4th HEFS workshop, 09/19/2013

Seminar C: results from the scientific validation of the Hydrologic Ensemble Forecast Service (HEFSv1)

James Brown

james.brown@hydrosolved.com

Contents

1. Overview of the phased evaluation of the HEFSv1

Motivation and scope

What it aimed to do

- Test critical features and screen for bugs/issues
- Demonstrate unbiasedness and skillfulness
- Provide guidance on expected quality
- Support early field applications (e.g. NYCDEP)

What it did not aim not do

- Benchmark HEFS against operational forecasts
- Cover a broad range of basins and use cases
- Provide guidance on calibration of HEFS

Three initial phases

For completion by the end of 2013

Phase I: medium-range (1-14 days), GFS (discontinued)

- Selected basins in four RFCs (AB, CB, CN, MA)
- Report available now (hyperlink)

Phase II: long-range (1-330 days), GEFS+CFSv2+CLIM

- Selected basins in MA and NE (in support of NYCDEP)
- Report on track for 30th September 2013

Phase III: medium-range, GEFS (as in Phase I)

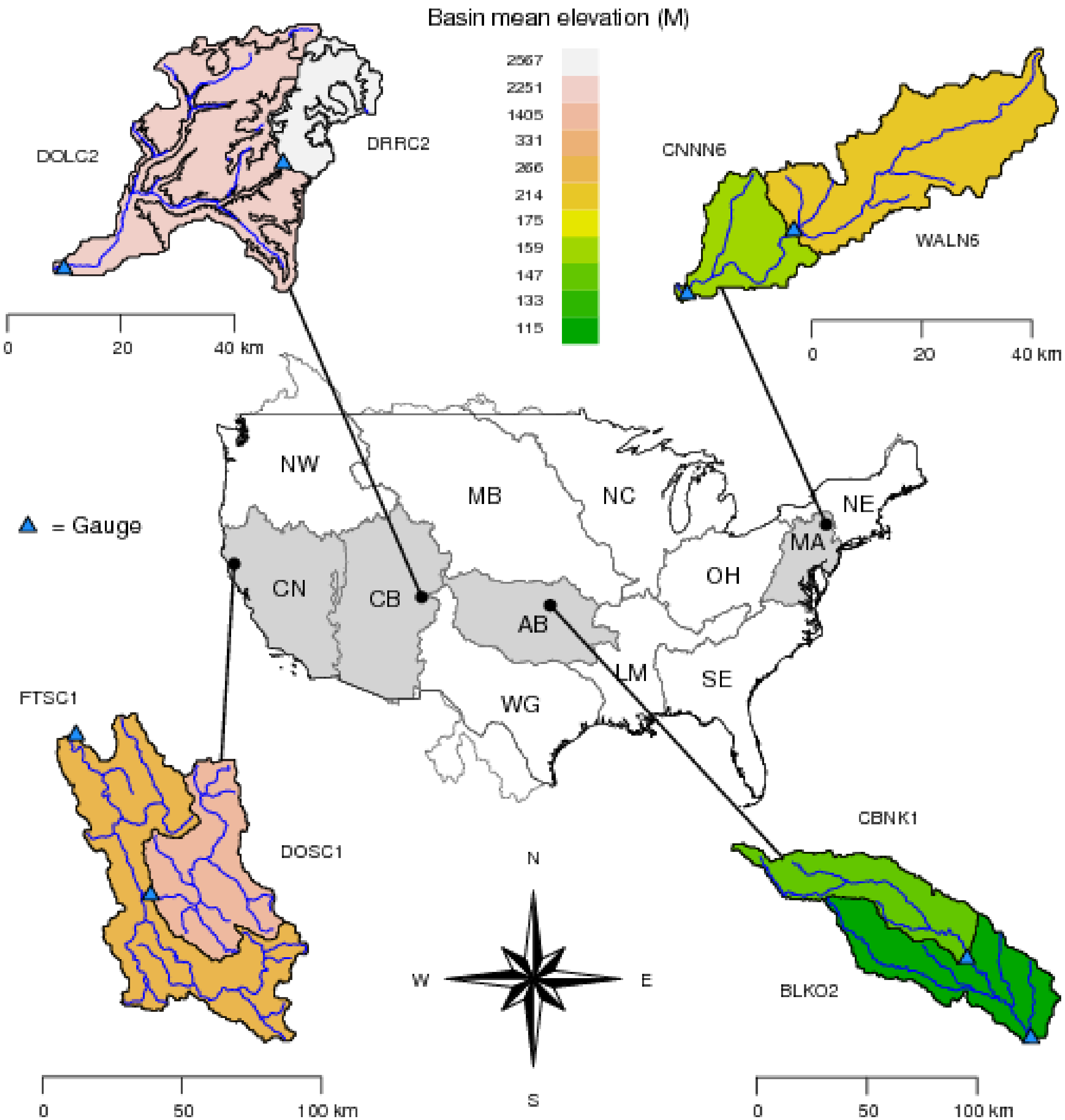
- Same design as Phase I, to establish gain from GEFS
- Report due 31st December 2013

2. Phase I: medium-range with frozen GFS

Phase I basins

Basins

- Four RFCs
- Hindcasts: 1979-1999
- Upper/lower pairing
- USGS gauge at the outlet of each basin
- Relatively small basins (largest 2000 sq. miles)
- Low elevations in AB and MA
- Higher elevations in CB and CN
- CB and CN have MAT/MAP sub-basins

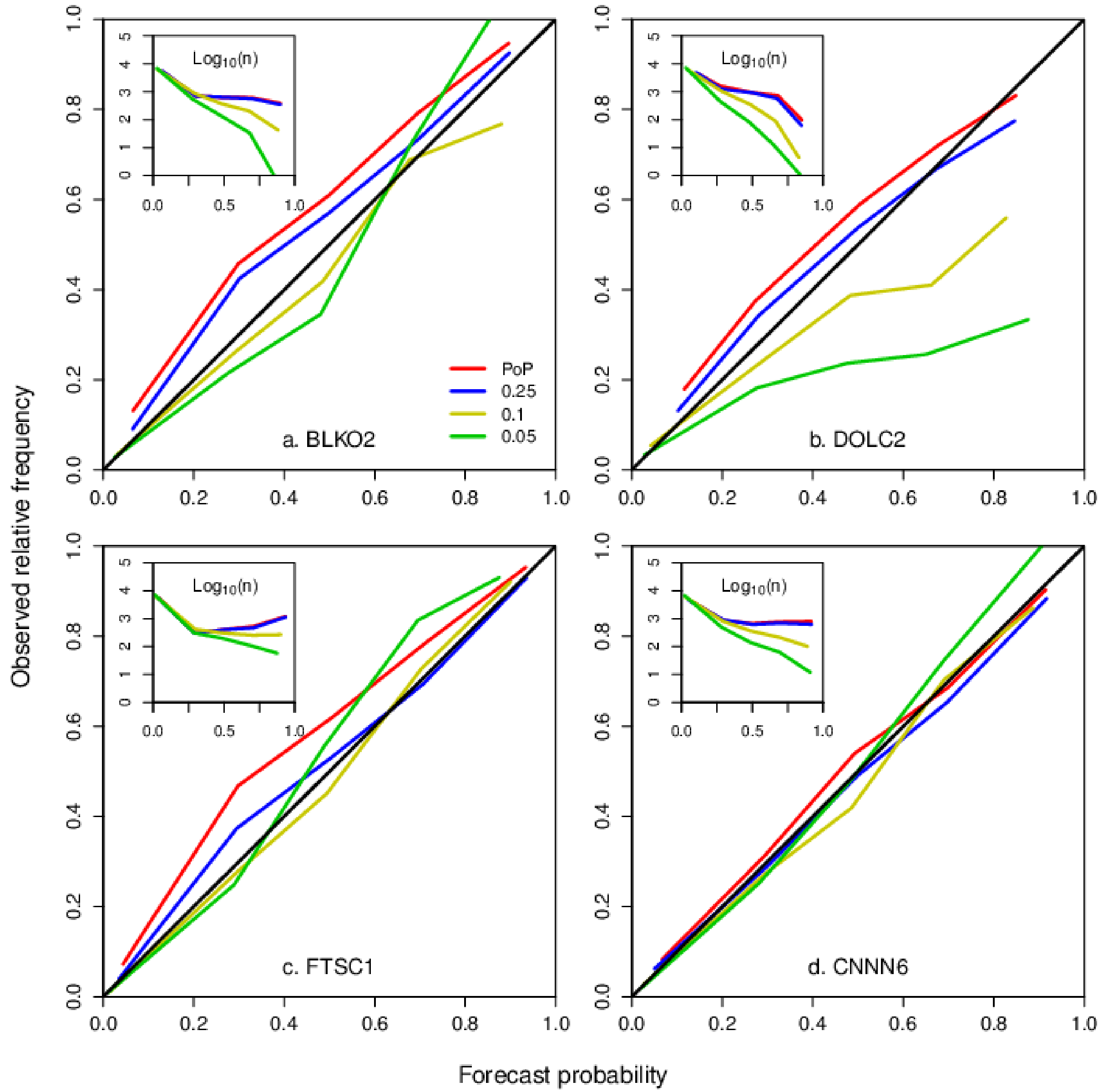


MEFP-GFS precipitation forecasts

HSL

Precipitation reliable

- Lead time of 1 day
- "0.05" = daily precip. exceeded 5% of time
- Moderate and high precipitation amounts generally show reliable probabilities
- Tendency for "dry bias" in PoP, i.e. forecast prob. too low
- Sample size becomes an issue in upper tail, so good to look at "raw data" plots…

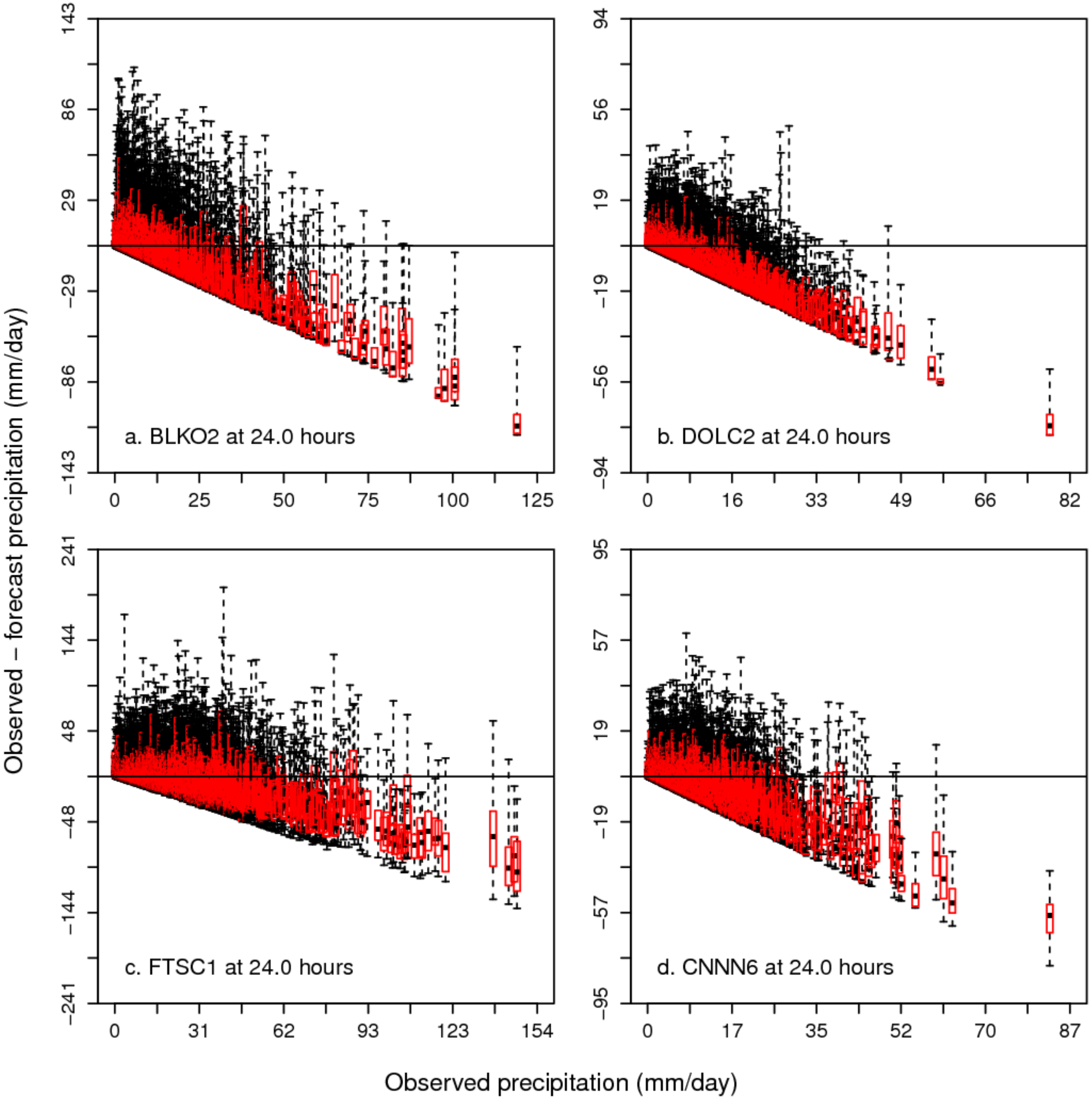


MEFP-GFS precipitation forecasts

HSL

Conditional bias

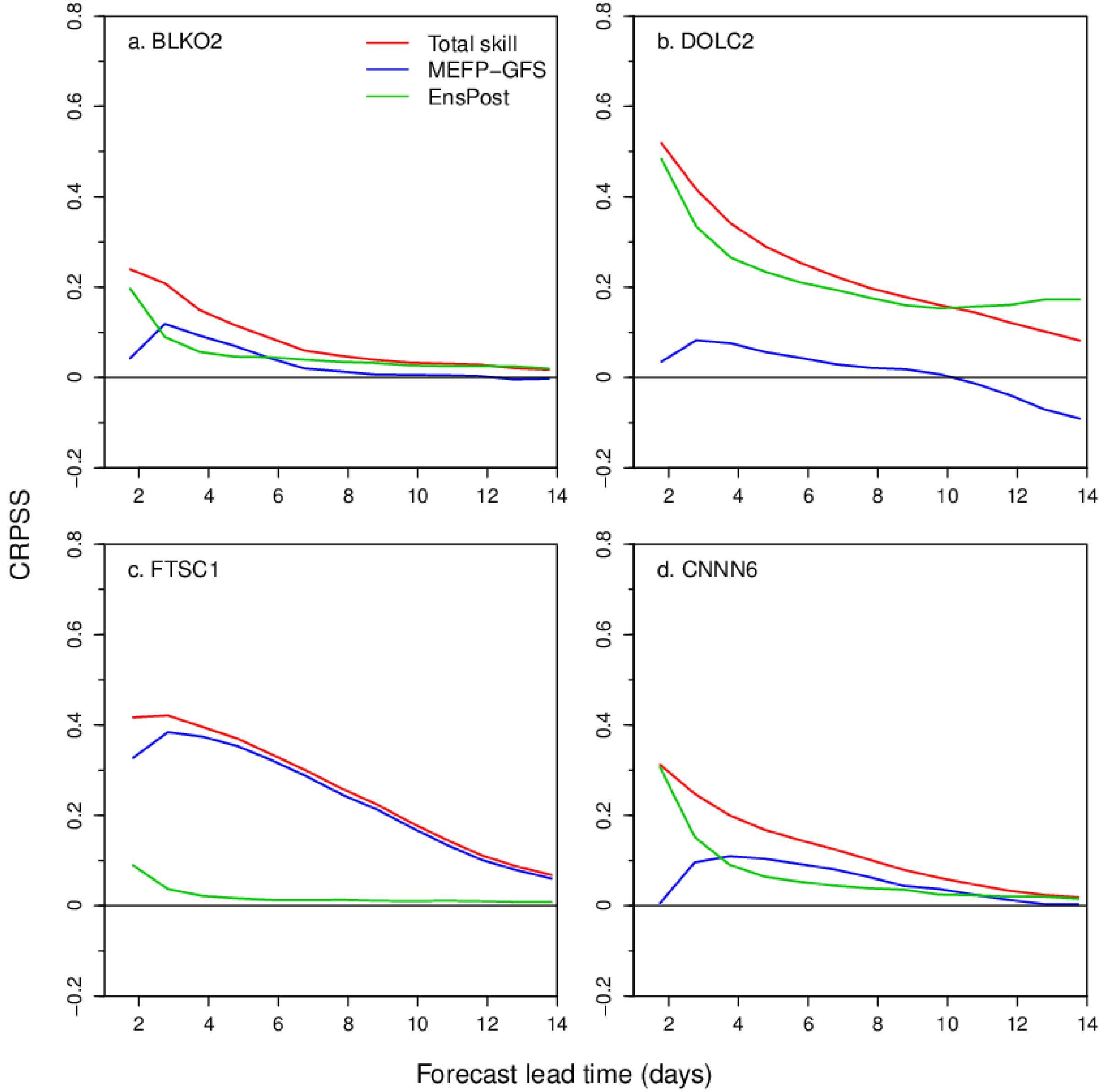
- Box plots ordered by observed amount at lead time of 1 day
- Tendency to under-forecast largest observed precipitation amounts
- In FTSC1, forecasts generally "capture" even largest amounts
- Conditional bias increases with lead time (not shown)



MEFP-GFS streamflow forecasts

Skill and its origins

- Skill (CRPSS) with climate forcing as baseline (akin to ESP)
- Apportioned skill from MEFP-GFS and EnsPost
- Skill in CN mainly from MEFP-GFS
- Skill in CB mainly from EnsPost
- Skill in AB and MA from both sources
- <u>**Big**</u> seasonal variation though (not shown)



Phase I main findings

Overall, results as expected

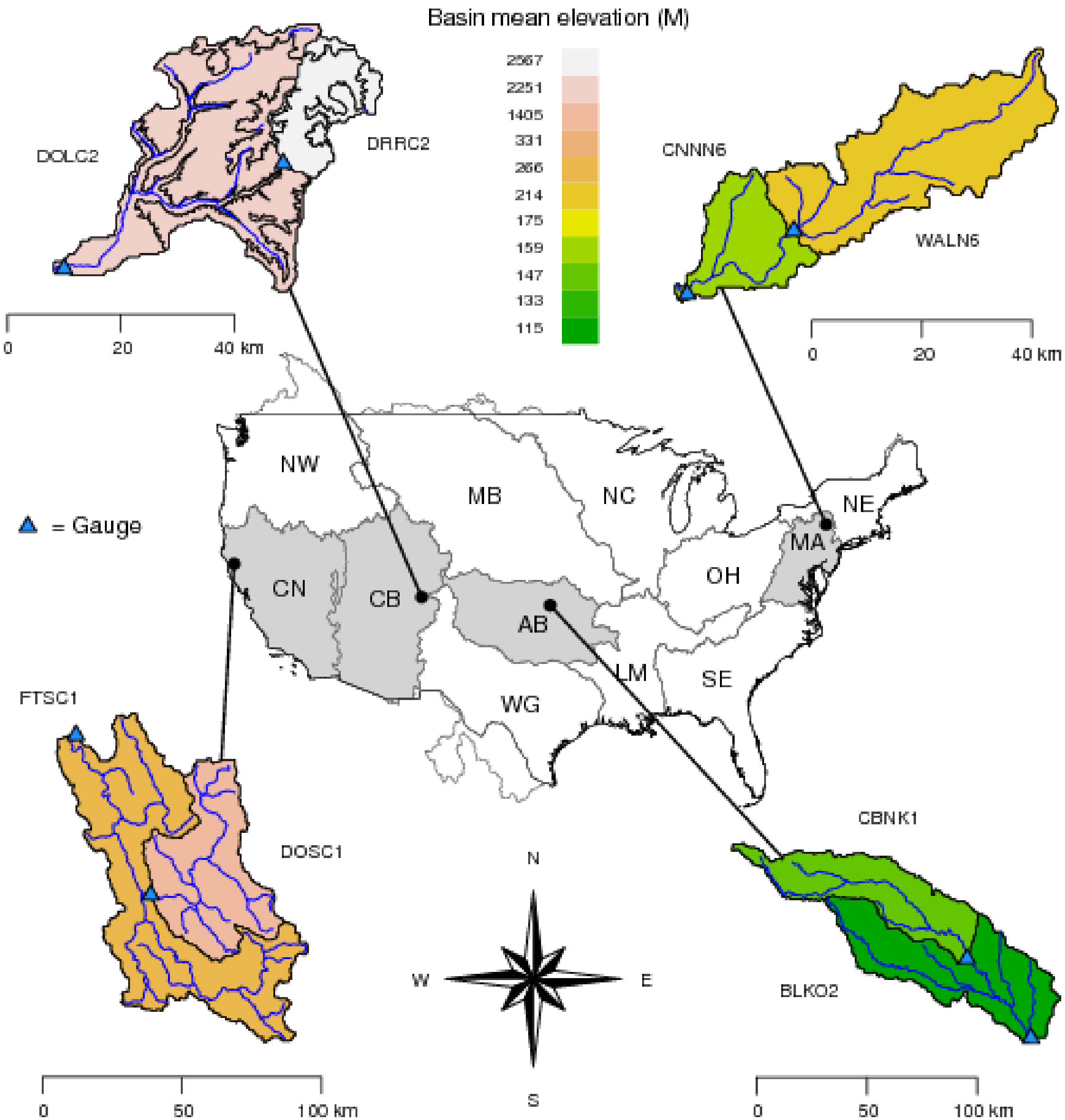
- MEFP preserves correlations of GFS, while reducing biases. Quality of GFS varies widely
- EnsPost adds skill by reducing bias (esp. low/ moderate flow). Difficulty of hydro. modeling varies
- Relative contributions from MEFP and EnsPost are highly conditional (on basin, season, flow etc.)
- Some issues to be explored
    - Conditional biases in PoP and heavy precipitation
    - Over-forecasting cold temperatures (GEFS is better)

3. Phase III: medium-range with latest GEFS, preliminary results

Phase III basins (same as Phase I)

Basins

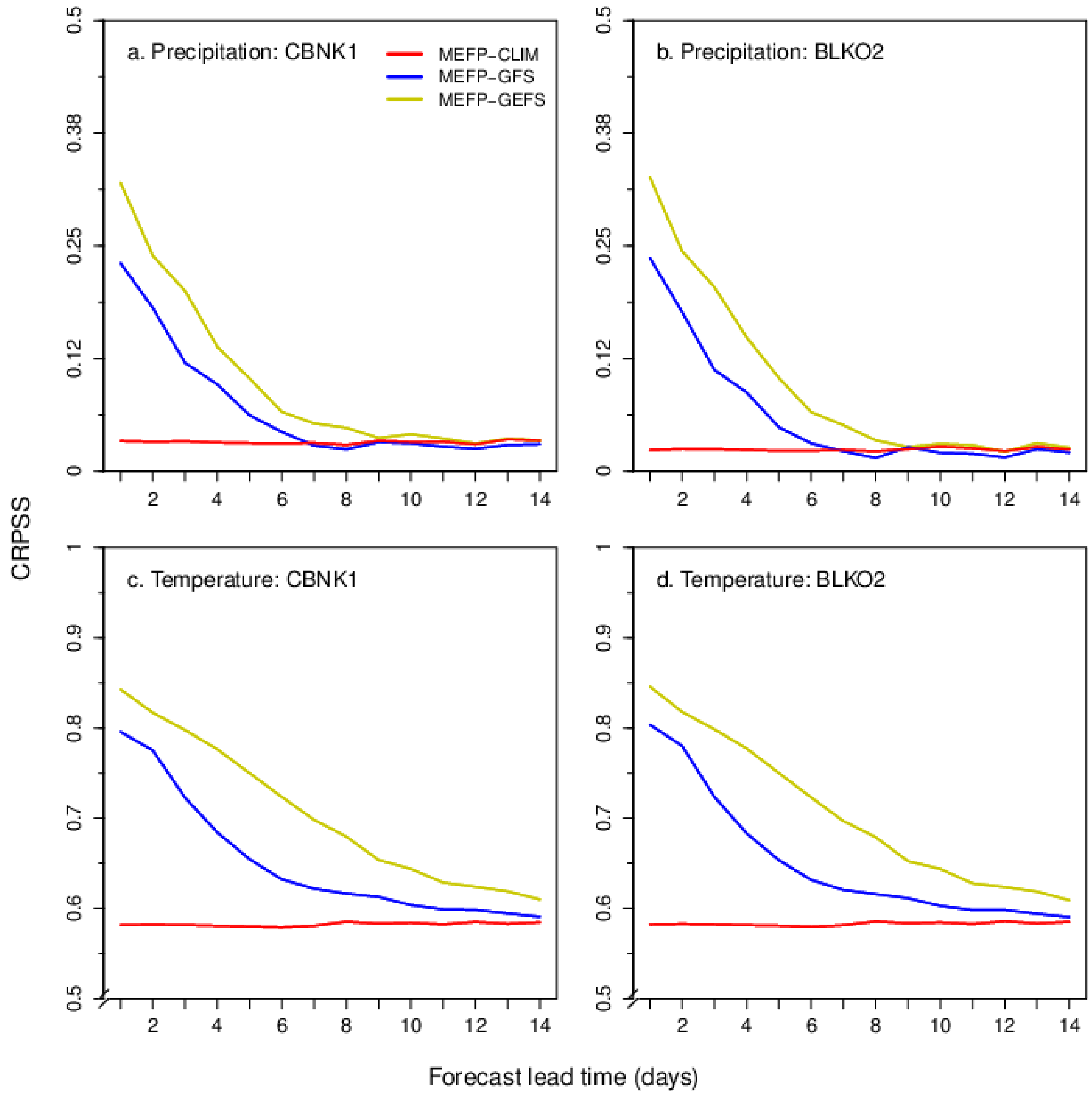
- Four RFCs
- Hindcasts: 1985-1999
- Upper/lower pairing
- USGS gauge at the outlet of each basin
- Relatively small basins (largest 2000 sq. miles)
- Low elevations in AB and MA
- Higher elevations in CB and CN
- CB and CN have MAT/MAP sub-basins



MEFP-GEFS: forcing

MEFP-GEFS adds value

- Preliminary verification results from MEFP-GEFS
- Skill (CRPSS) from two basins in ABRFC, precipitation (top) and temperature (bottom)
- Sample climatology as baseline (unconditional)
- Raw GEFS improves substantially on GFS and this is reflected in MEFP-GEFS results shown here
- Improvements particularly noticeable in first week, longer for temperature

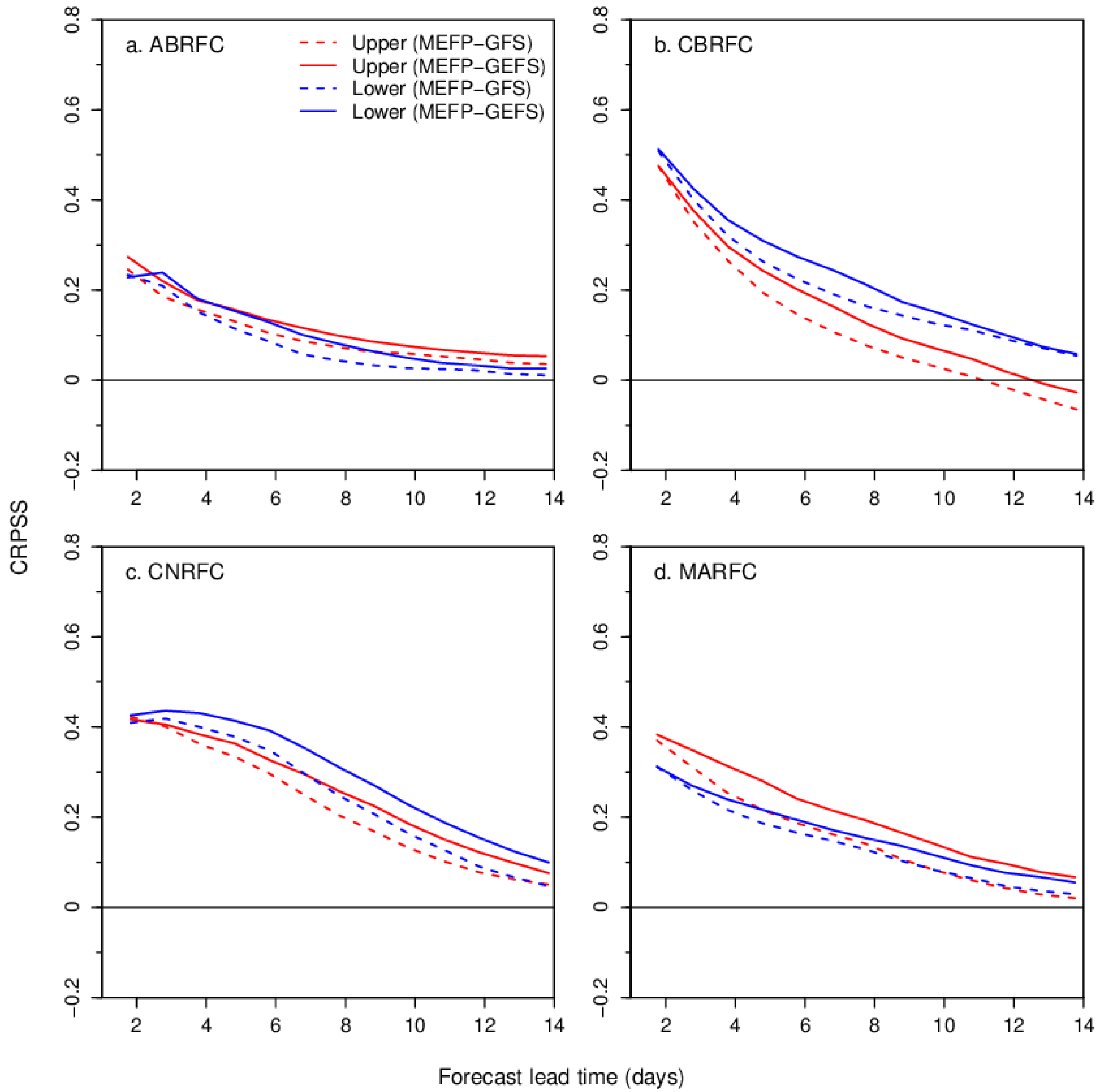


MEFP-GEFS: streamflow

HSL

Value also added to flow

- Streamflow with MEFP-CLIM baseline
- Skill shown for lower and upper basin
- Results include EnsPost
- GEFS consistently beats GFS (<u>statistically</u>)
- Skill from initial conditions and EnsPost dominates earliest times
- On time horizon of 4-10 days, GEFS adds ~1-2 days in lead time



Phase III preliminary findings

Forcing

- MEFP preserving correlations, reducing bias
- GEFS around 5-20% more skill than GFS in P (~1-7 days)
- As much as 50-75% more skill in T (~1-14 days)
- GEFS adds ~1-2 days lead time for P, and ~1-4 days for T

Streamflow

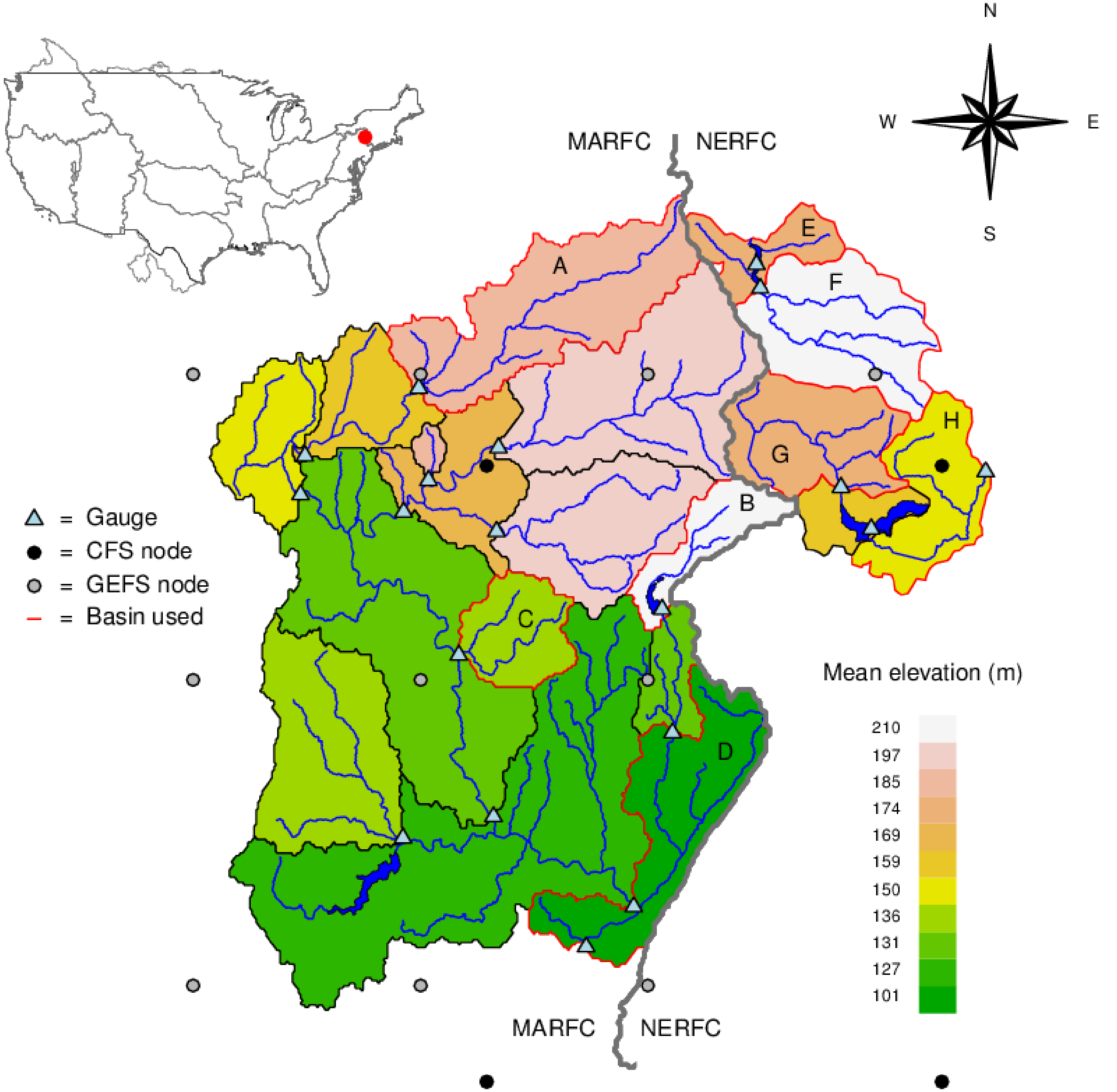
- Streamflow largely reflects P skill (T for snowmelt)
- Smaller added-value at early lead times (hydro. dominant)
- Once P washes through, GEFS adds ~1-2 days of skill

4. Phase II: long-range with GEFS+CFSv2+CLIM (GCC)

Phase II basins

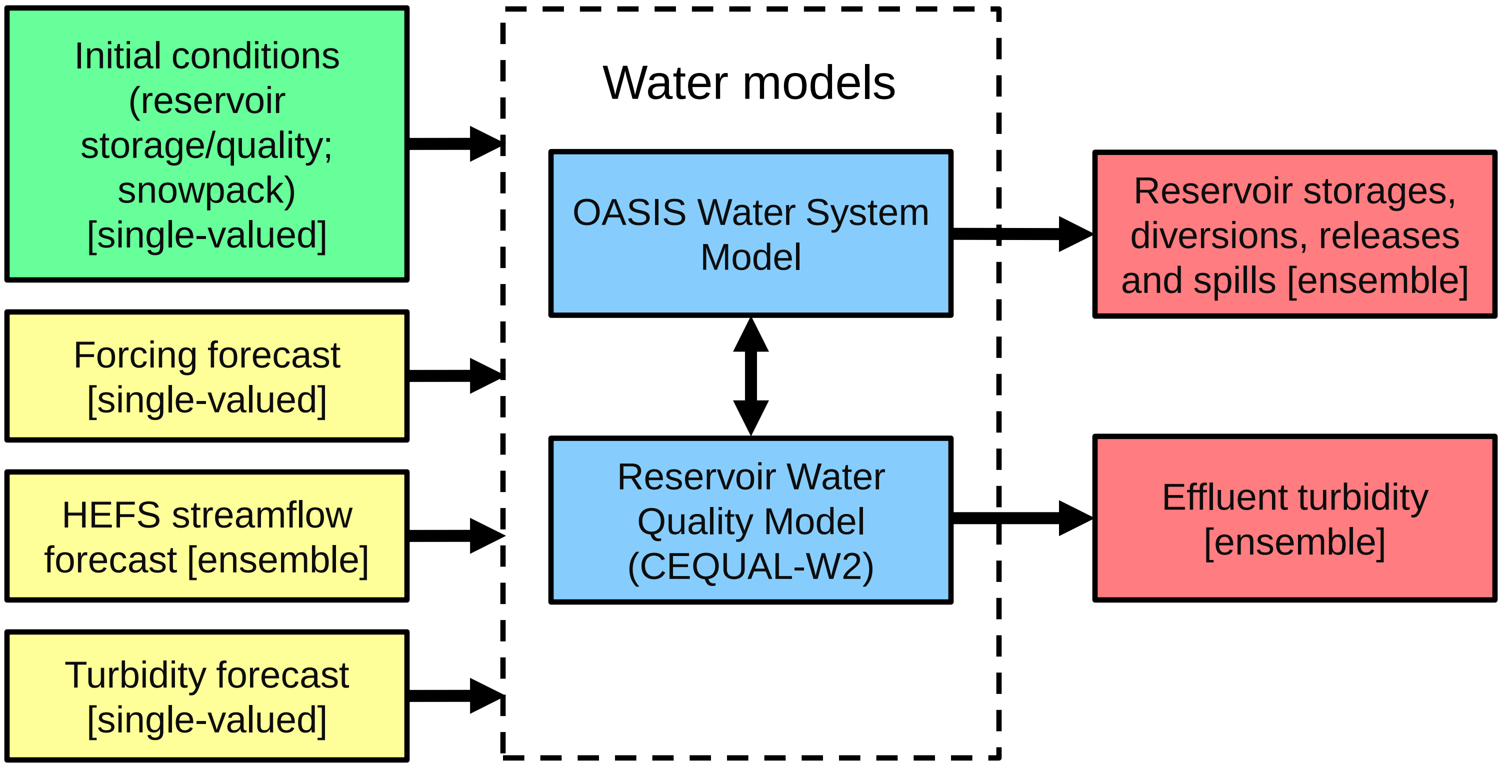
Basins

- MARFC and NERFC
- 22 basins
- Hindcasts: 1985-1999
- Verified 8 basins
  - MA-WALN6 (A)
  - MA-CCRN6 (C)
  - MA-MTGN4 (D)
  - MA-NVXN6 (B)
  - NE-MTRN6 (G)
  - NE-MRNN6 (H)
  - NE-PTVN6 (F)
  - NE-GILN6 (E)
- Most are subject to regulations (NYCDEP)



Motivation: NYCDEP

- HEFS inputs to NYCDEP Operational Support Tool (OST)

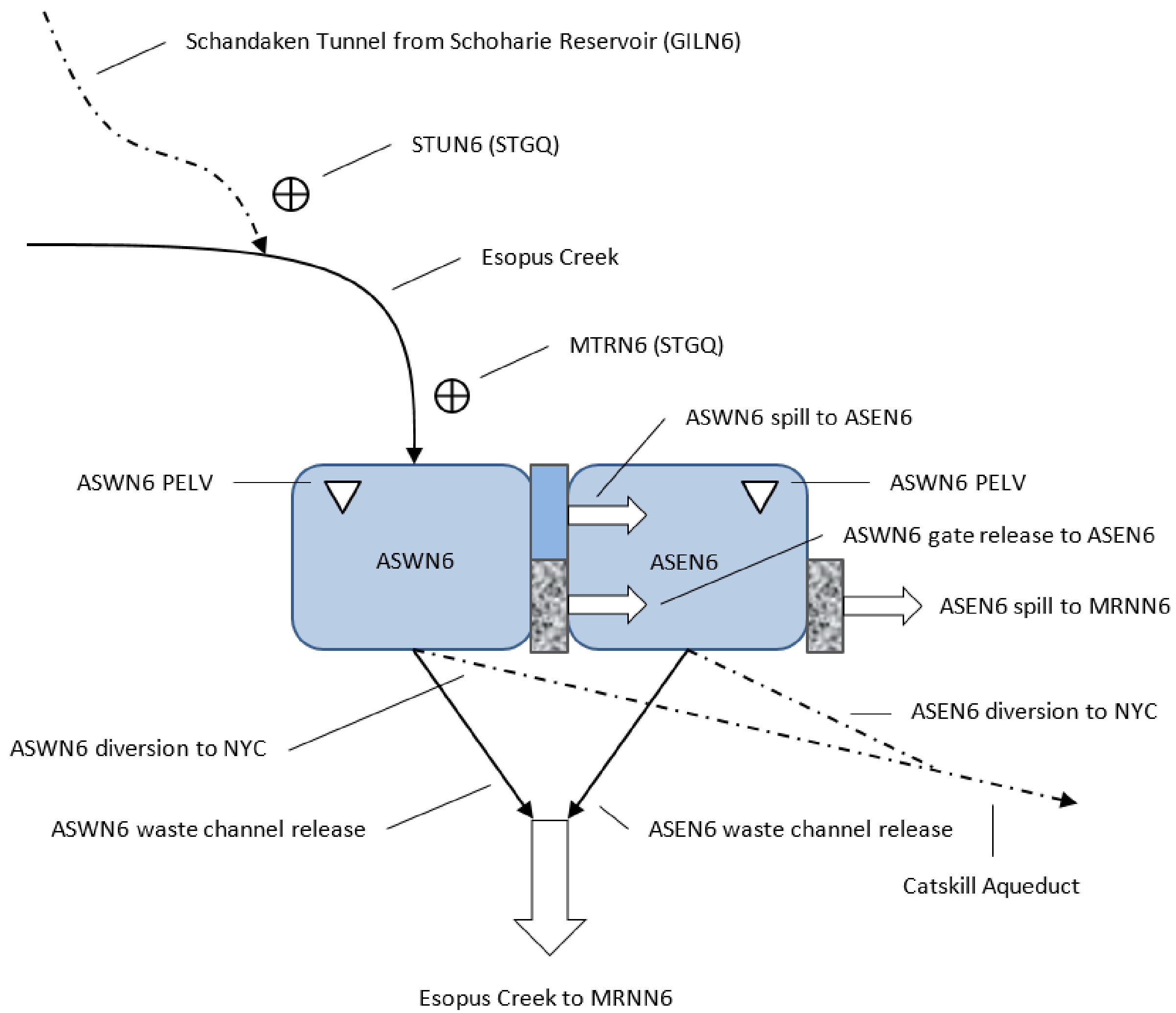


- Output: risks to volume objectives (e.g. habitat, flooding)
- Output: risks to quality objectives (NYC water supply)

Handling river regulations

Local flows verified

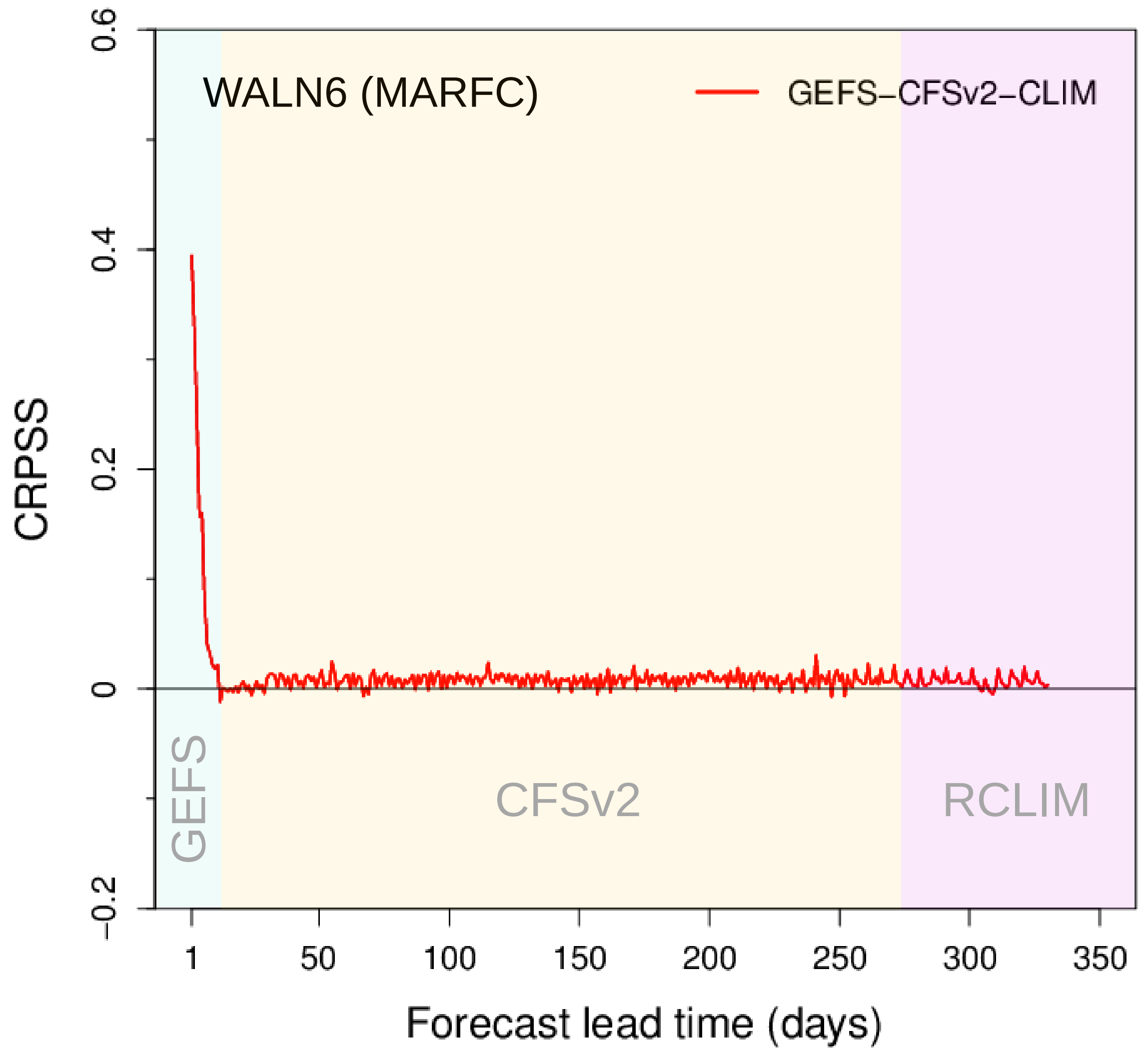
- Regulations often complex: see plot of Ashoken (NERFC)
- Adjust for diversions and releases in real-time
- In general, better to calibrate EnsPost on (estimated) natural flows
- Possible if regulations are known historically and in real-time
- Estimated local flows provided by NYCDEP
- EnsPost results not yet available



MEFP-GCC precipitation forecasts

Good skill in first week

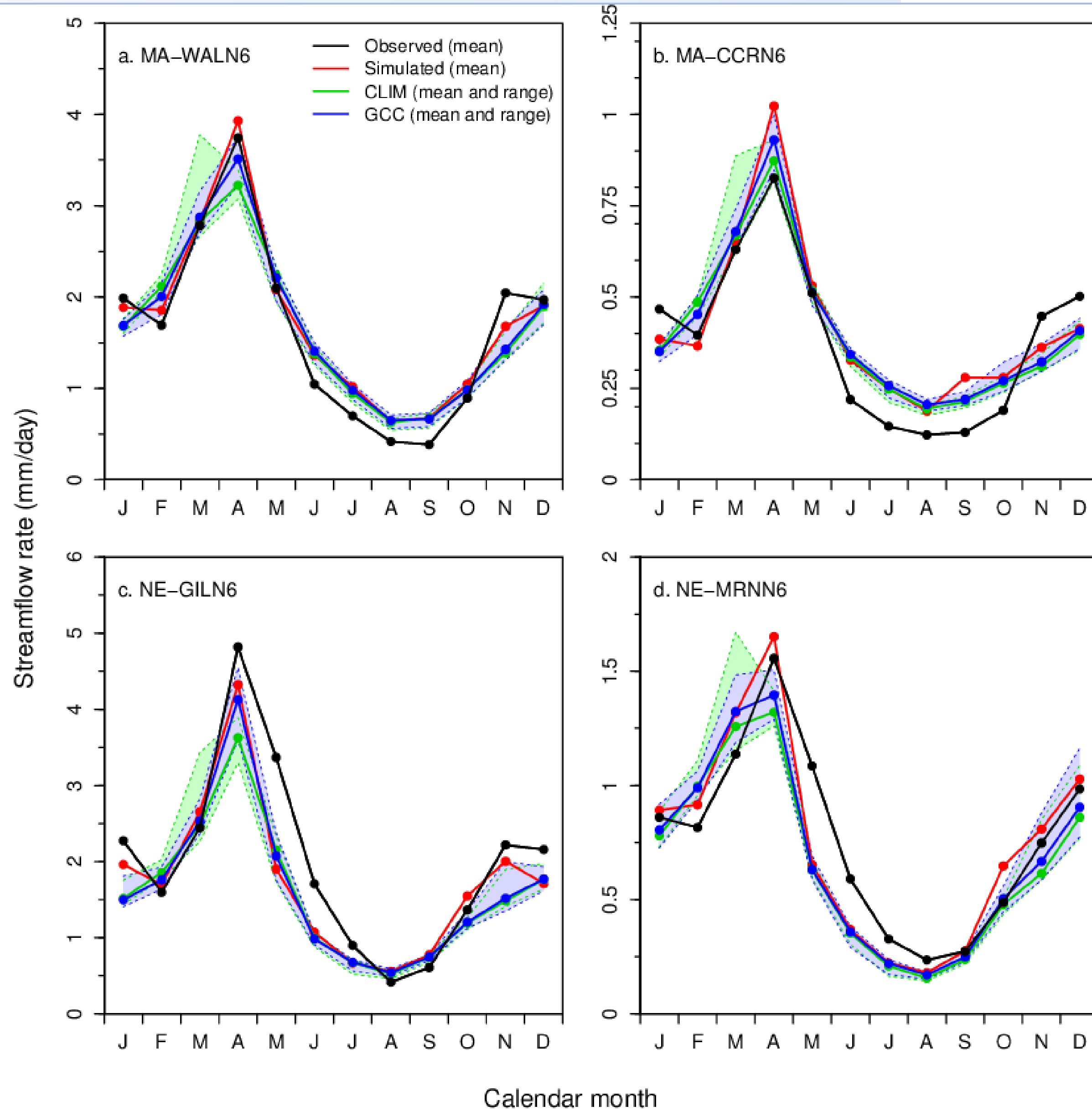
- Example of precipitation skill for WALN6
- Little skill in MEFP precipitation forecasts beyond ~one week (GEFS)…
- …to be expected as raw CFSv2 has limited skill, **except for specific regions and times of the year**
- Similar patterns seen for larger accumulation volumes (e.g. weekly, monthly)



Streamflow climatology

Some seasonal biases

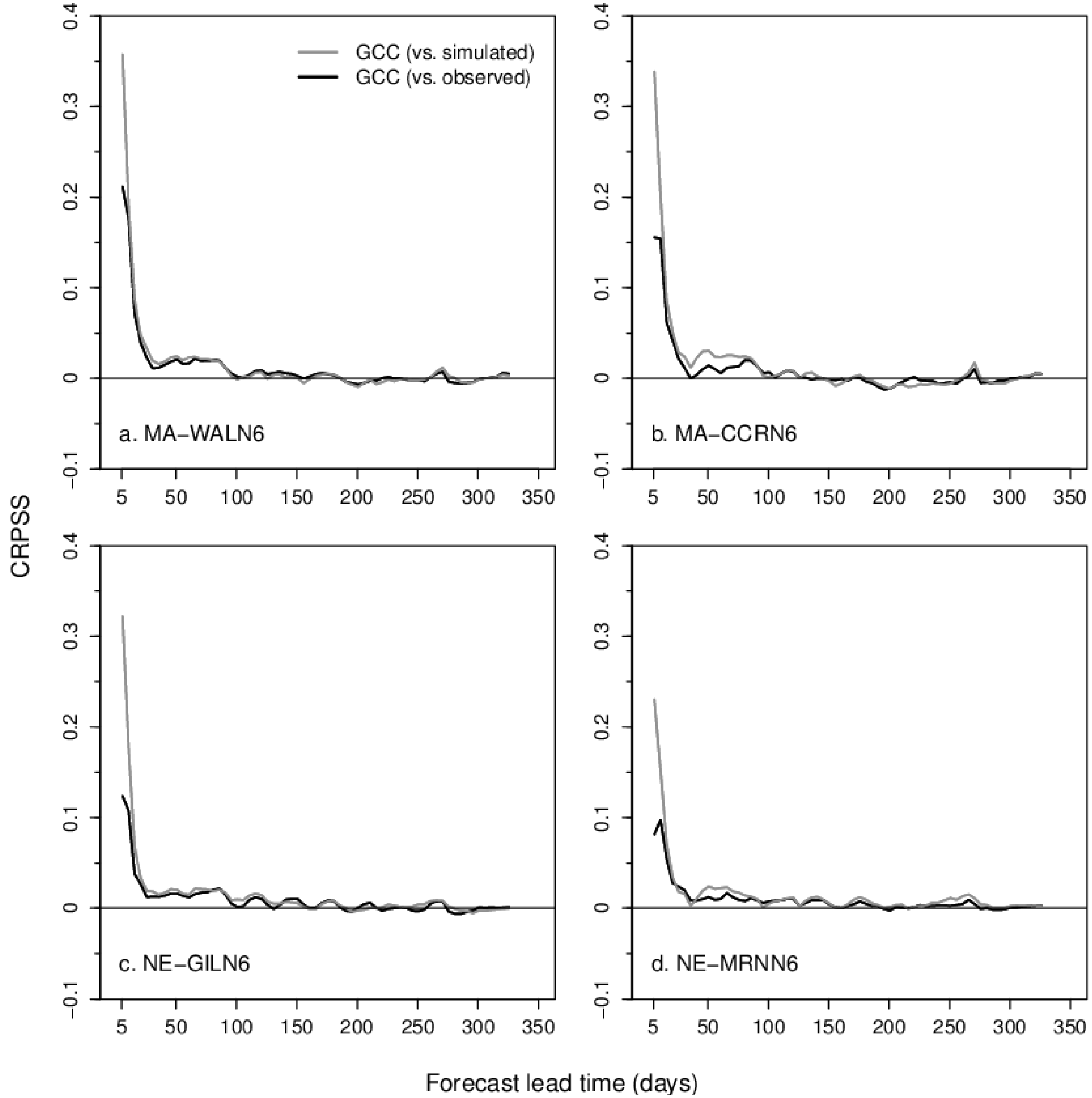
- Selected basins from MARFC and NERFC
- Mean forecast, observed and simulation by month
- Spread gives the range across forecast lead times (sample noise)
- Some systematic biases
- Low flow poorly captured in some basins, but spring peak reasonable
- Biases in MRNN6/GILN6
- Could remove these biases with EnsPost



MEFP-GCC: streamflow

Overall skill

- Total skill in streamflow when forced by MEFP-GCC versus MEFP-CLIM
- Verification against simulations: indicates skill without hydro. biases
- Overall, skill limited to period of GEFS forcing. But GEFS skill may last longer than 1-2 weeks
- EnsPost should add meaningful skill at early lead times
- Lack of forcing skill takes over at long lead times



Phase II main findings

Long-range precipitation problematic

- Very limited skill beyond ~1 week (GEFS)
- Similar story at aggregated periods (e.g. monthly)
- But: MEFP-GCC no worse than MEFP-CLIM (this is good)

Streamflow consistent with forcing

- Good skill for first 1-2 weeks (EnsPost will add further)
- No appreciable skill in long-range as GEFS washes out
- But: limited basins in north-east (CFSv2 poor)
- But: EnsPost will add skill when calibration is poor

5. Issues, gaps and recommendations (focused on science validation)

Main issues (from validation)

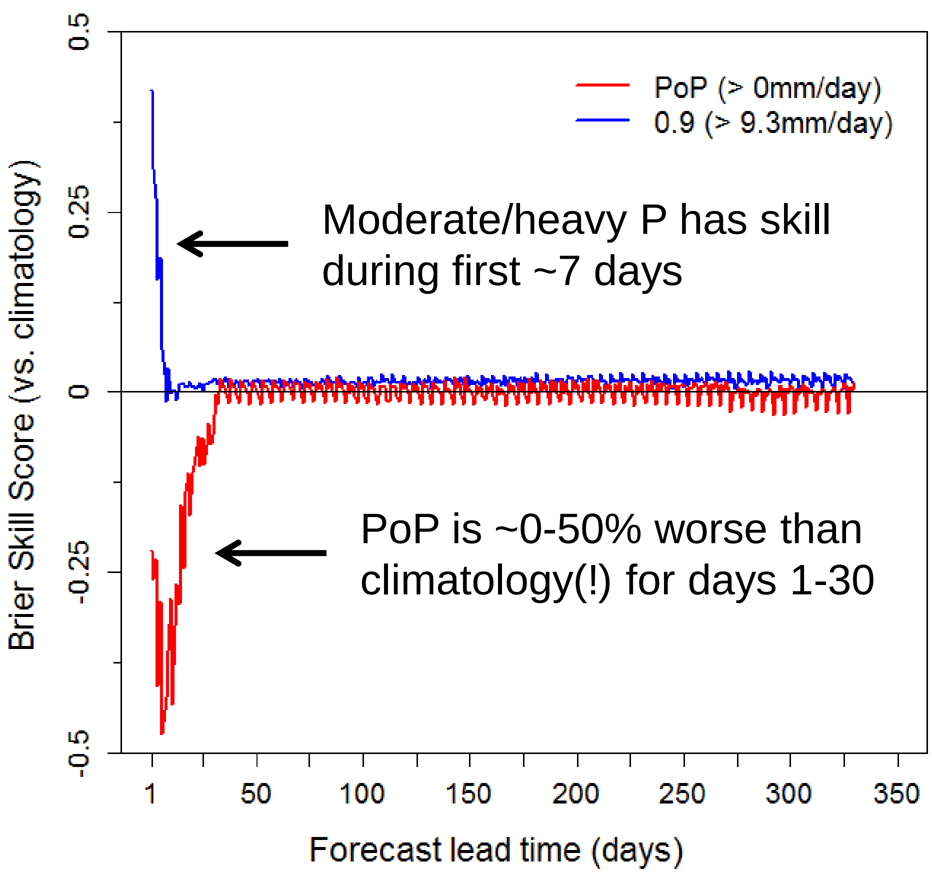
1. Biased forecasts of PoP

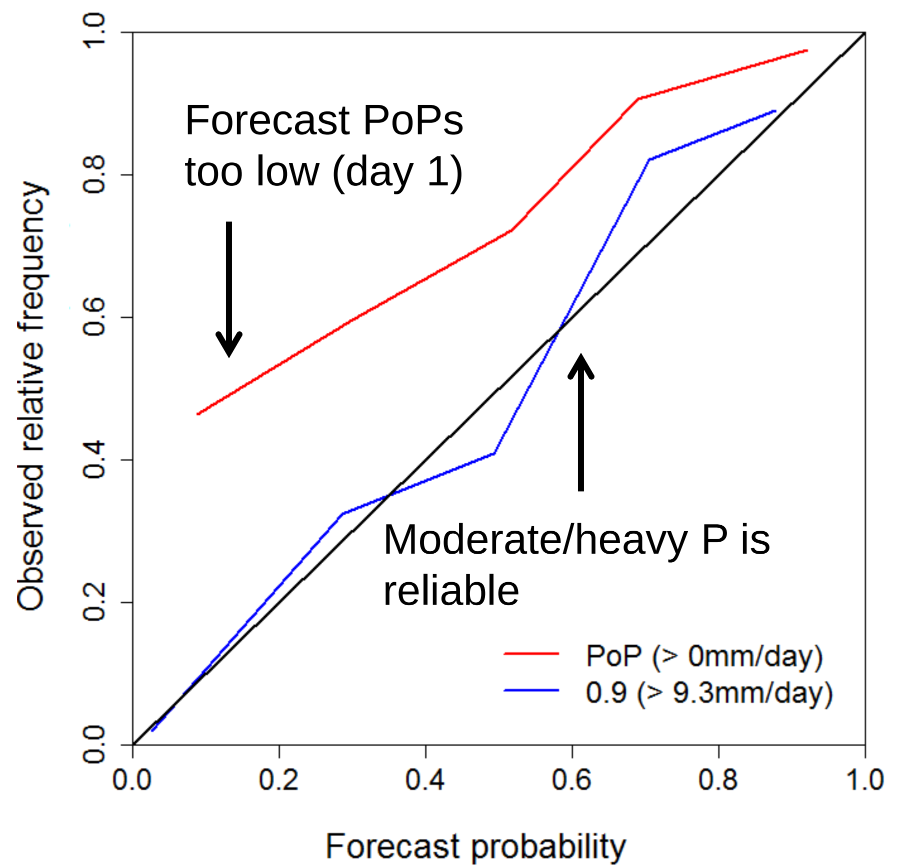
- Too many zero/light precipitation members (PoP too low)
- Particularly during first 30 days of long-range forecasts
- May be partly related to choice of threshold for PoP
- Was not seen in early versions (hopefully, simple fix)

2. “Discontinuities” in forecast horizon

- Abrupt features in verification results for P and T
- Live forecasts for T reveal shifts on monthly multiples
- Confident this is related to “canonical events”

WALN6, 1-330 days, GEFS+CFSv2+CLIM





Discontinuities

Possible causes of abrupt changes

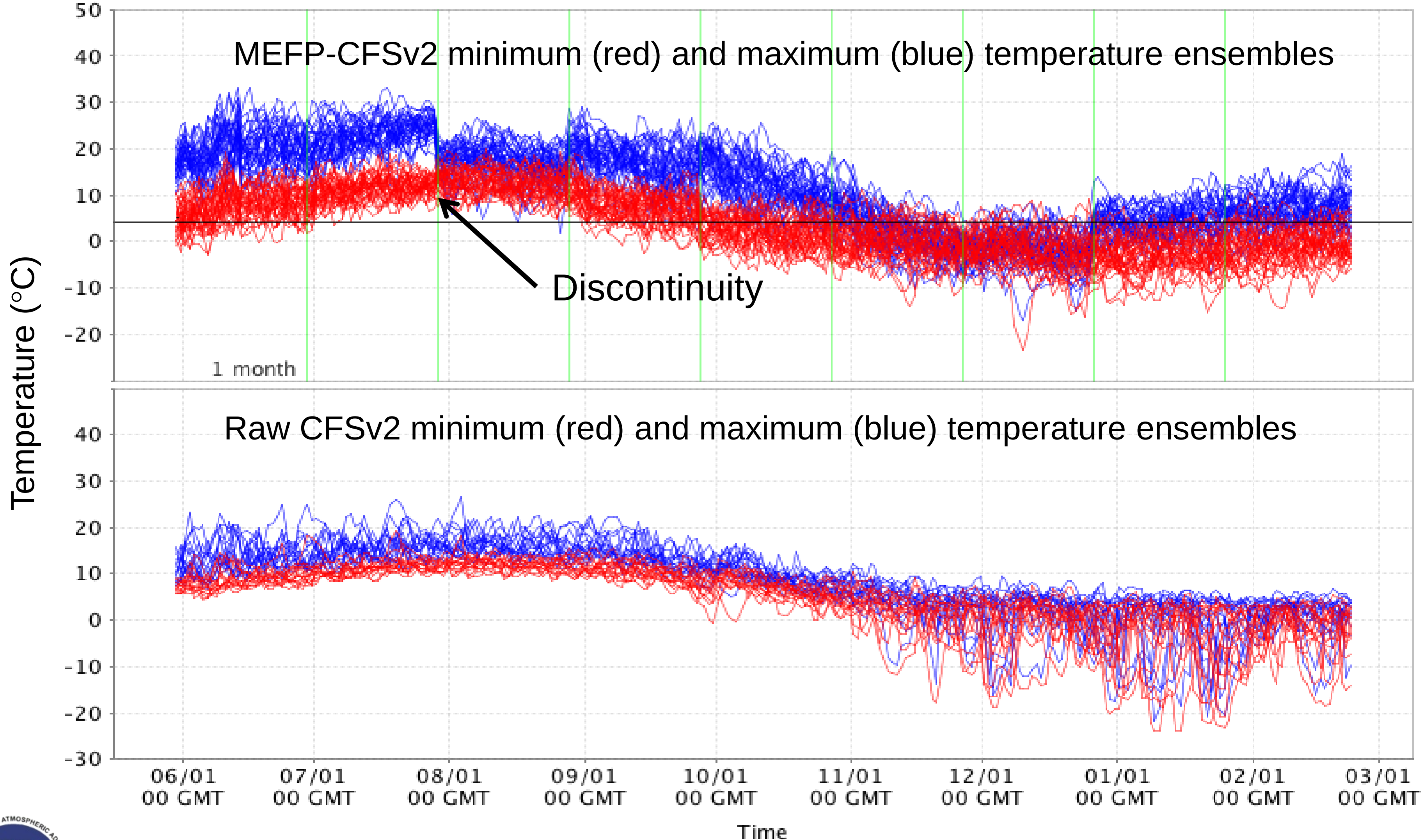
- Sudden changes in weather (real)
- Transition between raw forcing sources (artificial)
- Canonical events (artificial)

Canonical events

- Designed to capture skill in raw forcing at multiple scales
- Sequentially adjust climatology per event → final forecast
- Events operate on different parts of forecast horizon
- Limited sample data, so transitions may be abrupt

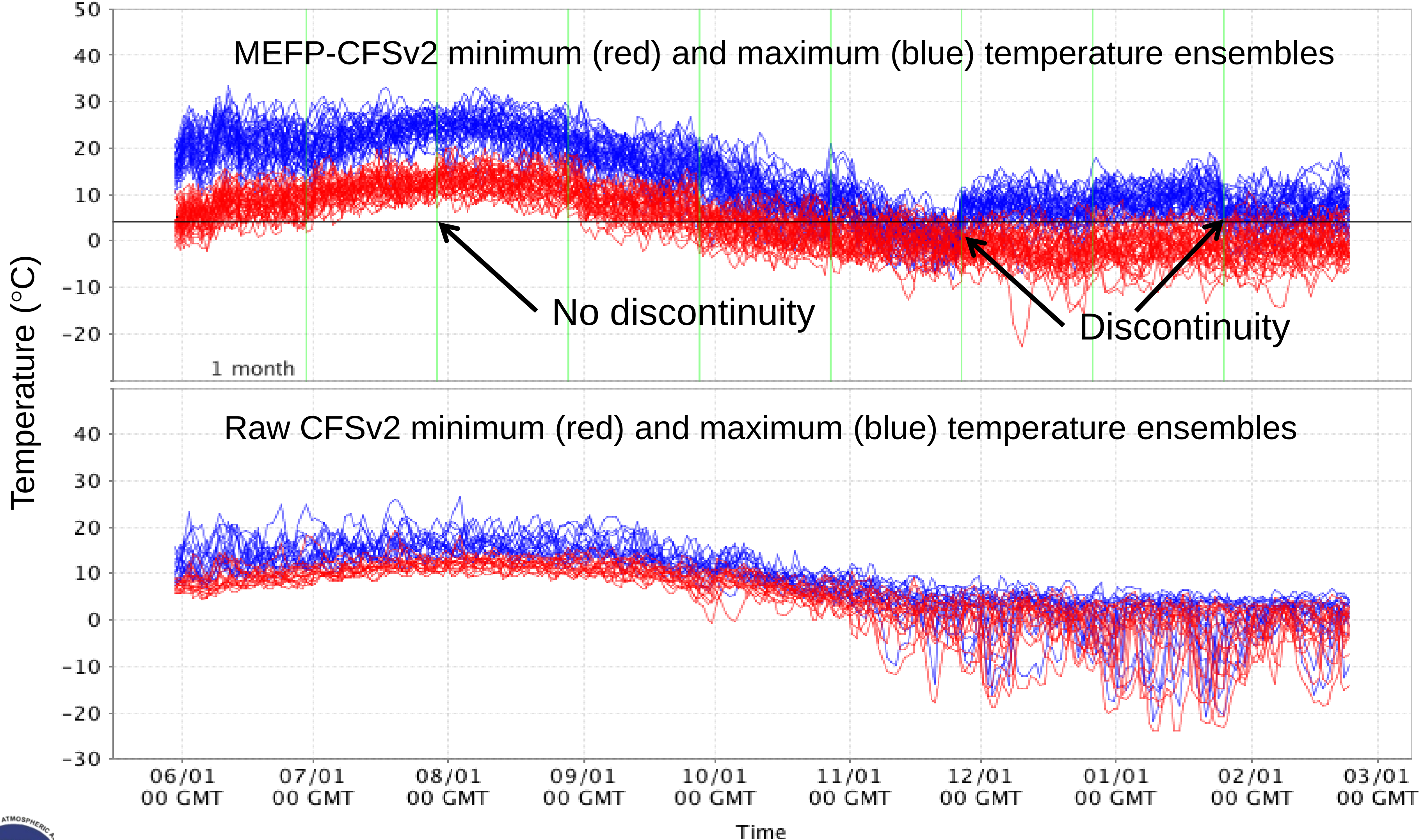
Canonical events: example (CREC1)

MEFP-CFSv2 temperature (daily min/max) using all canonical events



Canonical events: example (CREC1)

MEFP-CFSv2 temperature (daily min/max) WITHOUT "modulation events"



Gap analysis underway

Science and software gaps

1. Science gaps
   - “Known gaps” (recent validation & the “v1” in HEFSv1)
   - “Unknown gaps” (need further science validation)
2. Software gaps (also known/unknown)

Examples of science gaps

- Benchmark HEFS to operational forecasts
- Improve long-range forcing skill (climate indices?)
- Better accounting for hydro. uncertainties (e.g. DA)

Recommendations

Results so far broadly as expected

- Complex RTO project
- Prior testing was limited, mainly of individual components
- So, having no major surprises is a positive thing!

Proceed with planned rollout

- Some issues known (may be fixes in rollout timeframe)
- Can expect other issues with further evaluation
- Rollout will also raise issues (scientific and practical)
- CONOPS and training essential to smooth this transition

HSL

Further evaluation needed

- Limited basins and operating conditions so far
    - A few (mainly headwater) basins in four test RFCs
    - Limited testing of total flows downstream
    - Limited testing in regulated rivers (regulated ESP?)
    - Limited evaluation of high impact events
- No benchmarking against existing operations
    - ESP/statistical models for long range
    - RFC single-valued forecasts for short-range

Centralized infrastructure for evaluation

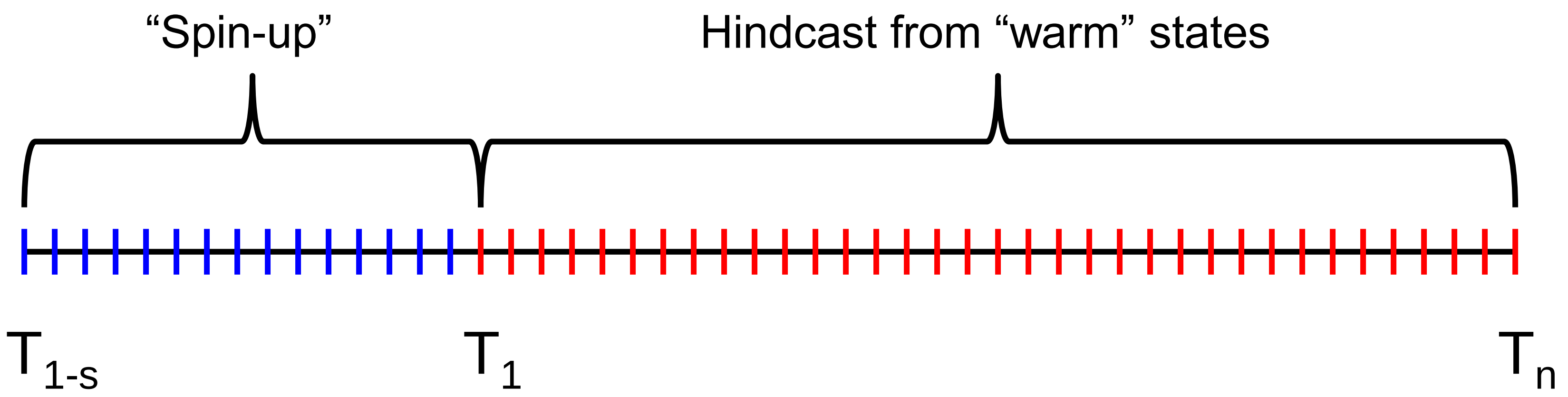
- Ad-hoc hindcasting is extremely difficult!
- NWC: opportunity to build low-latency archive & hindcasting/verification capability from ground-up
- Consistent, long-term, archive of observations, operational forecasts and hindcasts
- More work on diagnostics (reduce re-runs!)
- Testbeds to benchmark new techniques/data
- But evaluation must proceed in the interim

Questions?

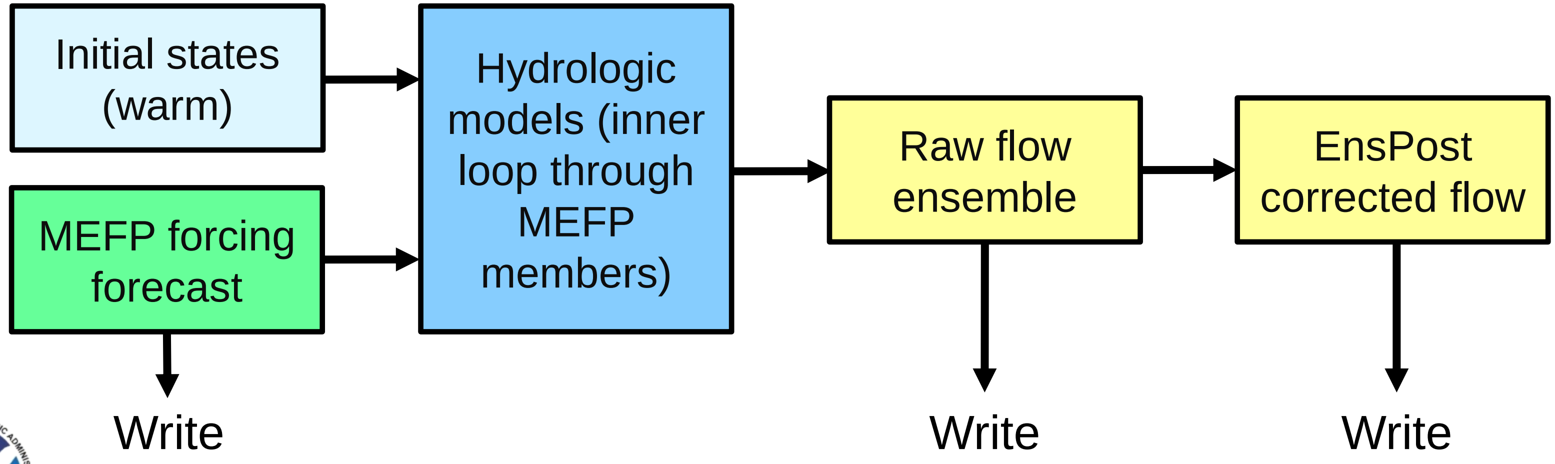
Extra slides

Hindcasting mechanics (CHPS)

Two step process: generate warm states, then hindcast

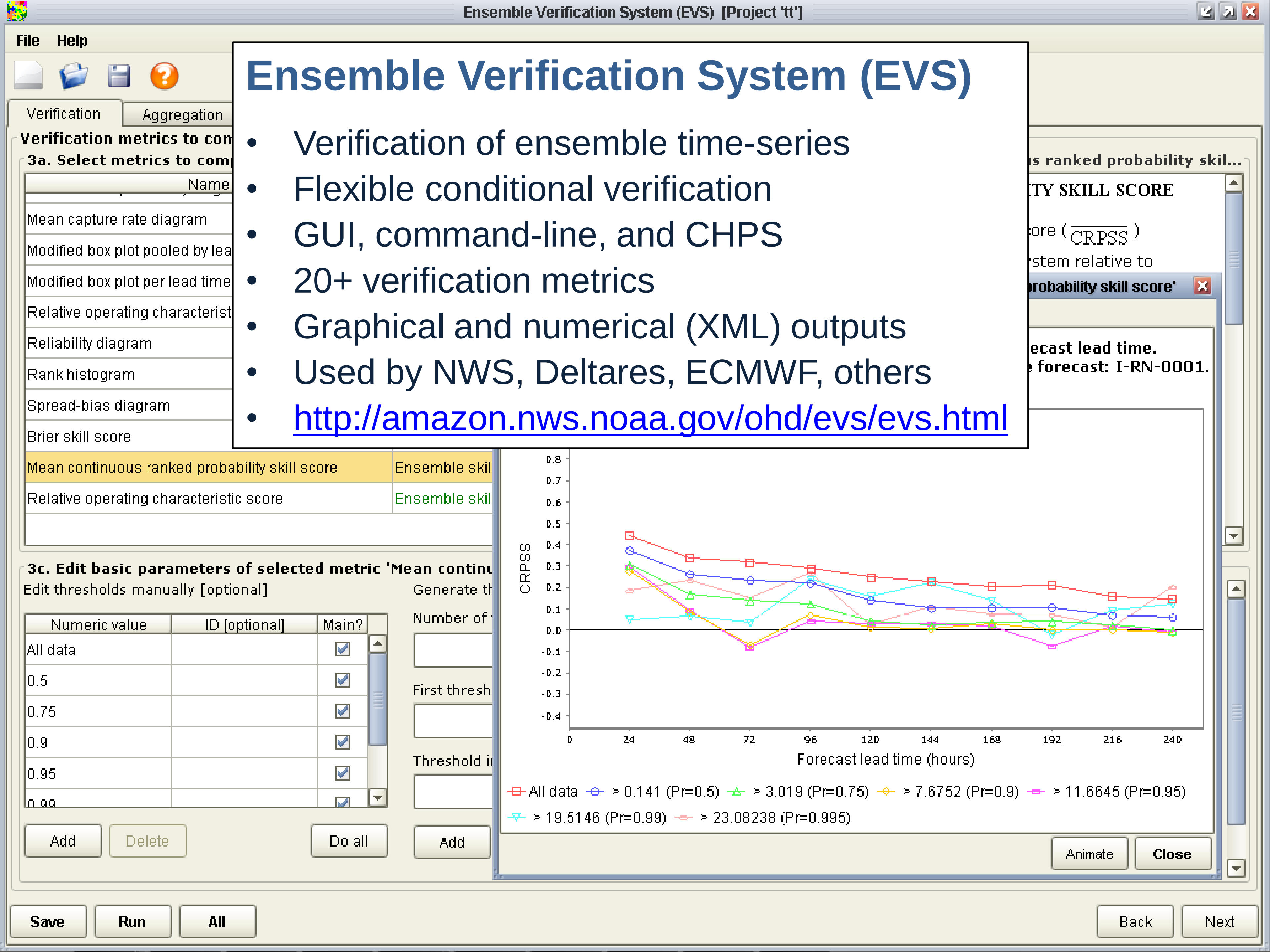


For each of $T_1,\ldots,T_n$:



Ensemble Verification System (EVS)

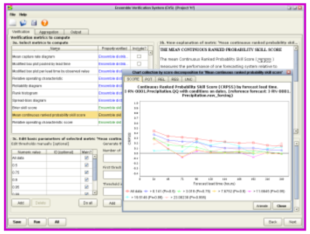
- Verification of ensemble time-series
- Flexible conditional verification
- GUI, command-line, and CHPS
- 20+ verification metrics
- Graphical and numerical (XML) outputs
- Used by NWS, Deltares, ECMWF, others
- http://amazon.nws.noaa.gov/ohd/evs/evs.html



How does the EVS operate?

1. Configure → 2. Execute → 3. Analyze

1. Configure



a. EVS GUI

EVS project file

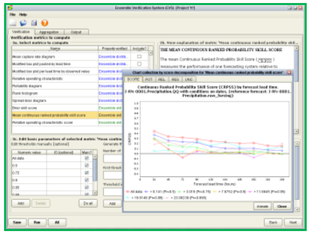
- Basins?
- Data?
- Metrics?
- ….

```
for fname in *.PIT
do
    for d in `seq 1 135`
    do
        DT=`printf "%04d" $d`
        newname=`echo $fname | sed "s/I-RN-0001/I-RN-$DT/"`
        echo 'Moving to file' $newname
        sed "s/I-RN-0001/I-RN-$DT/g" $fname > tmp.tmp
        mv tmp.tmp $newname
    done
done
```

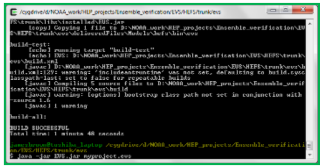
b. Shell scripting

Configure CHPS

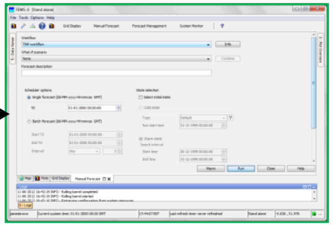
2. Execute



a. EVS GUI

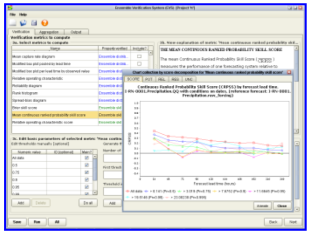


b. Command line

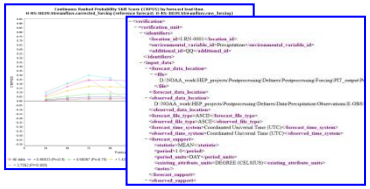


c. CHPS (hindcast)

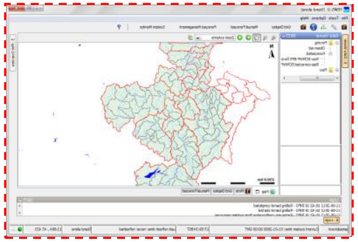
3. Analyze



a. EVS GUI



b. external tools



c. CHPS/GraphGen

Reminder of dates/deliverables

1. Delivery of NYCDEP hindcasts

- Delivered by 4th July 2013 (final by 4th September)
- Using MEFP “as-is” (mitigated issues as far as possible)

2. Delivery of science validation

- Phased-evaluation completed by 30th September
- Covers only a small fraction of locations and scenarios

3. Delivery of HEFSv1 software

- Version 1.01 on 24/09, maintenance release in mid-Nov.
- Rolled out to other RFCs in 2014

Reminder of version differences

Two versions of MEFP in active use

1. "Legacy" MEFP: EPP3 (Fortran) with updated hindcaster
- Hindcaster (CFSv2/GEFS): NYCDEP & science validation
- Forecaster: used by CNRFC **but** pre-HEFS version

2. "Recoded" MEFP: Java version with hindcaster/forecaster
- Used real-time, including NYCDEP & 150 basins in CN

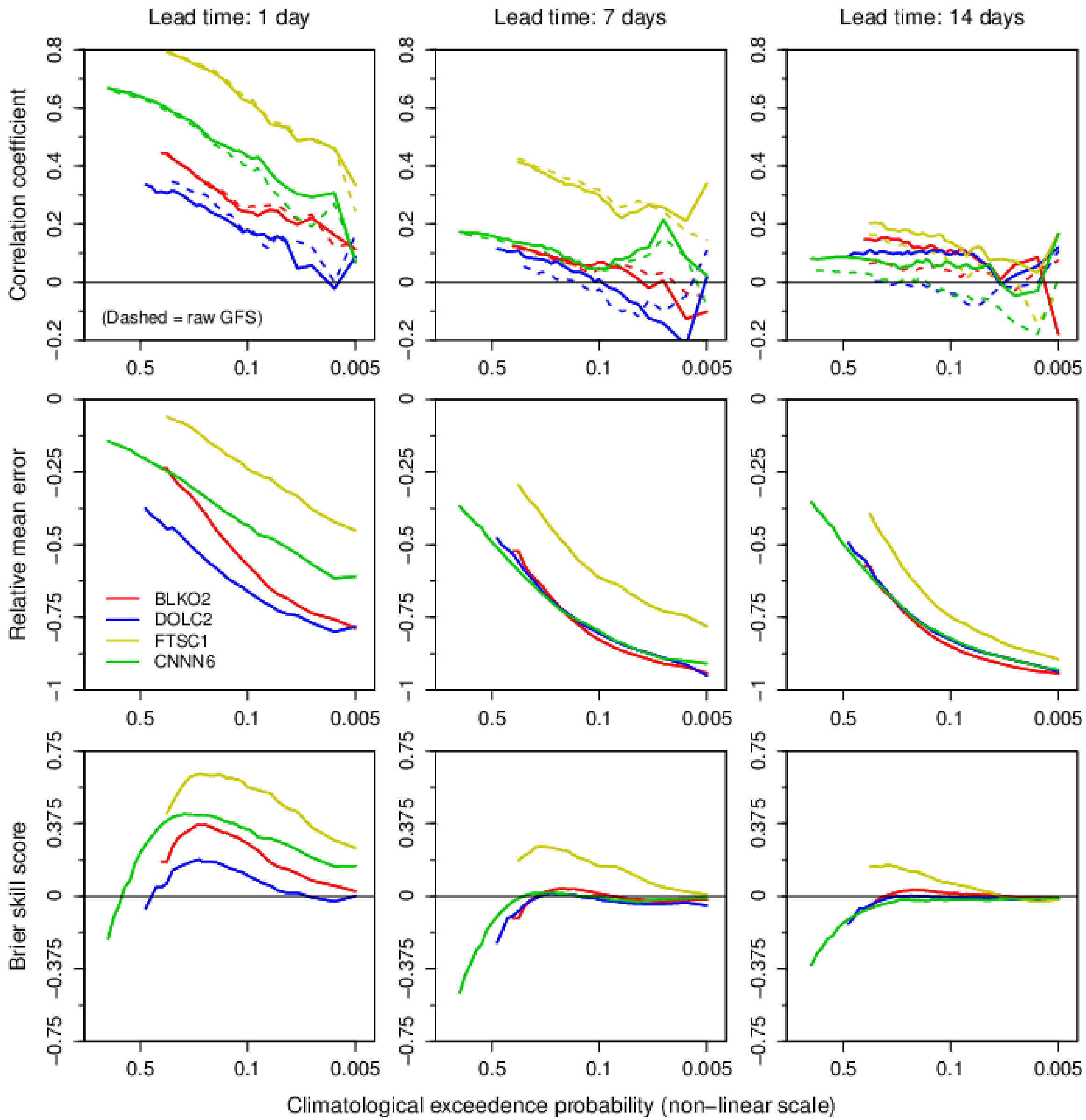
Equivalent (within some tolerance)

- Comparisons at OHD (software and hindcasts in 4 basins)
- Comparisons at CNRFC: some differences, can explain

Medium-range (GFS): example

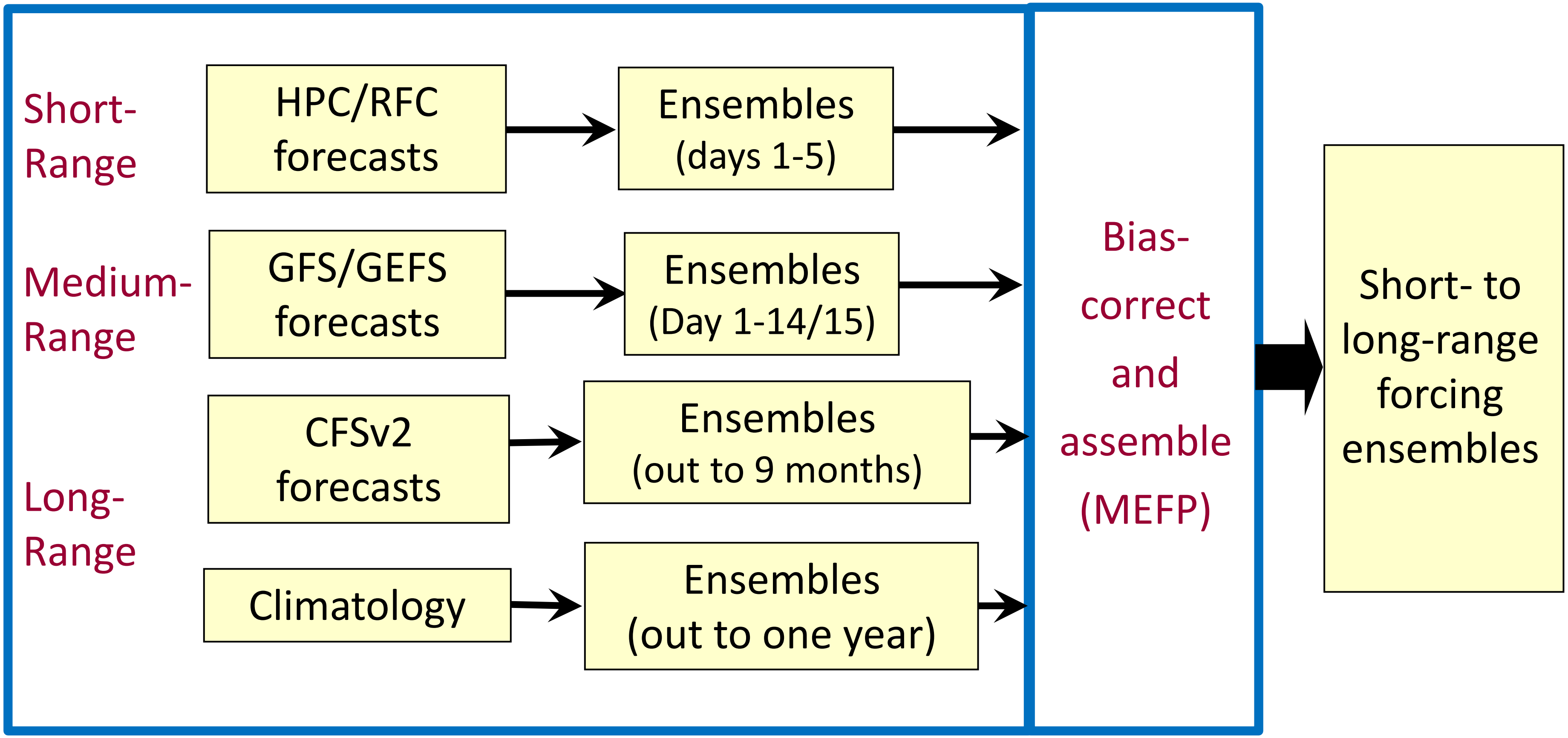
MEFP-GFS is skilful

- Subset of basins, 4 RFCs
- MEFP-GFS correlations similar to, or better than, raw GFS for all amounts across all forecast lead times (top)
- Biases increase for heavier precipitation (middle), but this is to be expected
- Some biases with PoP/light precipitation that reduce skill (bottom), which is not expected and points to an issue with the MEFP for PoP.



MEFP

MEFP data sources

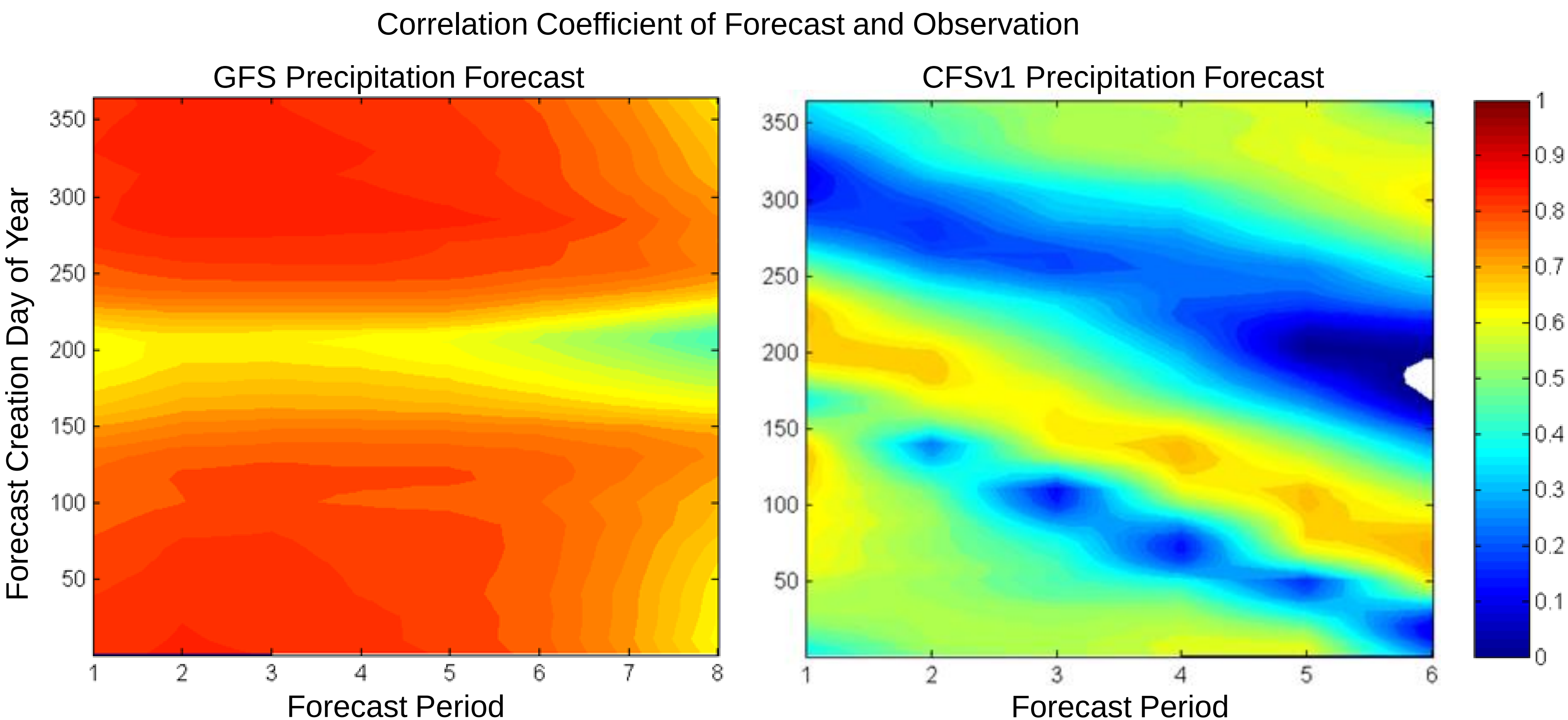


Methodology: key steps

- Partition data: forecast horizon broken into several units of variability or aggregation periods known as “canonical events” (see later) to extract maximum skill from raw forecasts
- Calibrate: for each forecast data source and canonical event, model the joint probability distribution between the single-valued forecasts and the corresponding observations
- Generate ensembles: given the live, single-valued, forecast, obtain the conditional probability distribution of the observed variable (take a “slice” through the joint distribution), then sample members
- Recover space-time and cross-variable relationships by applying the Schaake Shuffle
- Assemble the forecasts from the different sources

Canonical events: GFS and CFSv1

Correlation of GFS and CFSv1 precipitation forecast for NFDC1 in CNRFC

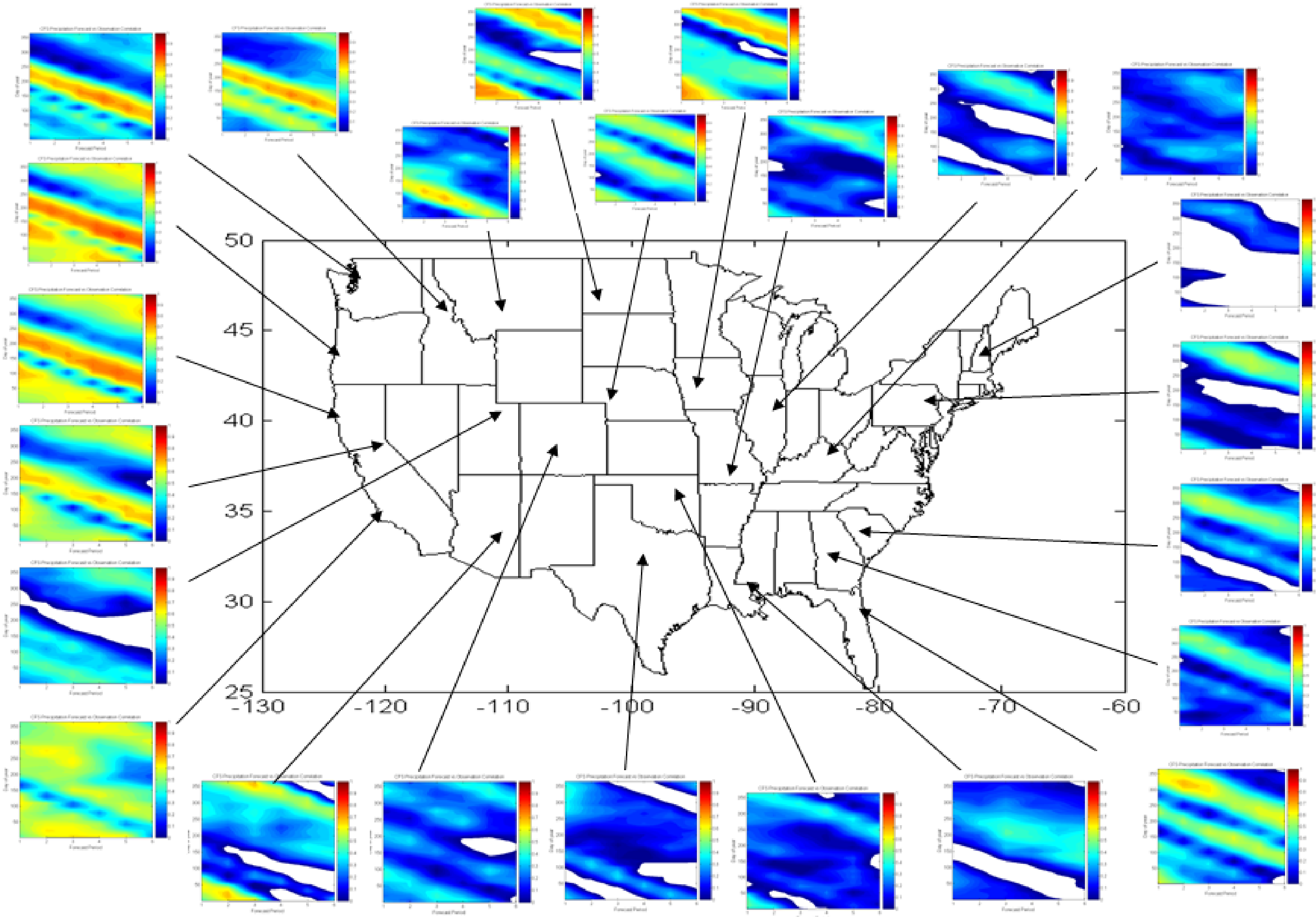


Period	1	2	3	4	5	6	7	8
Days	1	1-2	1-3	1-4	1-5	1-7	1-10	1-14

Period	1	2	3	4	5	6
Months	1-3	2-4	3-5	4-6	5-7	6-8

Canonical events: CFSv1

Correlation of forecast and observation for CFSv1 precipitation forecasts



Potential skill of CFSv1 precipitation forecast for 24 basins

Meta-Gaussian model

- Consider the joint distribution of forecast and observation:

  $F(x,y) = P(X \leq x, Y \leq y)$ *X*: Forecast *Y:* Observation

- The meta-Gaussian distribution constructed from the forecast and observation (Kelly and Krzysztofowicz, 1997):

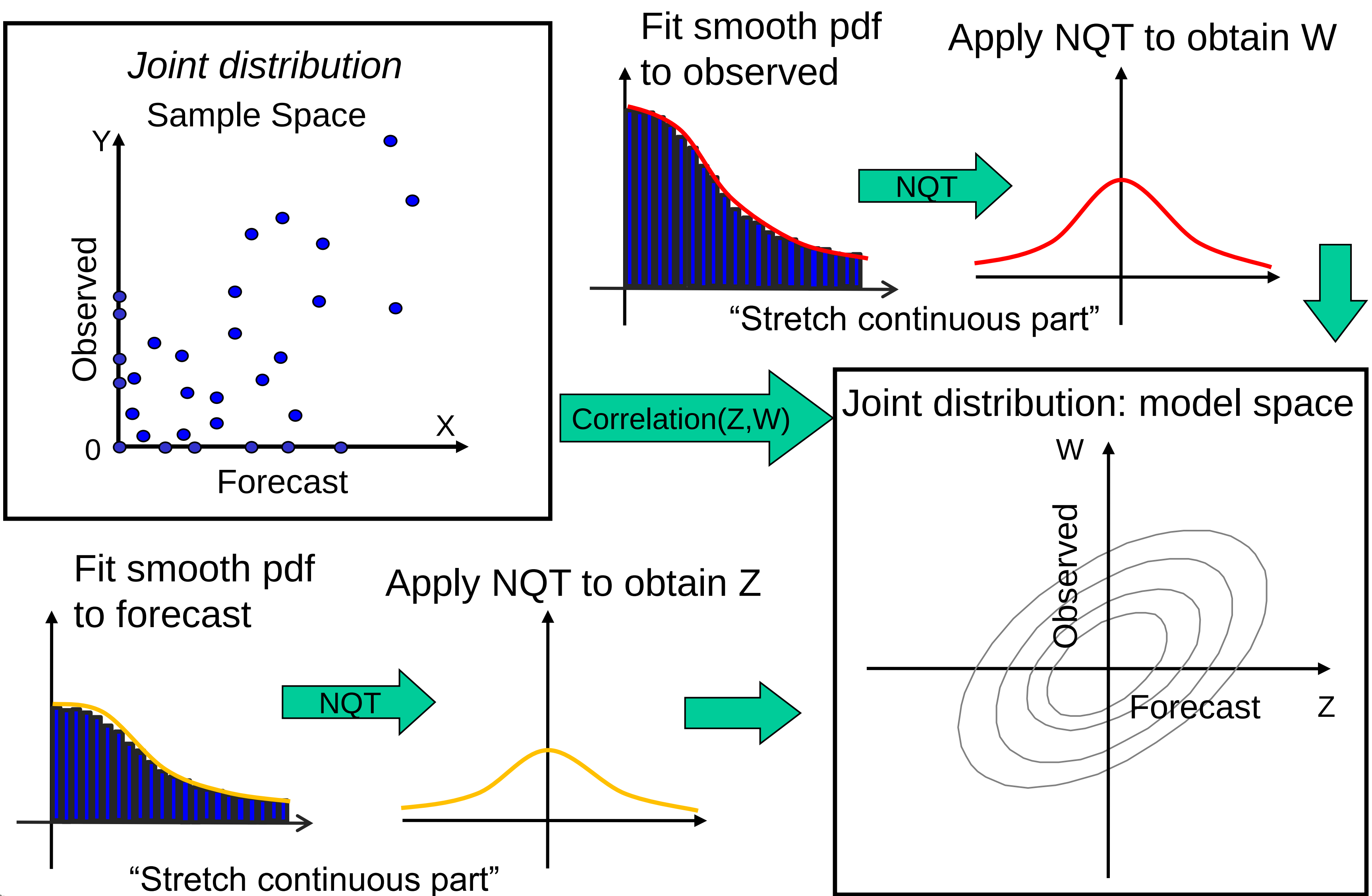
  $H(x,y) = B(Z, W; \rho)$, where
  $Z = Q^{-1}(F_X(X))$
  $W = Q^{-1}(F_Y(Y))$
  } Normal Quantile Transform (NQT)

  B is bivariate standard normal distribution function.
  Q is standard normal distribution function.
  ρ is correlation coefficient between *Z* and *W*.

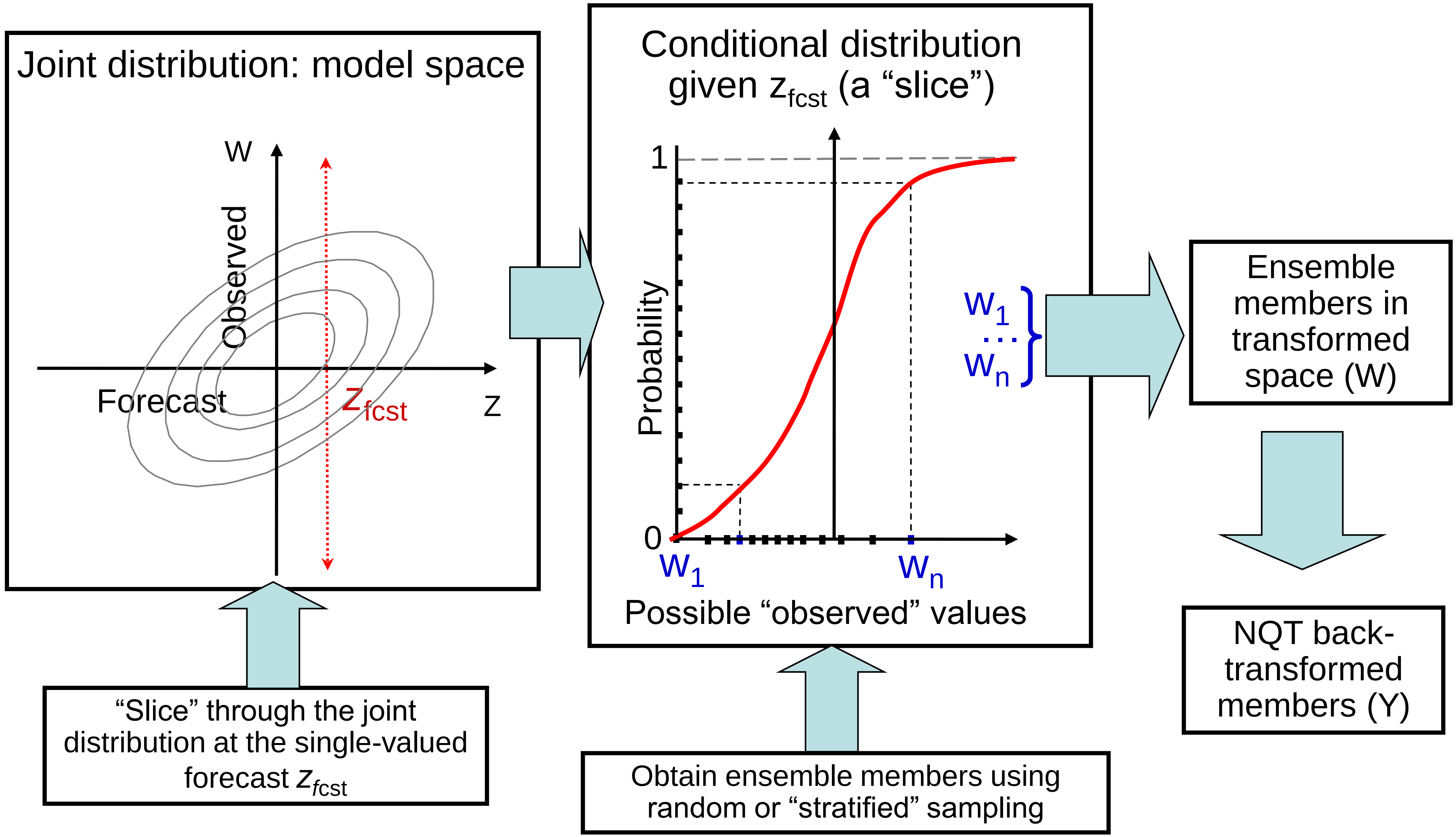
- Our assumption is that *F*(*x,y*) can be well approximated by *H*(*x,y*).

- Real-time forecast is then given by conditional distribution *H*(*y*|*x*). The members sampled from this must be back-transformed (inverse NQT).

Meta-Gaussian model: calibration



Meta-Gaussian model: ensembles



Precipitation intermittency

- Problem: NQT requires continuous variables, precipitation is mixed

  Solution: "explicit" or "implicit" treatment of precipitation

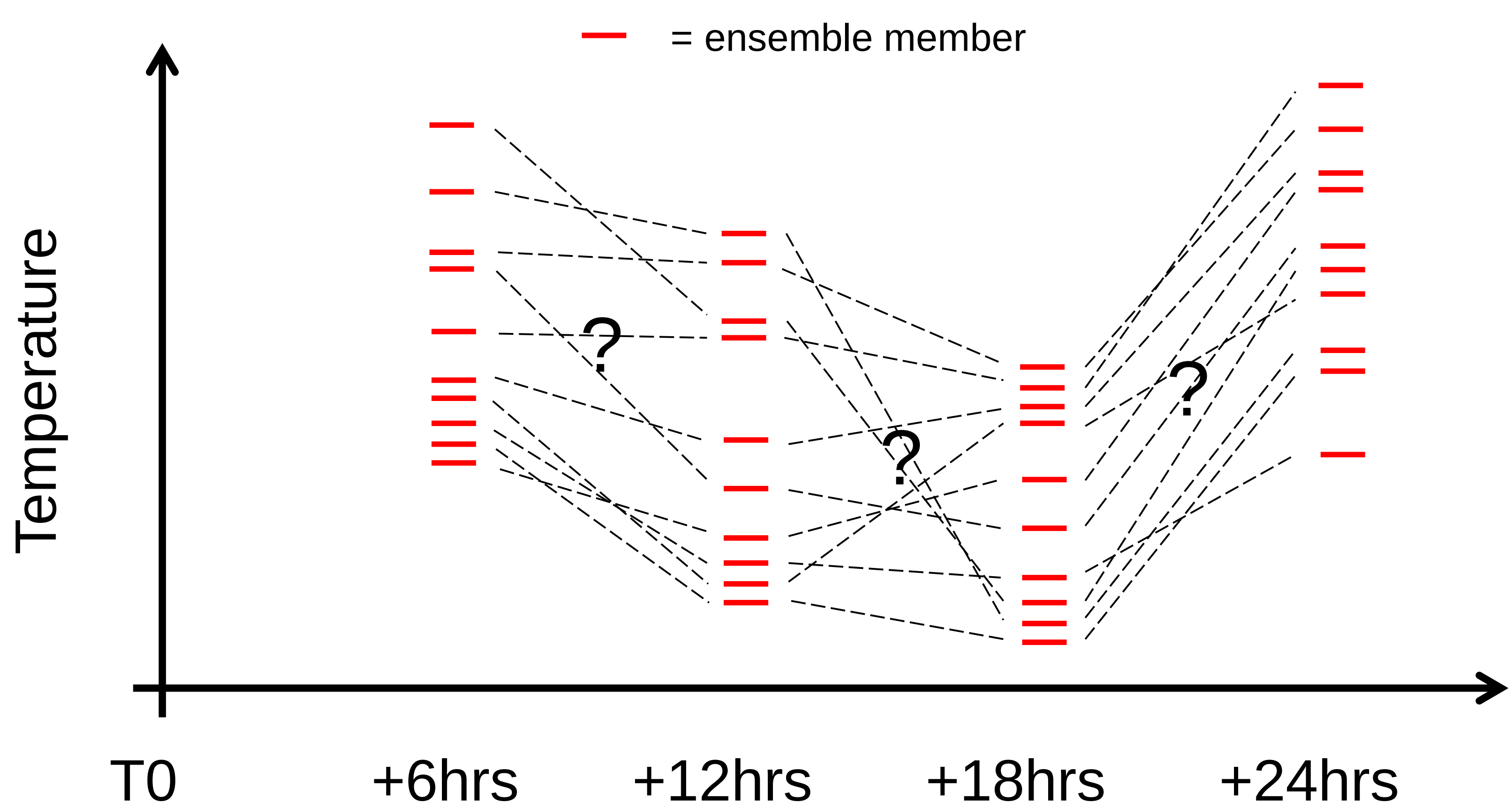
- Explicit treatment: the mixed-type meta-Gaussian model (Herr and Krzysztofowicz, 2005; Wu et al., 2011). Breaks the distribution into two parts. This approach works better for short time scales for which probability of precipitation (PoP) is low, i.e. dryer conditions

- Implicit treatment: similar to original meta-Gaussian model (Schaake et al., 2007; Wu et al., 2011). Defines a positive threshold above which continuous modeling occurs. May work better for longer aggregation periods and wet conditions where PoP is high.

Temperature ensemble generation

- Obtain daily minimum and maximum temperatures: convert observed and forecast time-series to daily minimum and maximum time series using a diurnal relationship.
- Apply MEFP to the daily minimum and maximum time series to produce daily maximum and minimum ensembles (using similar procedures for precipitation ensemble generation).
- The daily minimum and maximum ensembles are back-transformed to instantaneous values using the inverse of the diurnal relationship.

Preserving temporal patterns



- For precipitation too, and between precipitation and temperature. And in space. A lot of dots to join!

Schaake Shuffle: pragmatic choice

- Meteorological events are correlated in space and time.
  - Temperatures tend to be correlated from basin to basin and from one day to the next, as well as during the day
  - Large-scale storms can be more persistent in space and time than rain showers.
  - There are also relationships between variables. For example, precipitation may not occur on the days with the highest temperatures.

- These connections or correlations can be approximated by the rank structure of the historical observations for the same location and time period over multiple years

- The Schaake-Shuffle (SS) arranges the ensemble members to have the same rank structure as the historical observations, i.e. it “maps” the rank structure of the observations to the ensemble forecasts