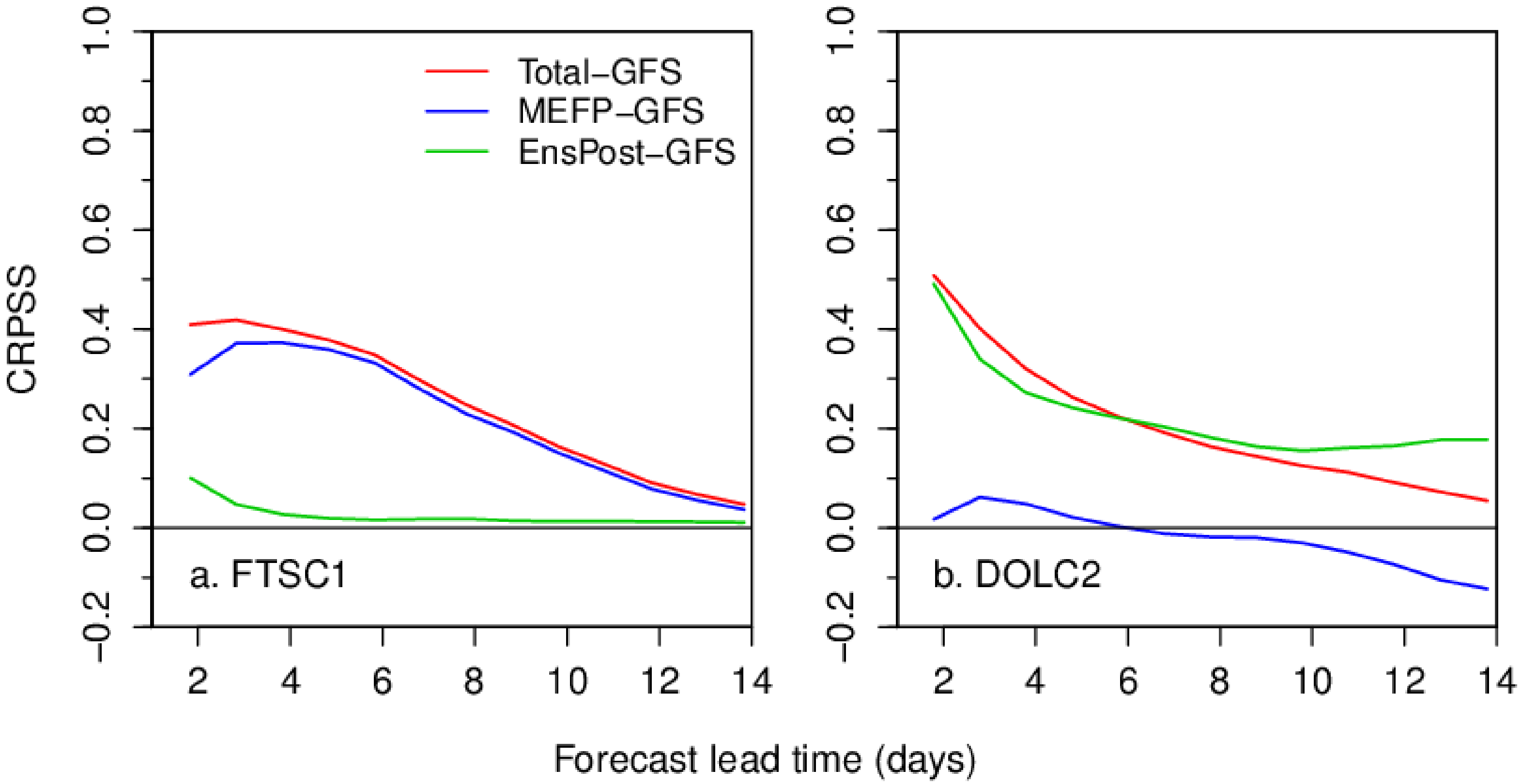
Forecasts incomplete if quality unknown

- 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

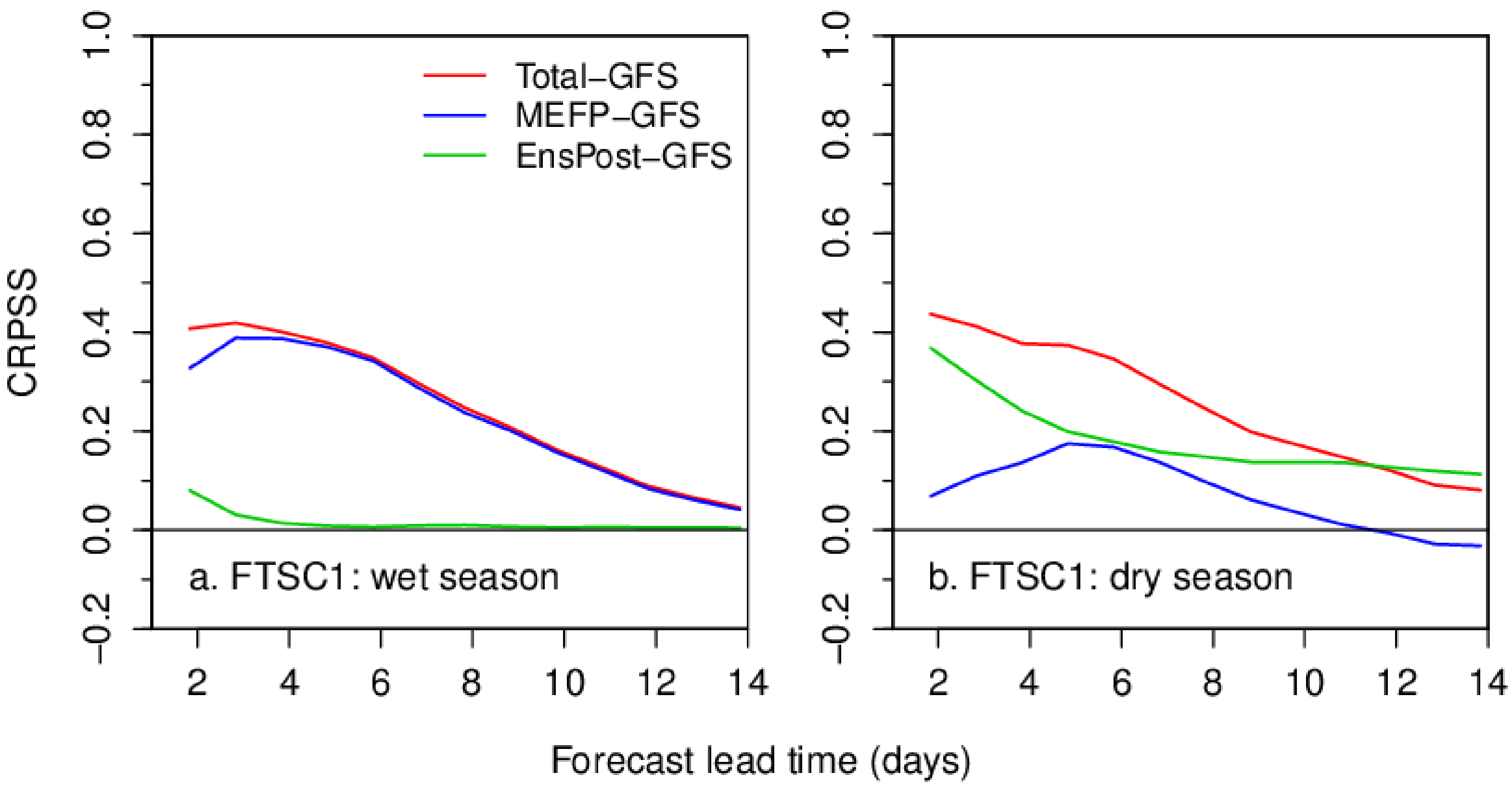
- Total uncertainty = meteorological + hydrological
- HEFS = MEFP + EnsPost
- Component error analysis can separate the two

Example: two very different basins



- Fort Seward, CA (FTSC1) and Dolores, CO (DOSC1)
- 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)

2. Data requirements

What data are required?

Datasets

- Hindcasts or archived forecasts (forcing and flow)
- Reliable observations (e.g. no major ratings biases)
- Hydrologic simulations for component error analysis
- Large sample (long record) and consistent record

Verification sample size depends on

- Period of record and frequency of T0s
- Aggregation period
- Sub-setting of data (“conditional verification”)

How to mitigate small sample?

Steps to reduce impacts

- Hindcasting (see earlier)
- Be careful with conditioning (i.e. avoid small subsets)
- Be careful with aggregation (e.g. monthly volumes)
- Choose verification metrics that summarize quality
- Can set minimum sample size in EVS (p.104 manual)

Steps to assess impacts

- Qualitative: check sample size plots in EVS
- Quantitative: compute confidence intervals (p.48)

Data quality control (QC)

Before hindcasting: QC input data

- Use MEFP/EnsPost data and parameter diagnostics
- Check for non-physical values and outliers

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
- Check verification pairs carefully…

Pairing mechanics and QC

- Pairing often requires assumptions/data manipulation
- For example, aggregation or re-timing of data
- E.g. Forecast (SQIN) vs. QME in ABRFC (GMT-6)
- <u>Always QC the pairs</u> (e.g. for 1-2 locations)!

3. Attributes of forecast quality

First, the big picture

Three separate, but related, concepts

- **Quality:** synonymous w/ verification (vs. observations)
- **Utility:** service is fit for purpose (includes quality)
- **Consistency:** forecasters not “gaming” the system

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)

Decades of publications on quality!

- Interested in forecast errors (forecast - observed)
- John Park Finley (1884): tornado verification
- Murphy and Winkler (1987): attributes of quality
- Books: Jolliffe and Stephenson (2011), Wilks (2006)
- http://www.cawcr.gov.au/projects/verification/
- http://hepex.irstea.fr/what-is-a-good-forecast/

HSL

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 <u>skill</u>
- Skill is valuable, but choice of baseline needs thought
  - 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)

Attributes of quality

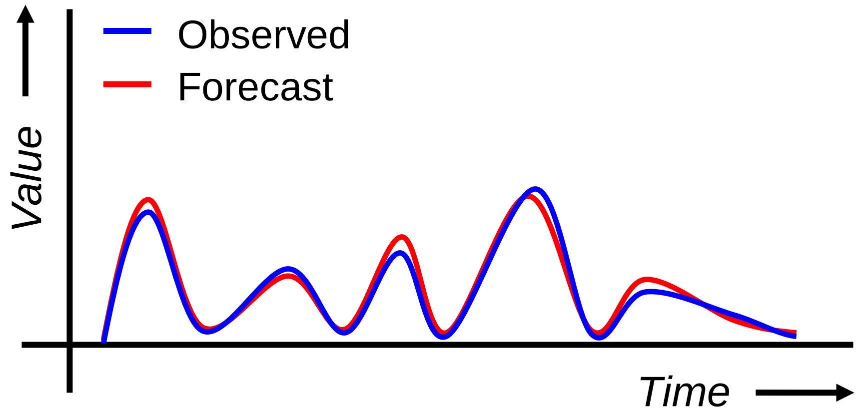
What is meant by attribute here?

- A "desirable" property of a forecasting system
- Specifically, a desirable relationship with observations
- A forecasting system has multiple attributes of quality
- Three, well-known from deterministic forecasting…

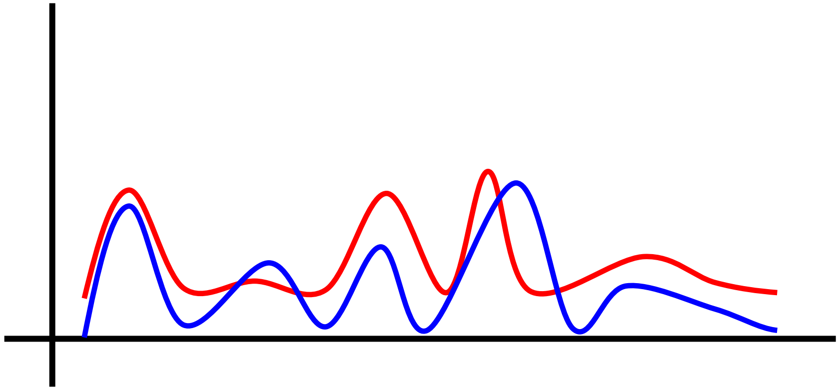
Accuracy, bias, and association

- Accuracy: generic term for <u>total</u> error (e.g. MSE)
- Bias: generic term for a <u>directional</u> error (e.g. ME)
- Association: generic for <u>correspondence</u> (e.g. COV)

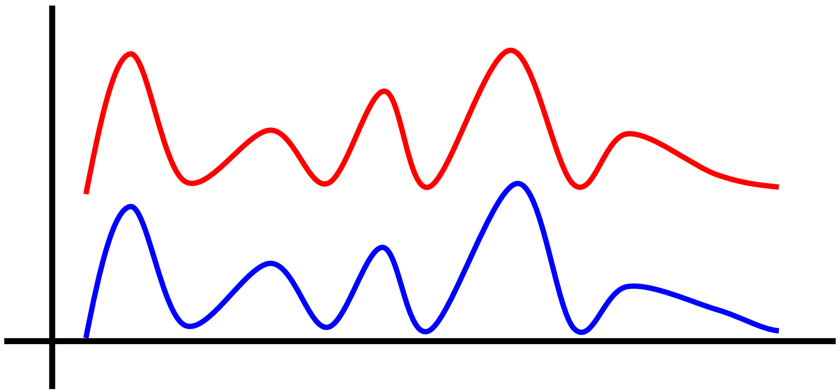
Attributes of quality: examples



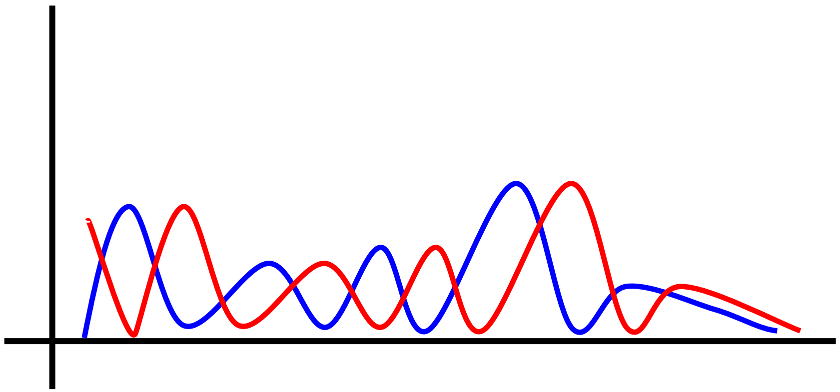
- Unbiased
- Strong association
- High accuracy (small total error)



- 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)

Unconditional vs. conditional quality

- **Unconditional**
  - All data, no subsets (except by forecast lead time)
- **Conditional**
  - Many possible conditions; season, flow amount etc.

Let’s look at some ensemble forecasts…

Ensemble forecasts: raw data

(X,Y)

({1.1,…,3.3}, 3.2)

({2.6,…,21.5}, 20.2)

({3.2,…,19.8}, 18.2)

({4.5,…,12.5}, 13.4)

({13.5,…,28.3}, 24.1)

({0.2,…,7.8}, 2.1)

({0.1,…,5.4}, 5.3)

({7.3,…,16.5}, 12.4)

({2.5,…,40.1}, 30.5)

({4.9,…,57.3}, 47.2)

…

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

→

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)

(f(5.3),o(5.3))

(0.0, 0.0)

(0.9, 1.0)

(0.8, 1.0)

(0.7, 1.0)

(1.0, 1.0)

(0.3, 0.0)

(0.1, 0.0)

(1.0, 1.0)

(0.9, 1.0)

(0.9, 1.0)

…

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:

$$\frac{1}{n}\sum_{i=1}^{n}\left(f_i(5.3) - o_i(5.3)\right) \approx 0$$

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
…	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:

$$\frac{1}{|f(5.3)=0.9|}\sum_{f(5.3)=0.9}\left(0.9-o(5.3)\right)\approx 0$$

In practice, n>>3 is needed!

Attributes of probability forecasts

$g(f,o)=r(o|f)s(f)$ "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. $u(o)$

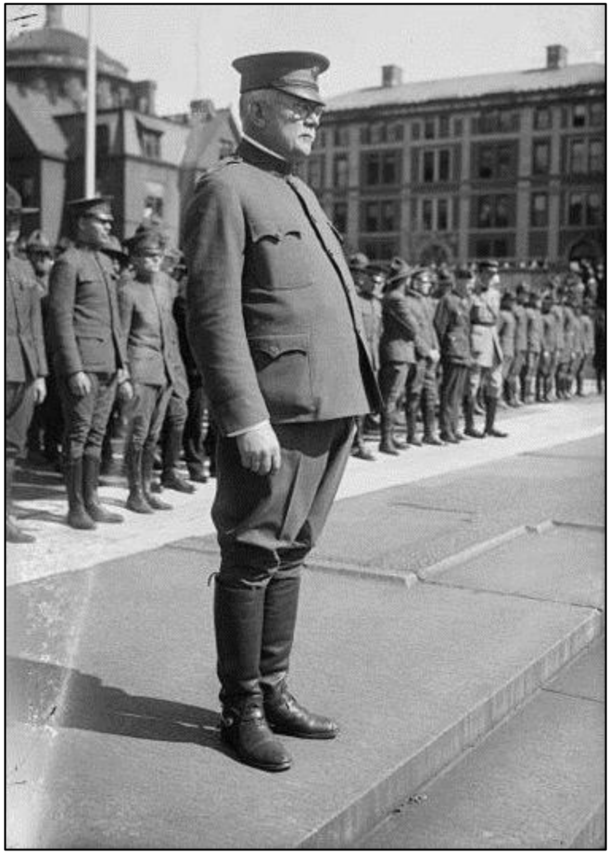
4. Measures of forecast quality

Tips on selecting measures

Things to consider

- The study may address specific users/applications
- But, do <u>not</u> rely on any single measure of quality
- Build a picture across several attributes of quality
  - Overall impression of accuracy (total error)
  - Unconditional and conditional biases (directional error)
  - Measures that are insensitive to bias (correlation, discrimination)
  - Skill relative to a baseline (remember skill reflects the baseline!)
- Be mindful of sample size issues
- Extreme events: be mindful of non-occurrences!...

Extreme events: tornado forecasts



John Park Finley: 1854-1943

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

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

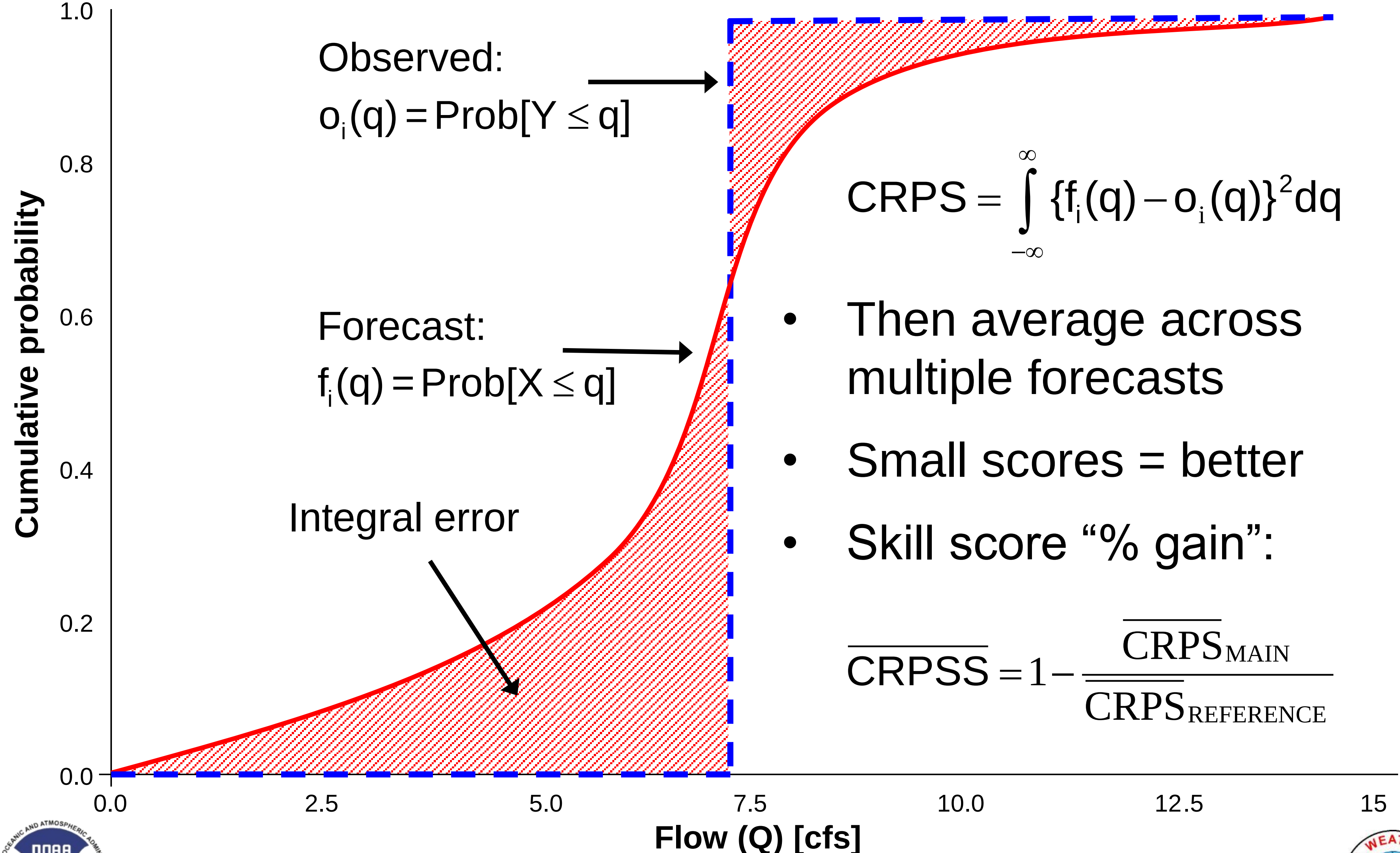
Correct if always forecasting “no tornado”:
72+2680/(28+72+23+2680)=98.1%!

Correct when tornado observed:
28/(28+72)=28%!

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

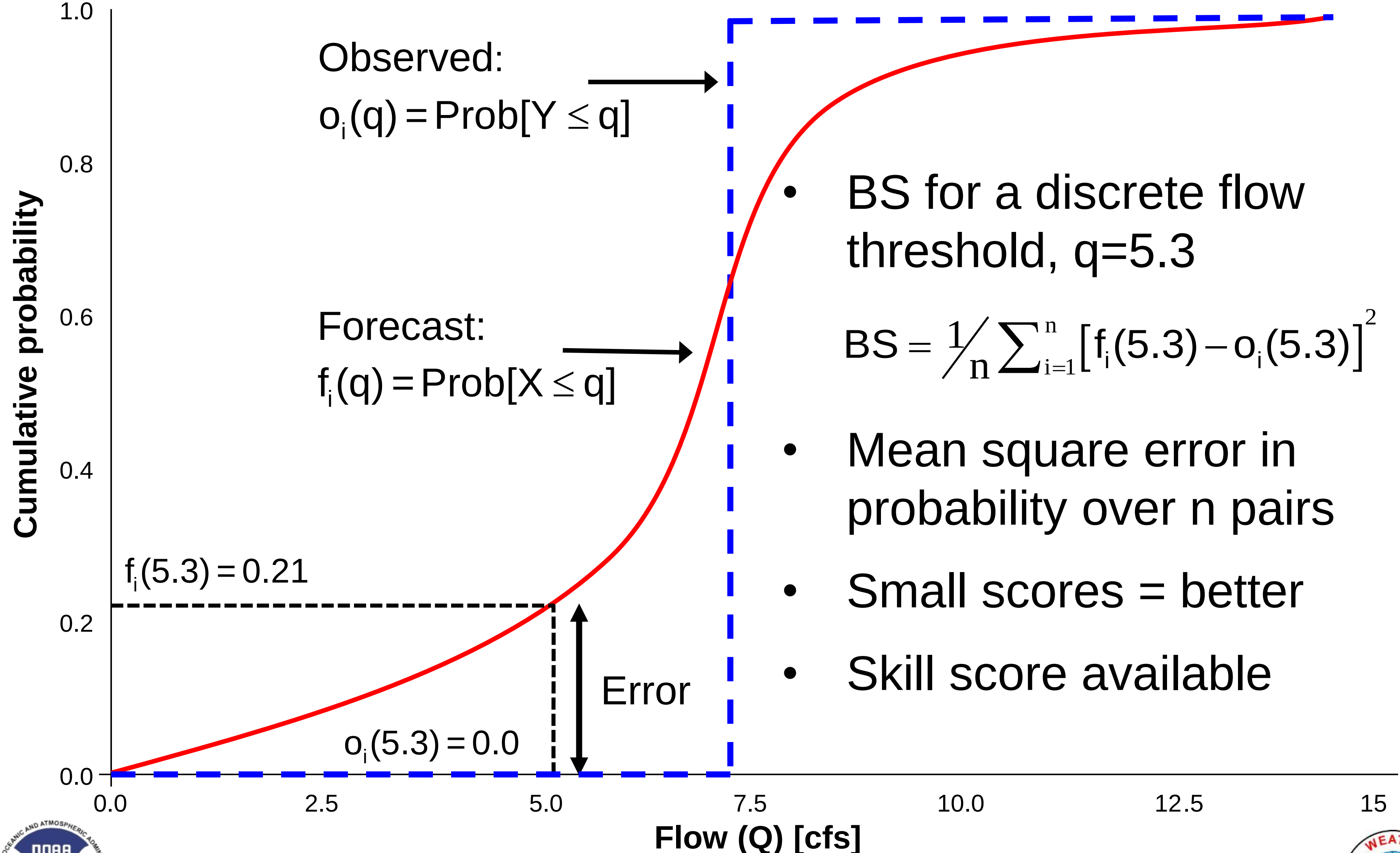


$$\mathrm{CRPS} = \int_{-\infty}^{\infty} \{f_i(q) - o_i(q)\}^2 dq$$

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

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

Accuracy (total error): Brier Score

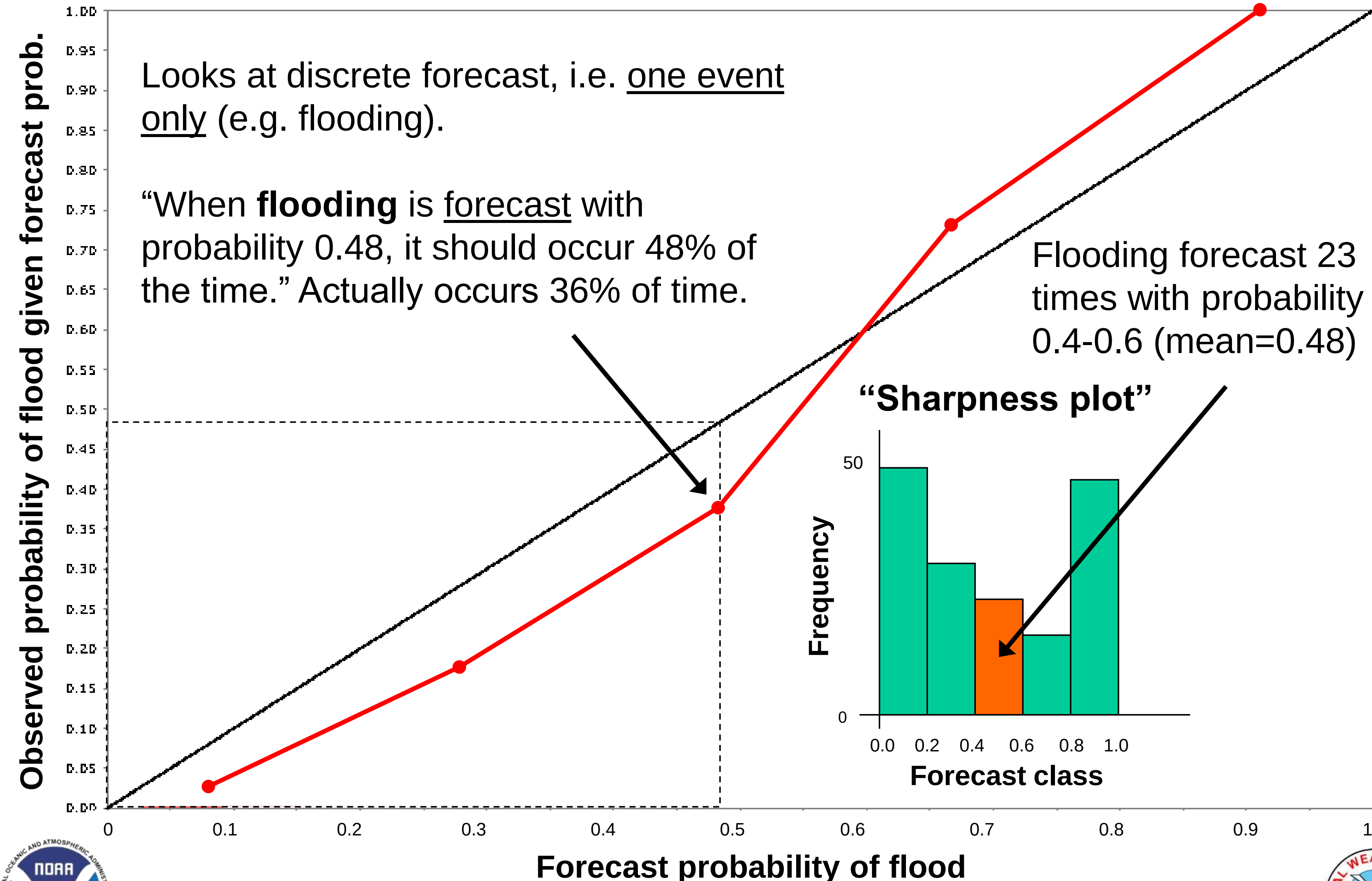


- BS for a discrete flow threshold, q=5.3

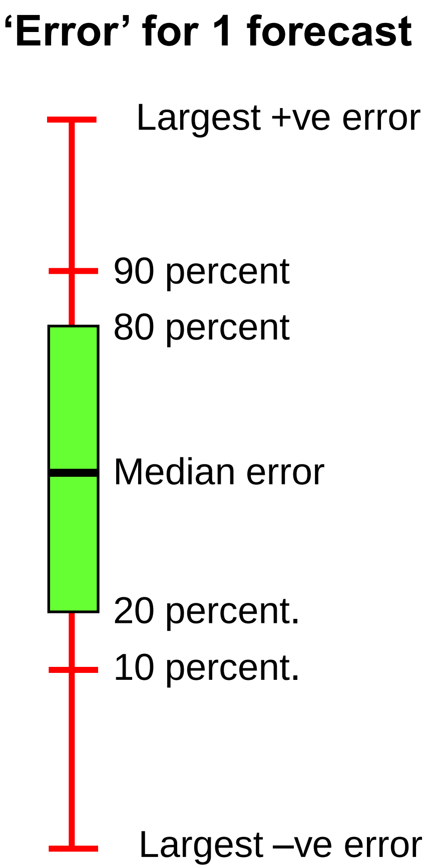
$$BS = {}^{1}\!/_{n} \sum_{i=1}^{n} [f_i(5.3) - o_i(5.3)]^2$$

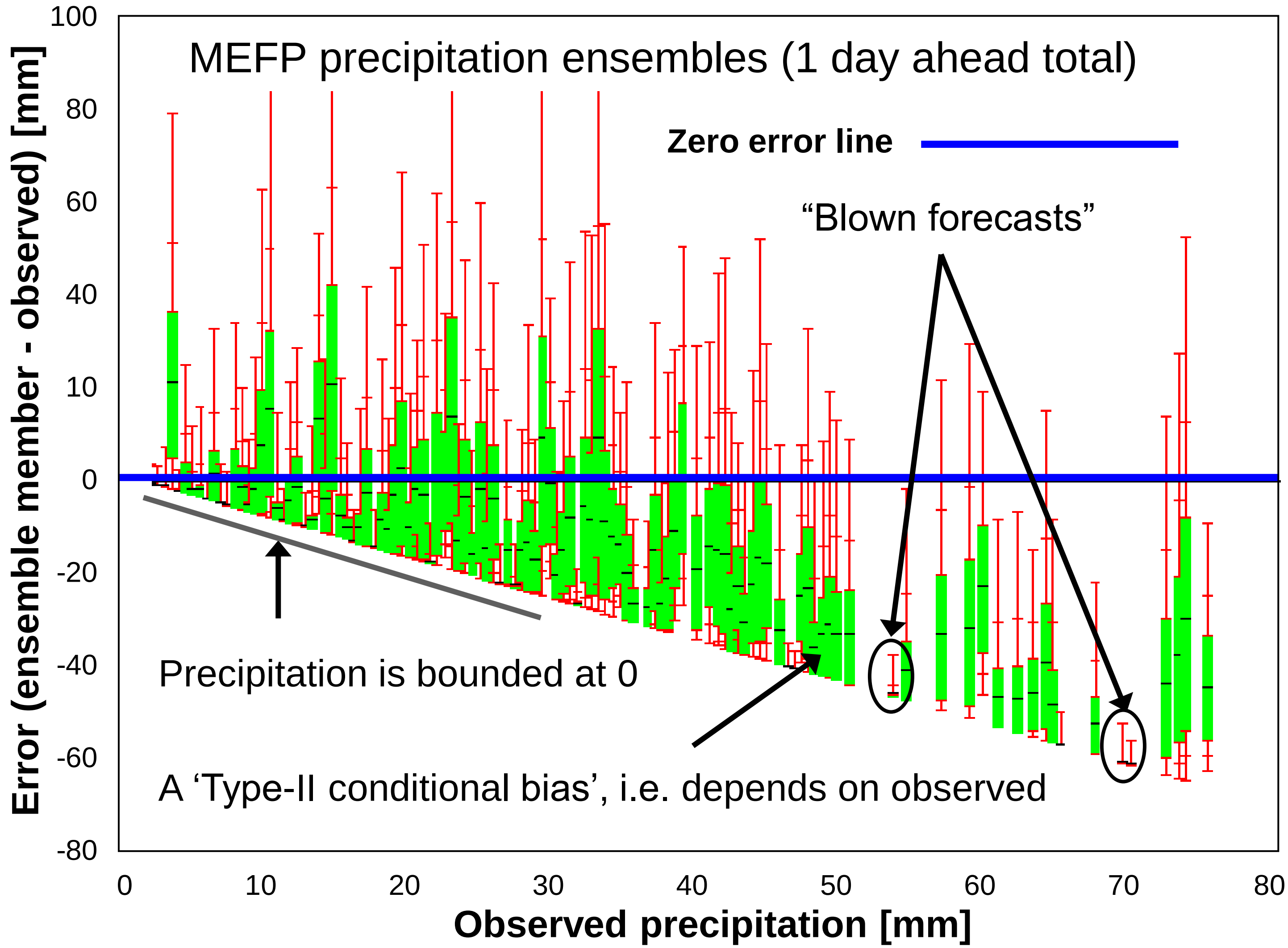
- Mean square error in probability over n pairs
- Small scores = better
- Skill score available

Conditional bias: reliability diagram

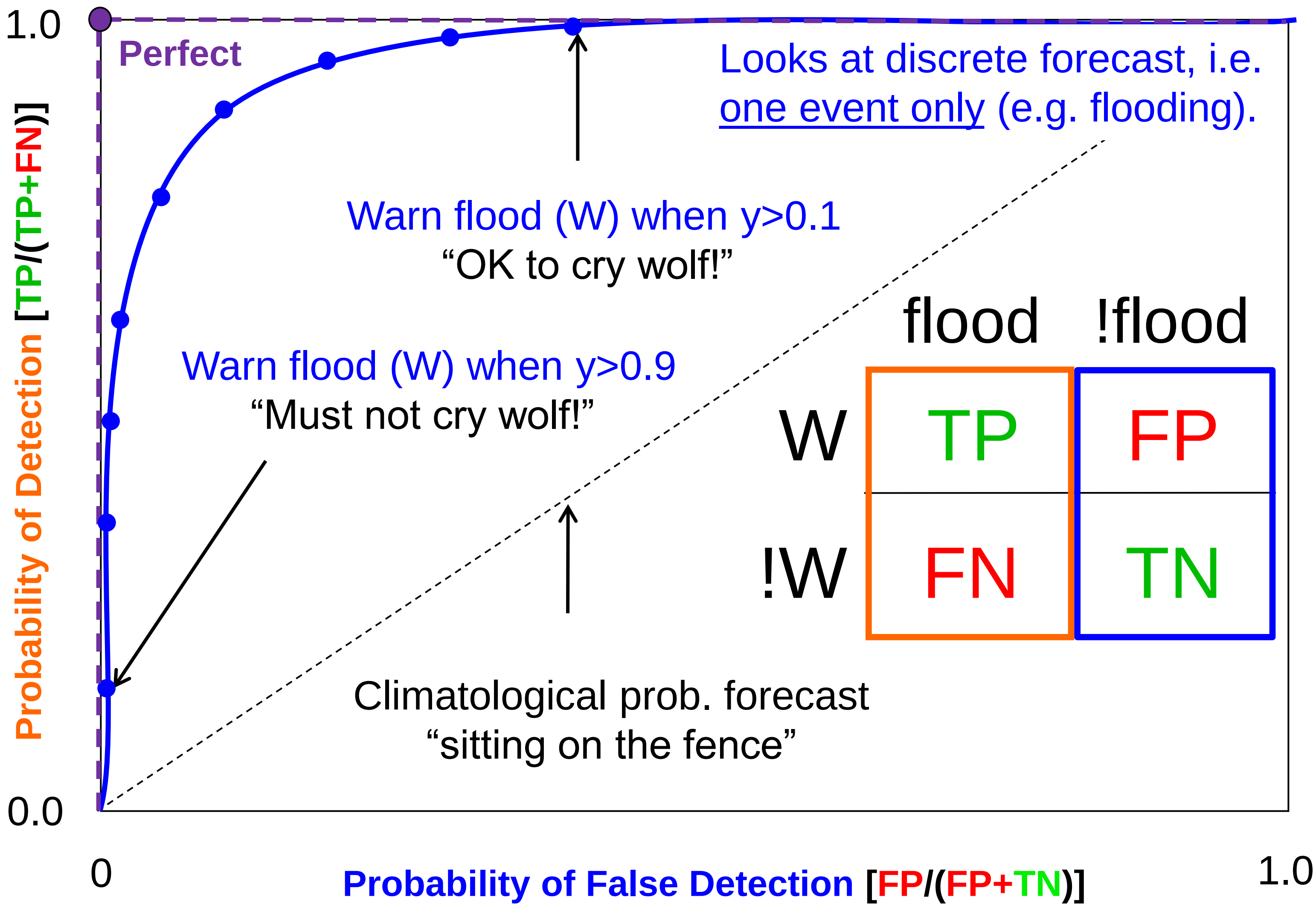


Conditional bias: box plots





Discrimination: ROC



5. Final thoughts and suggestions

Final thoughts

Things to consider

- Try to maximize period and consistency of record
- Due diligence before verification (data/calibration QC)
- Always QC the paired data, as mistakes easily made
- Identify the scope/users of the verification (questions)
- Consider several <u>attributes</u> and <u>measures</u> of quality
- Consider contrasting attributes (e.g. bias/association)
- Be mindful of sample sizes and verify accordingly
- 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

- 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.

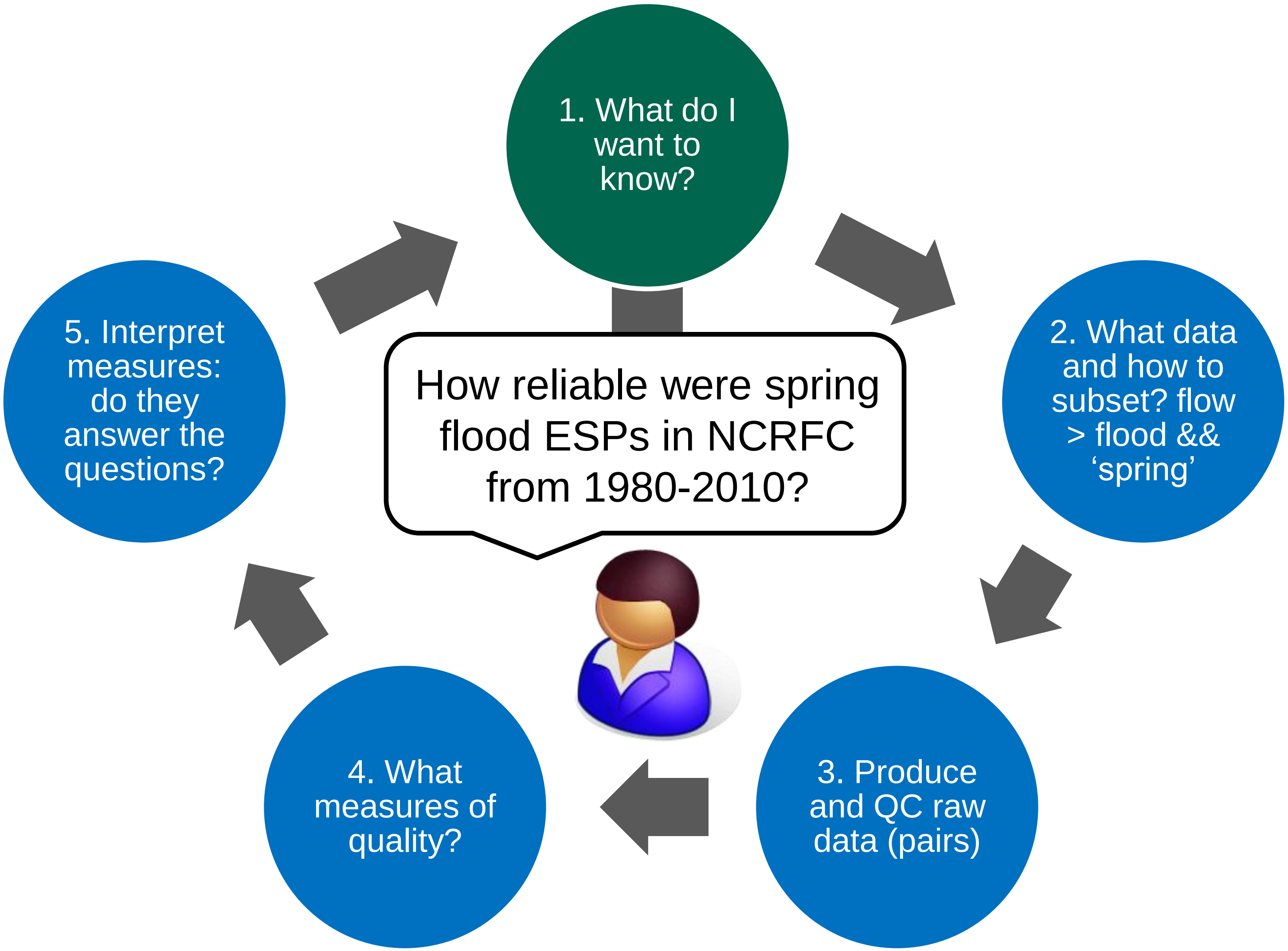
- Demargne, J., Brown, J.D., Liu, Y., Seo, D-J., Wu, l., Toth, Z. and Zhu, Y. (2010) Diagnostic verification of hydrometeorological and hydrologic ensembles. *Atmospheric Science Letters*, 11(2), 114-122.

- Jolliffe, I.T., and Stephenson, D.B. (eds). (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.

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)

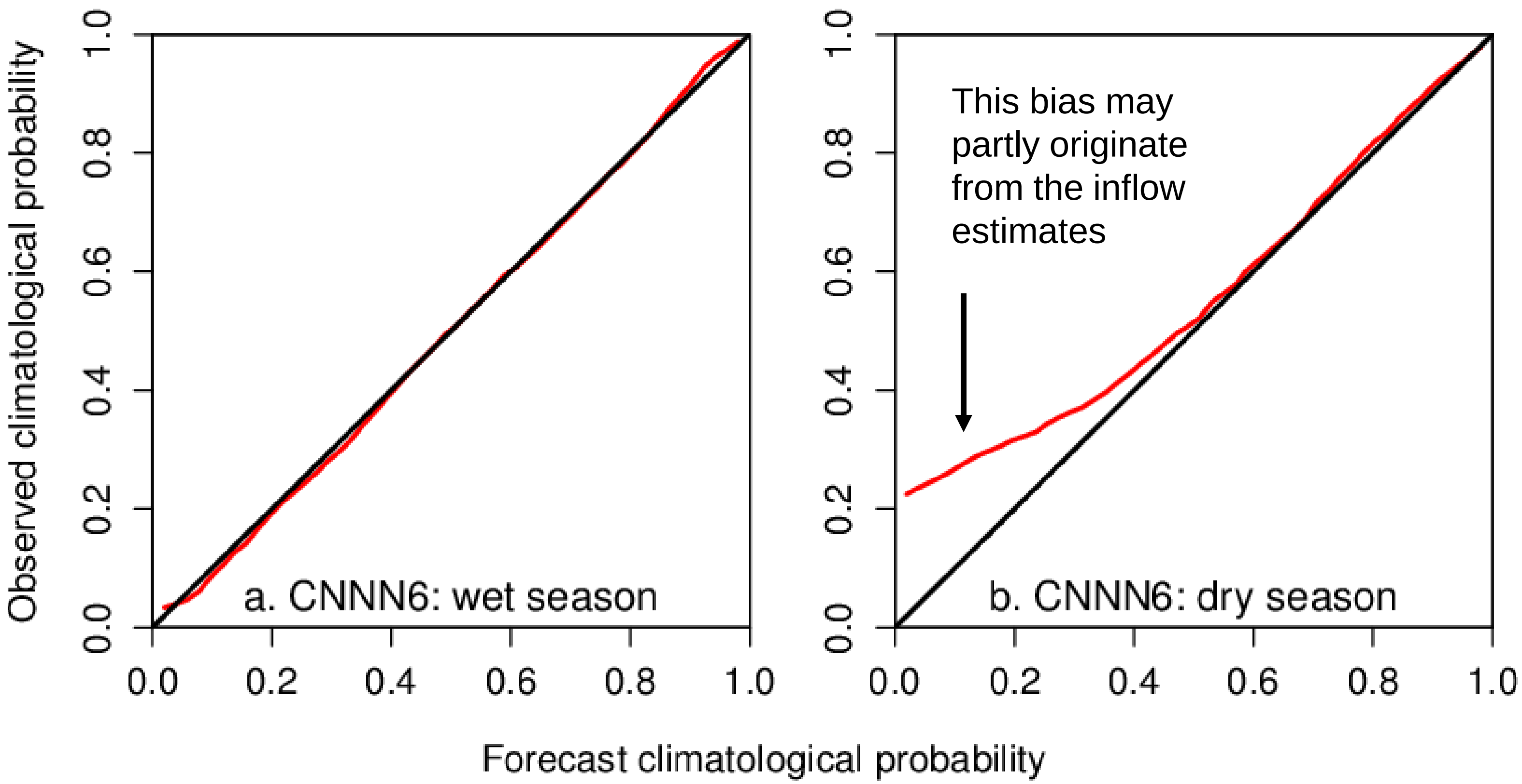
- Specify locations, data sources, metrics etc.

2. Aggregation (many locations): <u>option</u>

- Choose locations, aggregation method etc.

3. Output (graphical and numerical)

Data QC example



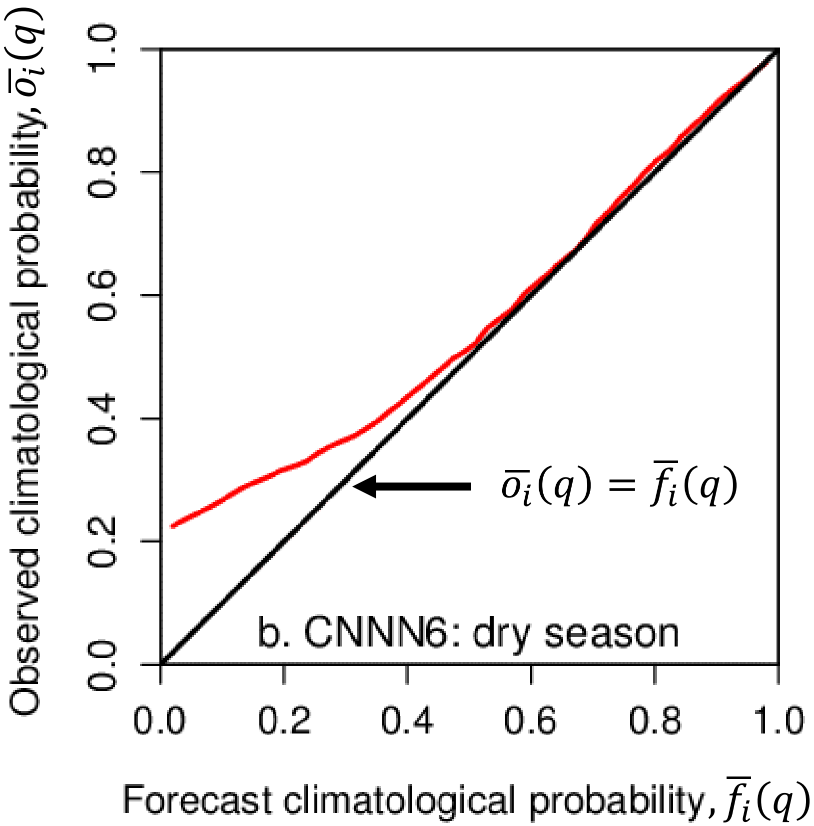
- Cannonsville, NY (CNNN6): reservoir inflows are estimated
- Inflow estimates do not include evaporation = biases in dry conditions
- Data QC problems can be insidious (e.g. masked by model errors)

HSL

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