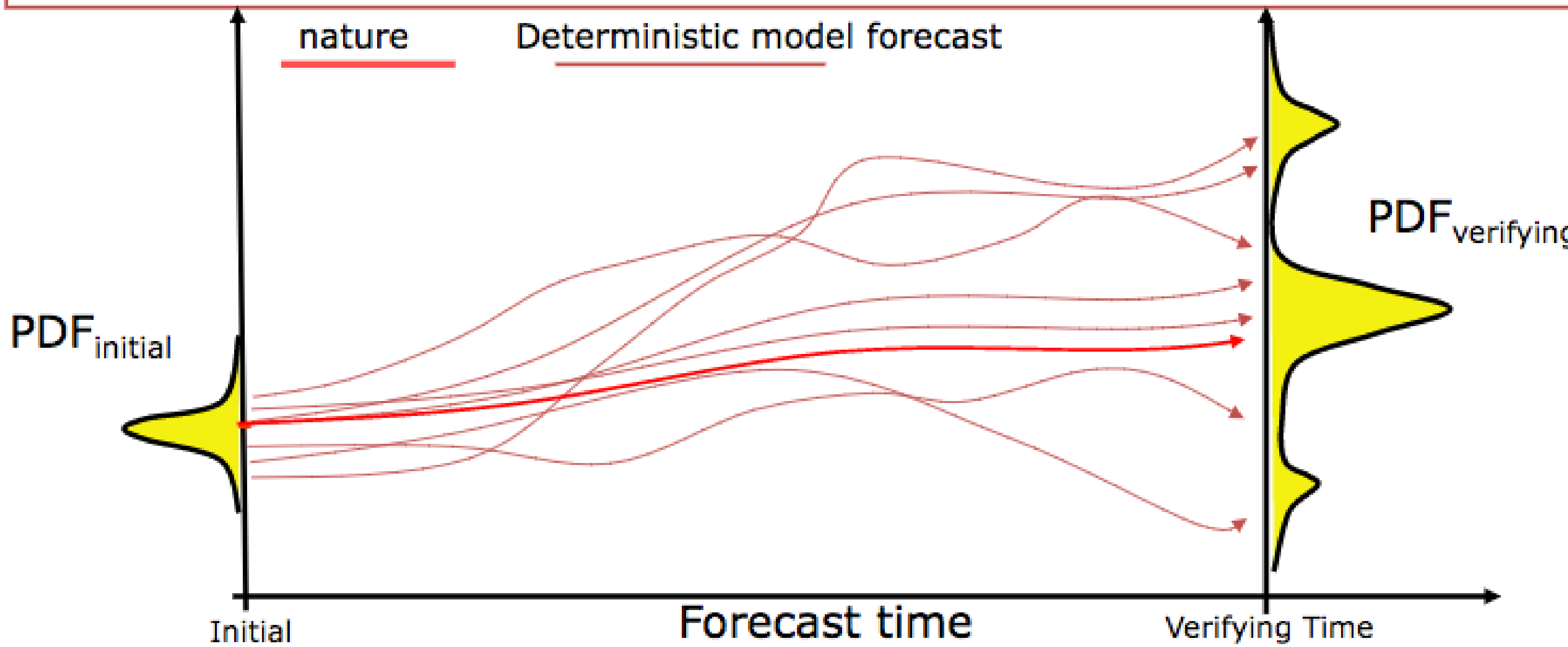
Assessing the Quality of Subseasonal/Seasonal Ensemble Forecasts

S. Schubert: 15 March 2017

Ensemble forecast

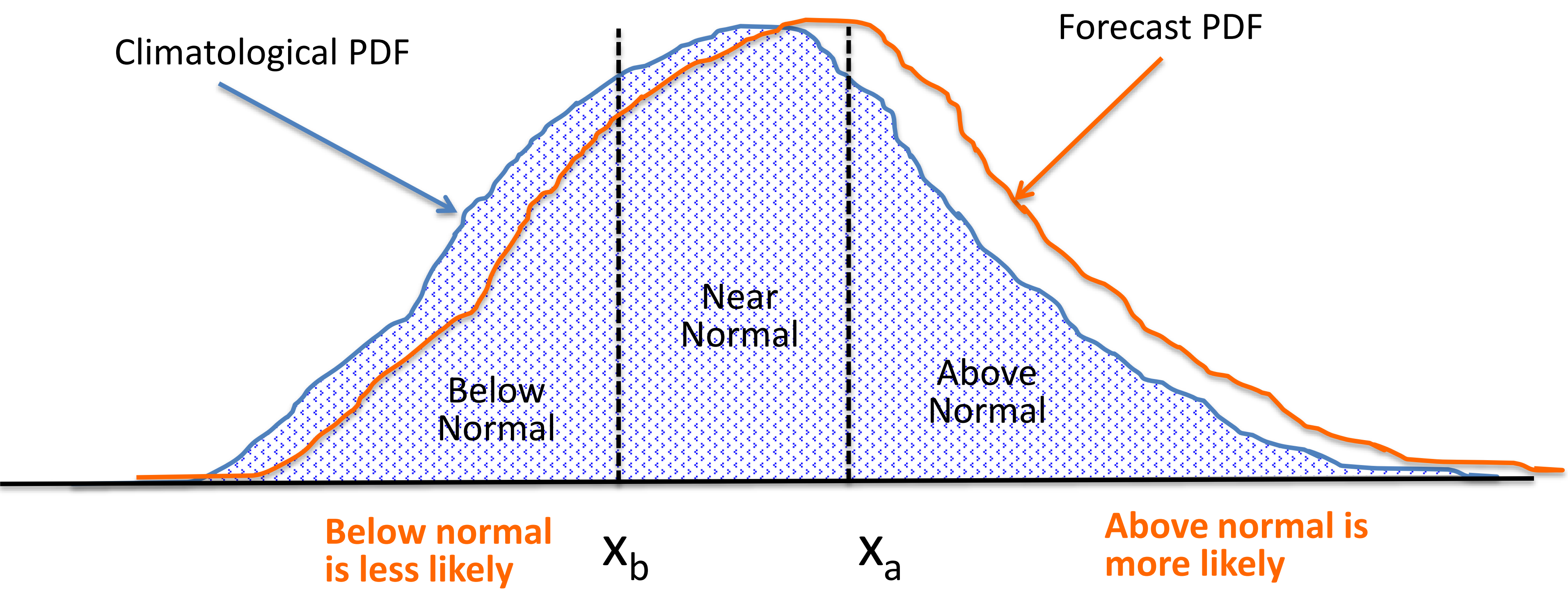
- A collection of two or more forecasts verifying at the same time
- Ensemble forecasting: propagate into de future the probability distribution function reflecting the uncertainty associated with initial conditions and model limitations
- Practical approach: running a set of deterministic runs whose initial conditions sample the initial PDF



8

Malaquías Peña
Environmental Modeling Center, NCEP/NOAA

More typical seasonal forecast
(Small shift in PDF)



Forecasts are evaluated in terms of integrated measures such as terciles e.g.: how much have the climatological probabilities (1/3, 1/3, 1/3) for below, near and above normal changed (e.g., 0.25, 0.30, 0.45)

Assessing the quality of a forecast system

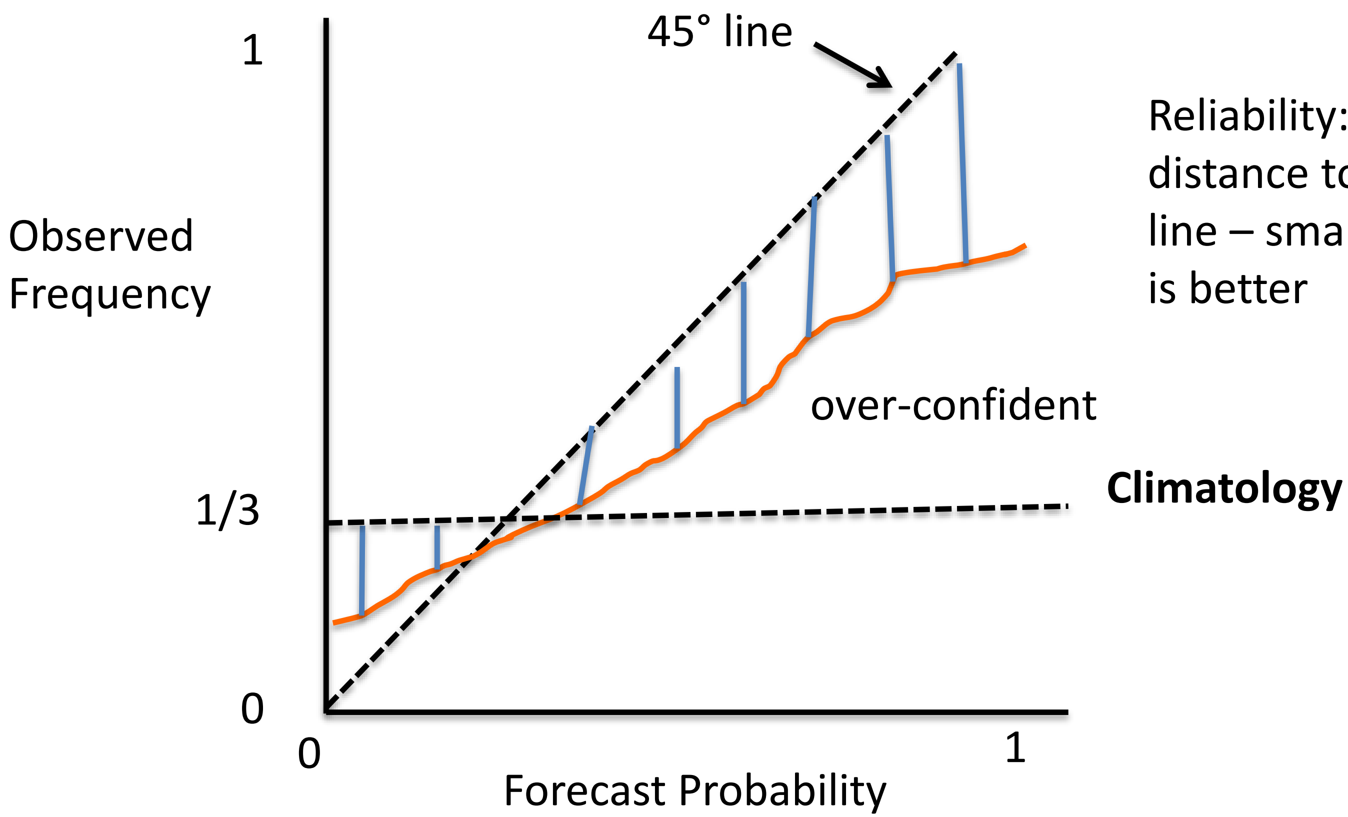
- Characteristics of a forecast system:
  - **Consistency**: Do the observations statistically belong to the distributions of the forecast ensembles? (consistent degree of ensemble dispersion)
  - **Reliability**: Can I trust the probabilities to mean what they say?
  - **Sharpness**: How much do the forecasts differ from the climatological mean probabilities of the event?
  - **Resolution**: How much do the forecasts differ from the climatological mean probabilities of the even, and the systems gets it right?
  - **Skill**: Are the forecasts better than my reference system (chance, climatology, persistence,…)?

Rank Histogram

Reliability Diagram

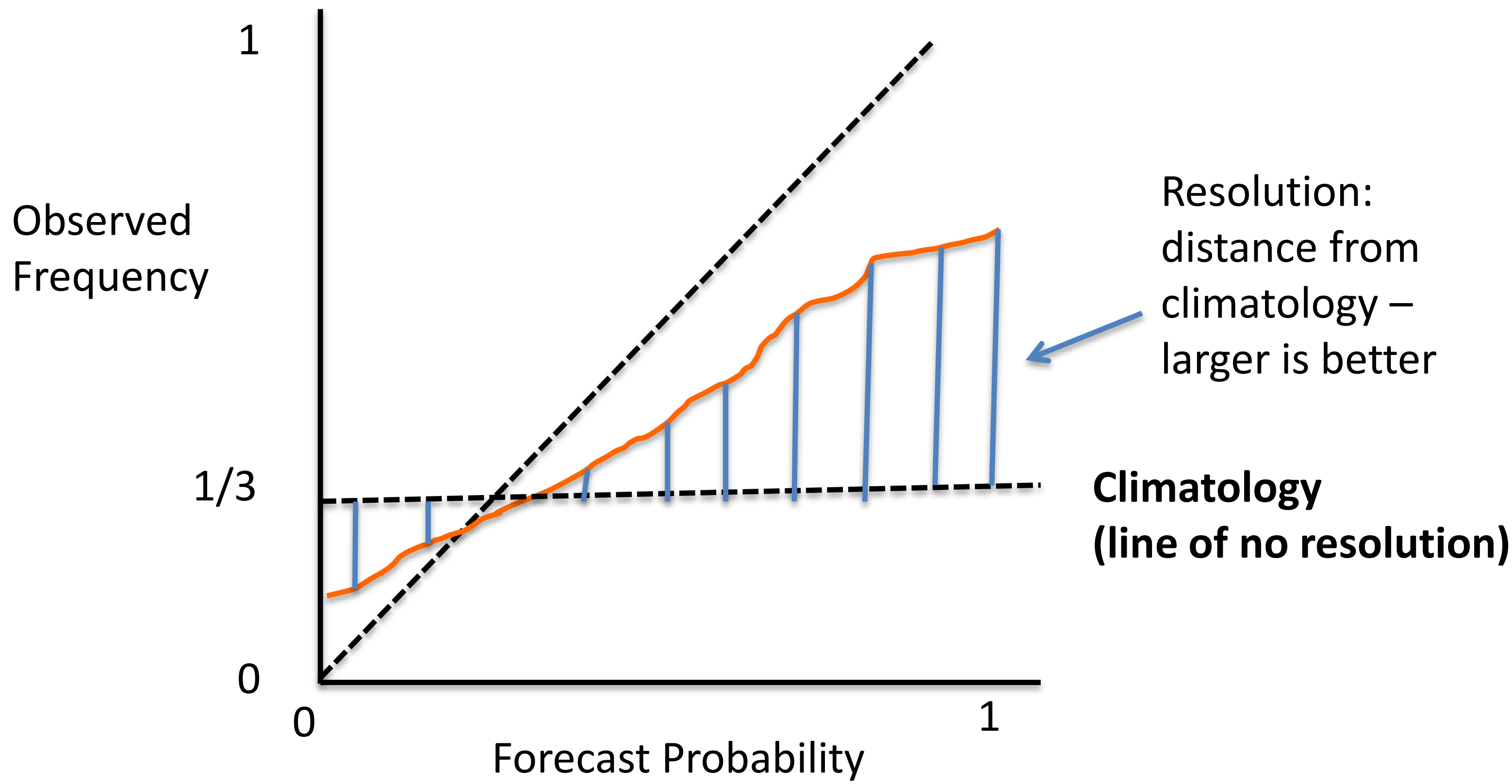
Brier Skill Score

Reliability
(e.g., that temperature falls in the upper tercile)



Reliability: distance to 45° line – smaller is better

Resolution



1
Observed
Frequency
1/3
0
0
1
Forecast Probability
Resolution:
distance from
climatology –
larger is better
Climatology
(line of no resolution)

Brier Score Decomposition Relates Reliability and Resolution

(*mean* squared difference between forecast probabilities and actual outcomes)

$$\mathrm{BS} = \frac{1}{n}\sum_{k=1}^{n}(f_k - x_k)^2$$

where

f_k = forecast probability on occasion k

x_k = observation (0 or 1) on occasion k

BS can be decomposed into 3 components that represent important properties of the forecasts:

$$\mathrm{BS} = \underbrace{\frac{1}{n}\sum_{i=1}^{I}N_i(f_i - \bar{x}_i)^2}_{\text{Reliability}} - \underbrace{\frac{1}{n}\sum_{i=1}^{I}N_i(\bar{x}_i - \bar{x})^2}_{\text{Resolution}} + \underbrace{\bar{x}(1-\bar{x})}_{\text{Uncertainty}}$$

Brier Skill Score (BSS)

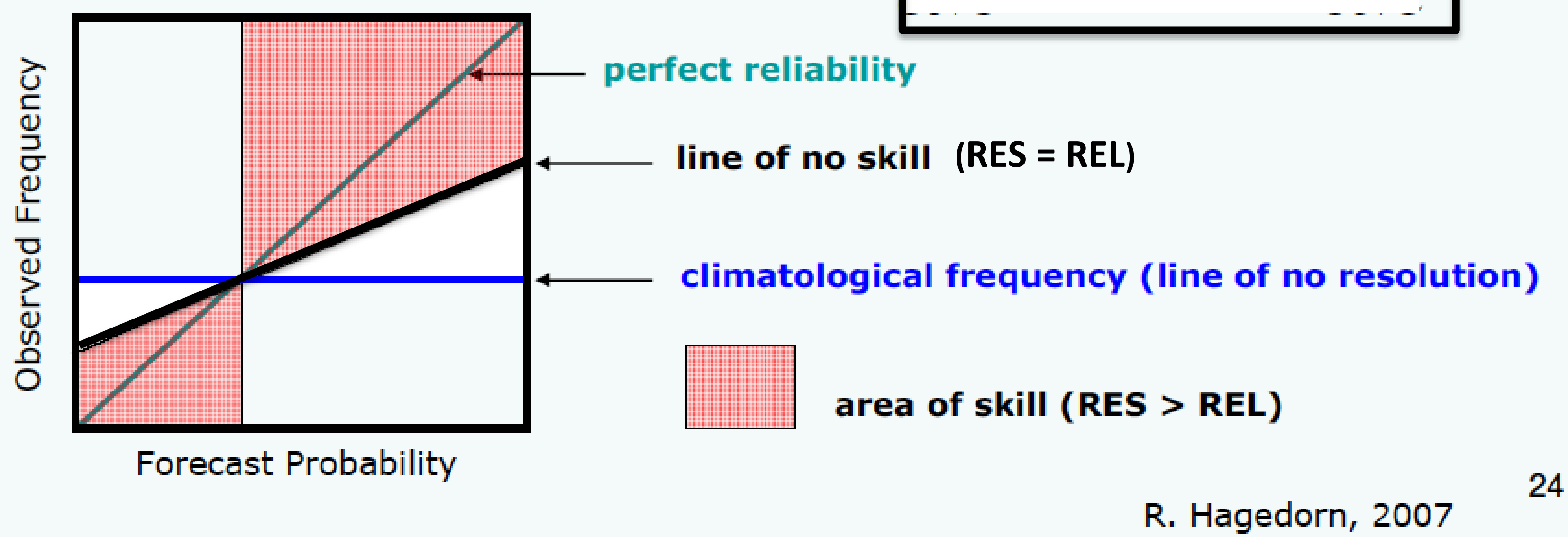
$$BSS = -\frac{\mathrm{BS} - \mathrm{BS}_{\mathrm{ref}}}{\mathrm{BS}_{\mathrm{ref}}}$$

If reference forecast is climatology then:

$$\boxed{\mathrm{BSS} = \frac{\mathrm{RES} - \mathrm{REL}}{\mathrm{UNC}}}$$

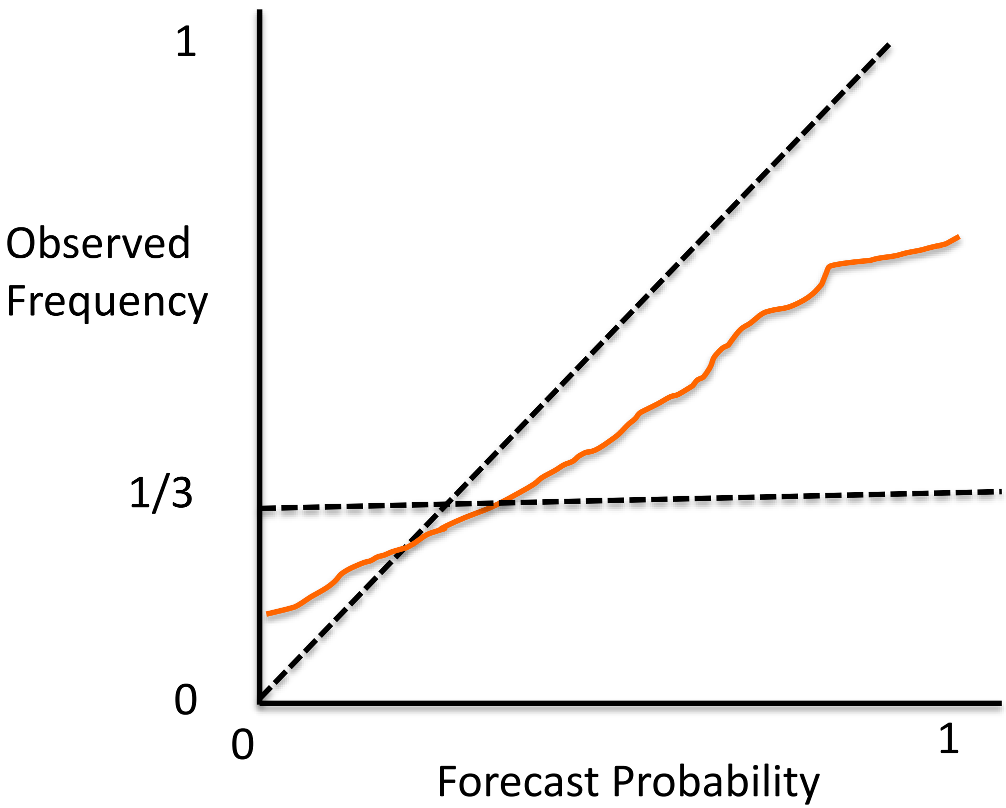
BSS and Reliability Diagram

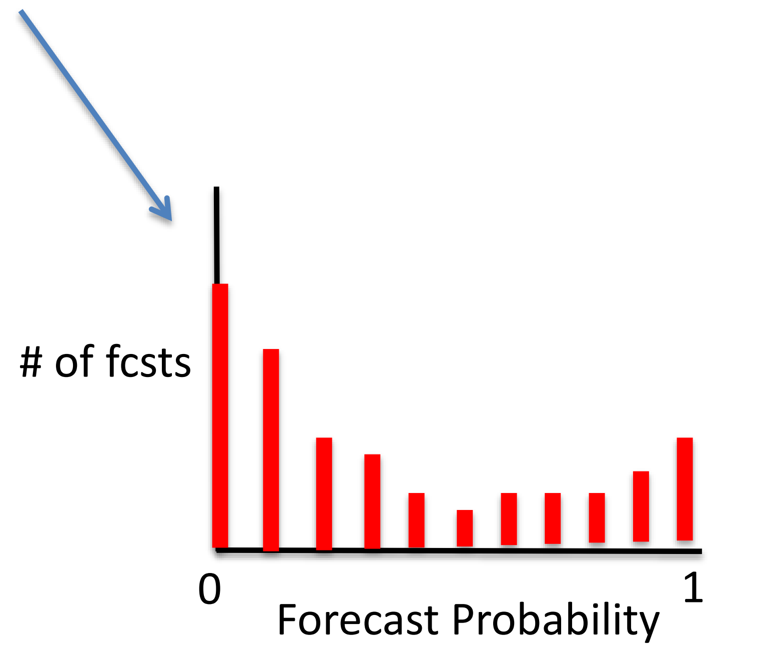
$$BSS = \frac{\text{Res} - \text{Rel}}{\text{Unc}}$$

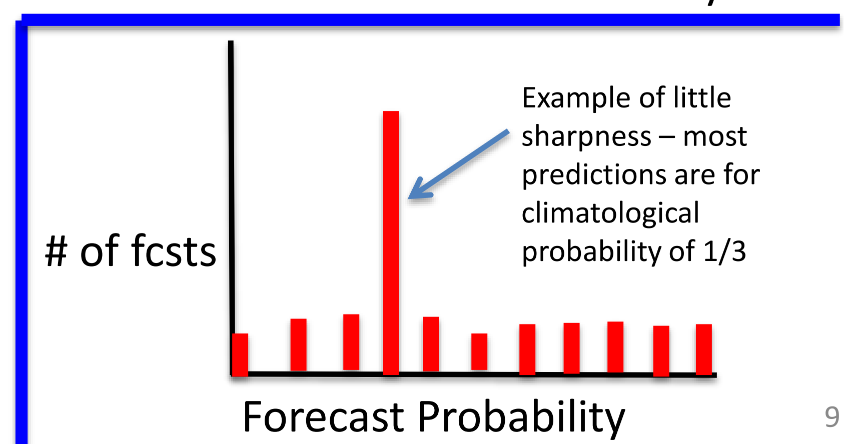


Perfect forecast: BSS=1
Reference forecast (climatology): BSS=0
Better (Worse) than reference forecast: BSS>0 (BSS< 0)

Sharpness (forecast frequency histogram)

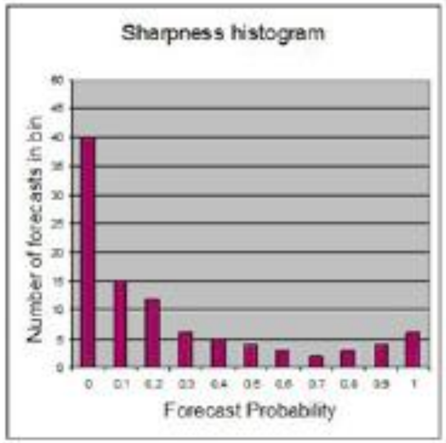


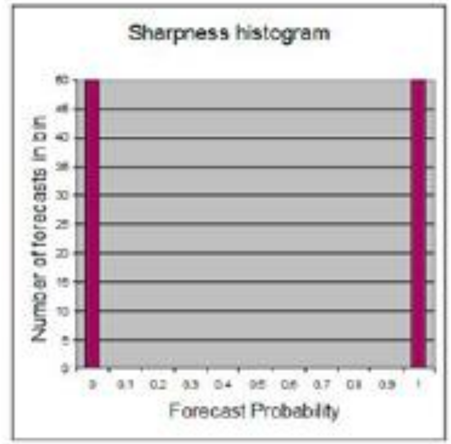




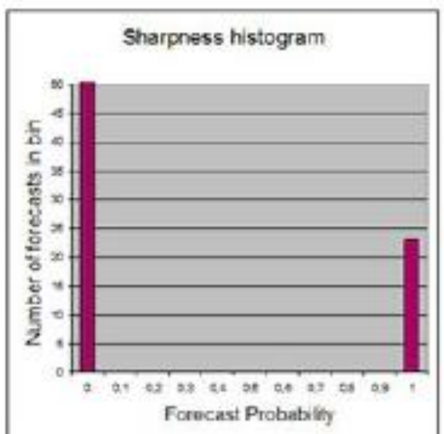
Sharpness Histogram Exercise

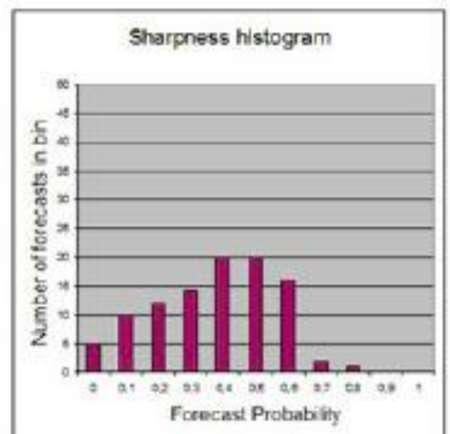
Reasonable sharpness





Maximum possible sharpness





Poor sharpness

Laurence J. Wilson
Atmospheric Science and Technology Branch
Environment Canada

Some Results from NMME

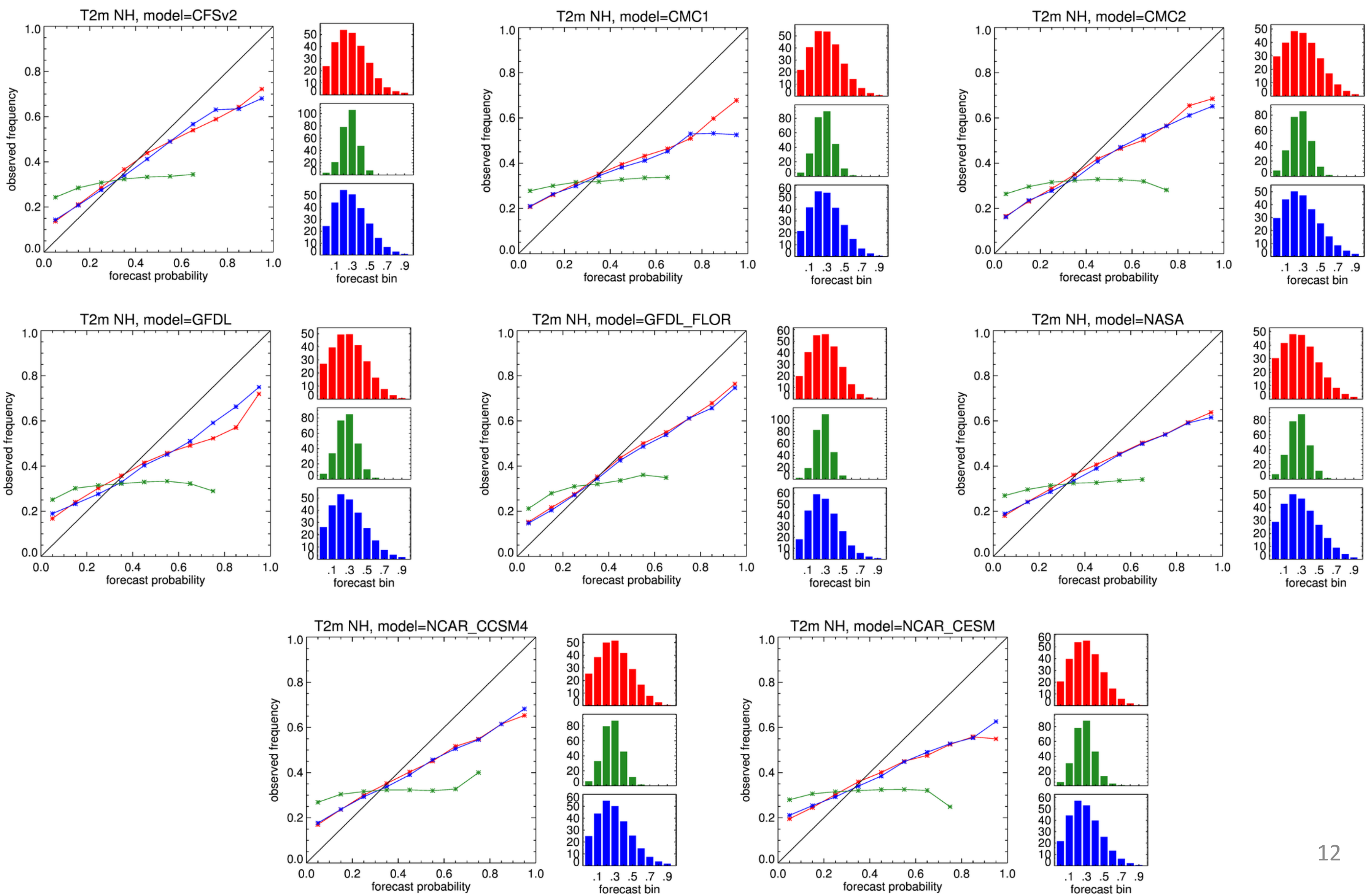
(courtesy Emily Becker NOAA/CPC)

- Reliability diagrams showing the three-category reliability and sharpness for each individual model, lead-1 seasonal forecasts.
- T2m and precip are aggregated over northern hemisphere (23N-75N) land, and SST is aggregated over the Nino3.4 region.
- 1982-2010 retrospective forecasts are from the CPC in-house NMME data set, leave-one-out cross-validation.
- T2m observational data is from GHCN+CAMS, precip obs data is CAMS, and SST obs data is OISSTv2.

lead-1 seasonal forecasts, 1982-2010
Northern Hemisphere (23N-75N) land

Reliability Diagram: **T2** NH, 3 category

Courtesy Emily Becker NOAA/CPC

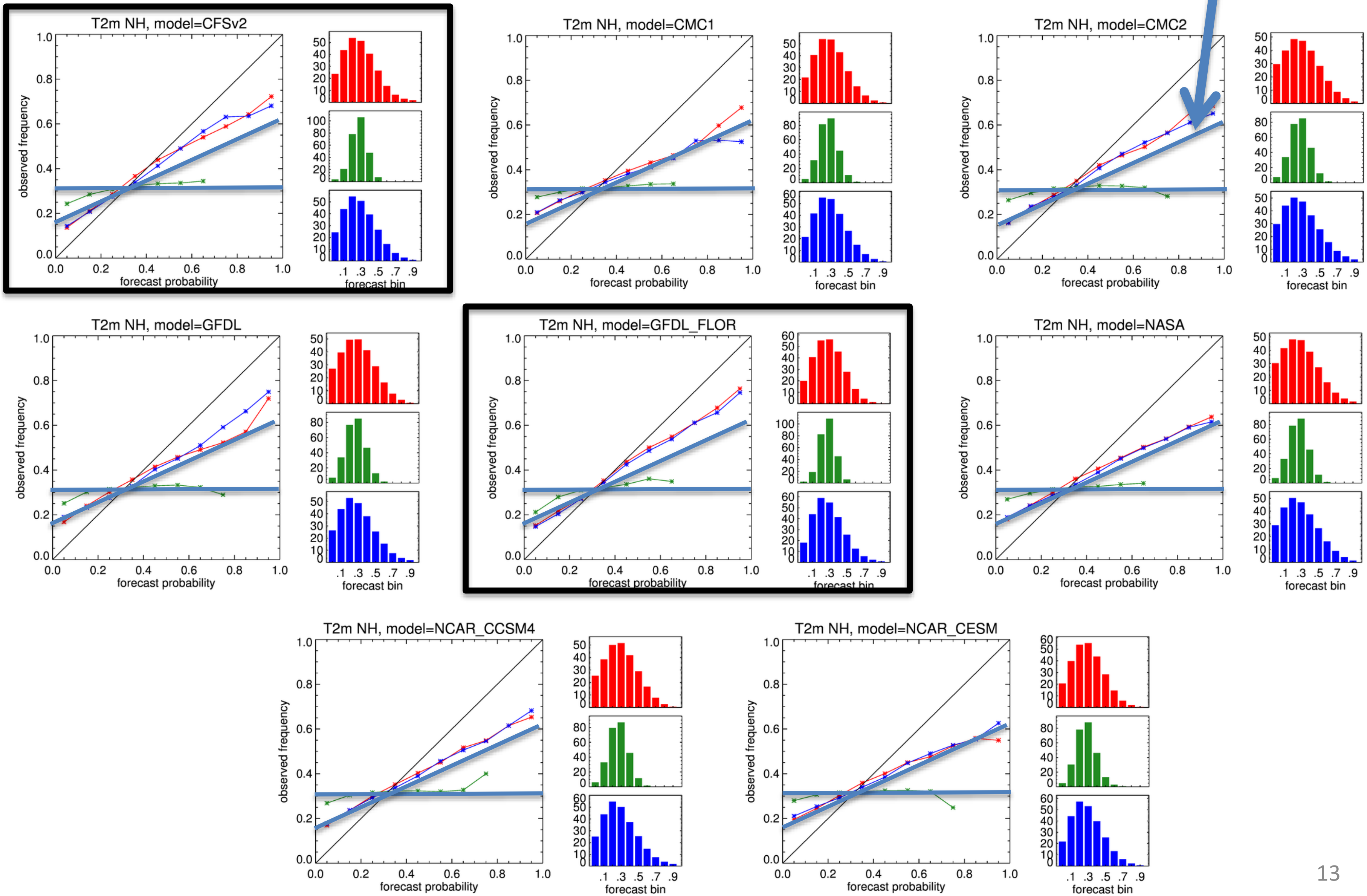


lead-1 seasonal forecasts, 1982-2010
Northern Hemisphere (23N-75N) land

Reliability Diagram: **T2** NH, 3 category,

Courtesy Emily Becker NOAA/CPC

Line of no skill

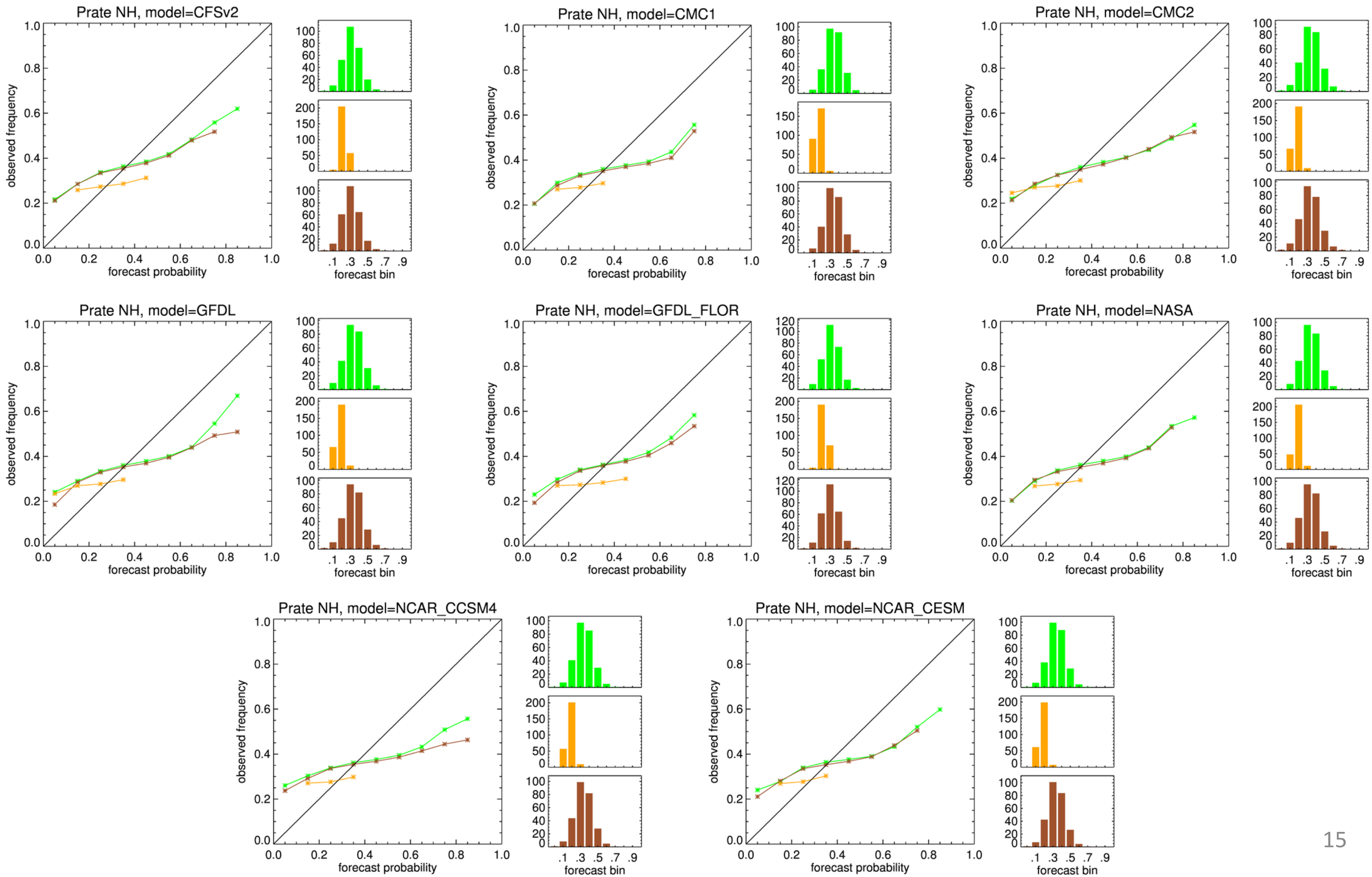


Model	Hindcast Period	No. of Members	Arrangement of Members	Lead (month)	Model resolution (atmos)	Model resolution (ocean)	Reference
Active							
NCEP/CFSv2	1982-2010	24 (28)	4 members (0, 6, 12, 18z) every 5^{th} day	0-9	T126L64	MOM4L40 .25deg Eq	Saha et al (2010)
GFDL/CM2.1	1982-2010	10	All 1^{st} of the month 0Z	0-11	2x2.5degL24	MOM4L50 .3deg Eq	Delworth (2006)
GFDL/CM2.5 (FLOR)	1982-present	24	All 1^{st} of the month 0Z	0-11	C18L32 (50km)	MOM5 L50 0.30 deg Eq 1degPolar1.5	Vecchi et al (2014)
CMC1-CanCM3	1981-2010	10	All 1^{st} of the month 0Z	0-11	CanAM3 T63L31	CanOM4L40 .94deg Eq	Merryfield et al (2013)
CMC1-CanCM4	1981-2010	10	All 1^{st} of the month 0Z	0-11	CanAM4 T63L35	CanOM4L40 .94deg Eq	Merryfield et al (2013)
NCAR/CCSM4	1982-2010	10	All 1^{st} of the month 0Z	0-11	0.9x1.25degL26	POPL60 .25deg Eq	Kirtman et al. (in prep)
NASA/GEOS5	1981-2010	11	4 mems every 5 days; 7 mems on last day of last month	0-9	1x1.25 deg L72	MOM4L40 .25deg Eq	Vernieres et al (2012)
NCAR/CESM1	1982-2010	10	All 1^{st} of the month 0Z	0-11	0.9x1.25degL30	POPL60 .25deg Eq	Tribbia et al.
Retired							
NCEP/CFSv1	1982-2009	15	1^{st} 0Z +/-2 days, 21^{st} 0z +/-2d, 11^{th} 0z +/-2d	0-8	T62L64	MOM3L40 0.30 deq Eq	Saha et al (2006)
NCAR/CCSM3	1982-2010	6	All 1^{st} of the month 0Z	0-11	T85L26	POPL42 0.3deg Eq	Kirtman and Min2009)
IRI-ECHAM4f	1982-2010	12	All 1^{st} of the month 0Z	0-7	T42L19	MOM3L25(1.5x0.5)	DeWitt (2005)
IRI-ECHAM4a	1982-2010	12	All 1^{st} of the month 0Z	0-7	T42L19	MOM3L25 (1.5x0.5)	DeWitt (2005)

lead-1 seasonal forecasts, 1982-2010
Northern Hemisphere (23N-75N) land

Reliability Diagram: **Precipitation** NH, 3 category

Courtesy Emily Becker NOAA/CPC

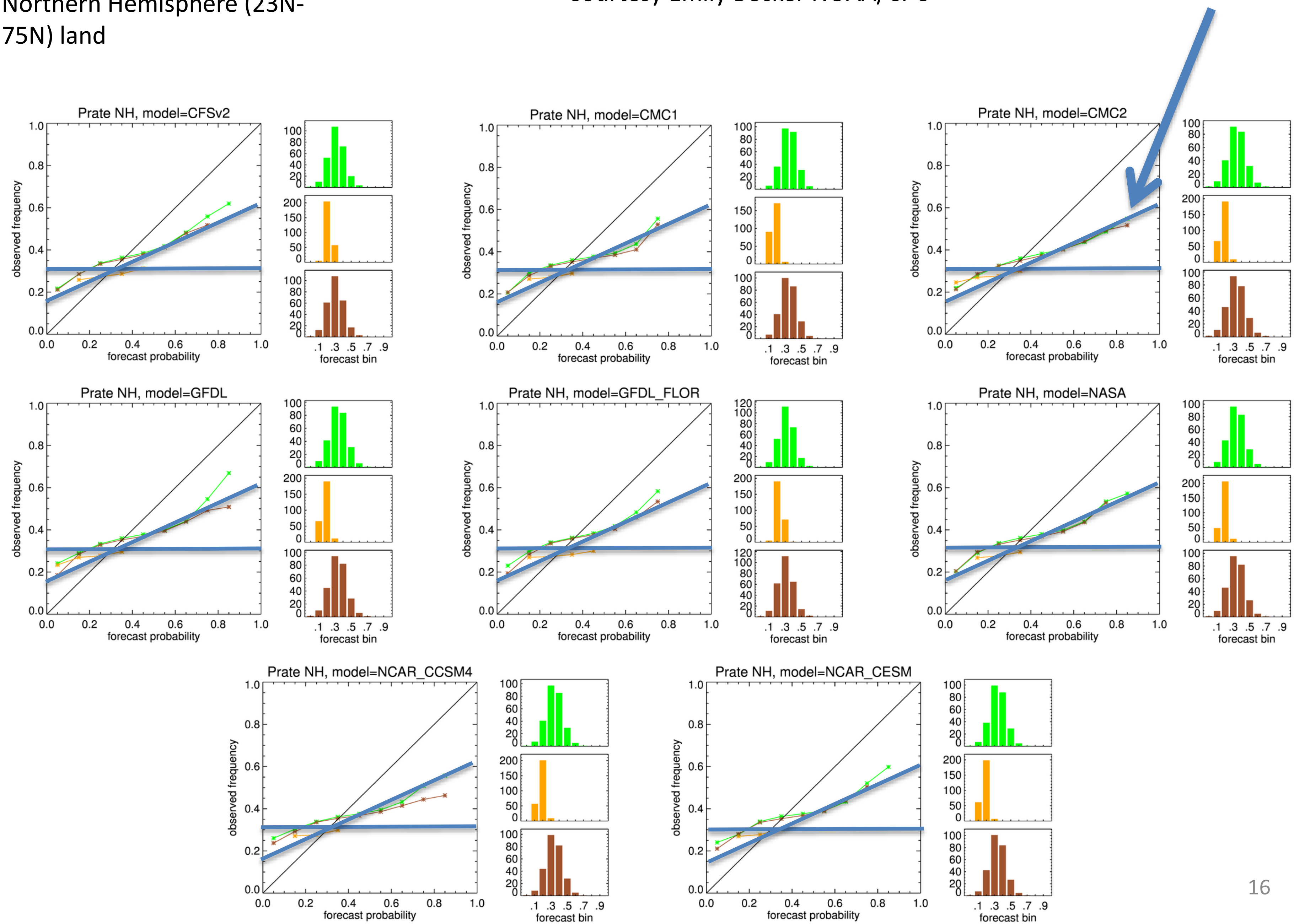


lead-1 seasonal forecasts, 1982-2010
Northern Hemisphere (23N-75N) land

Reliability Diagram: **Precipitation** NH, 3 category

Courtesy Emily Becker NOAA/CPC

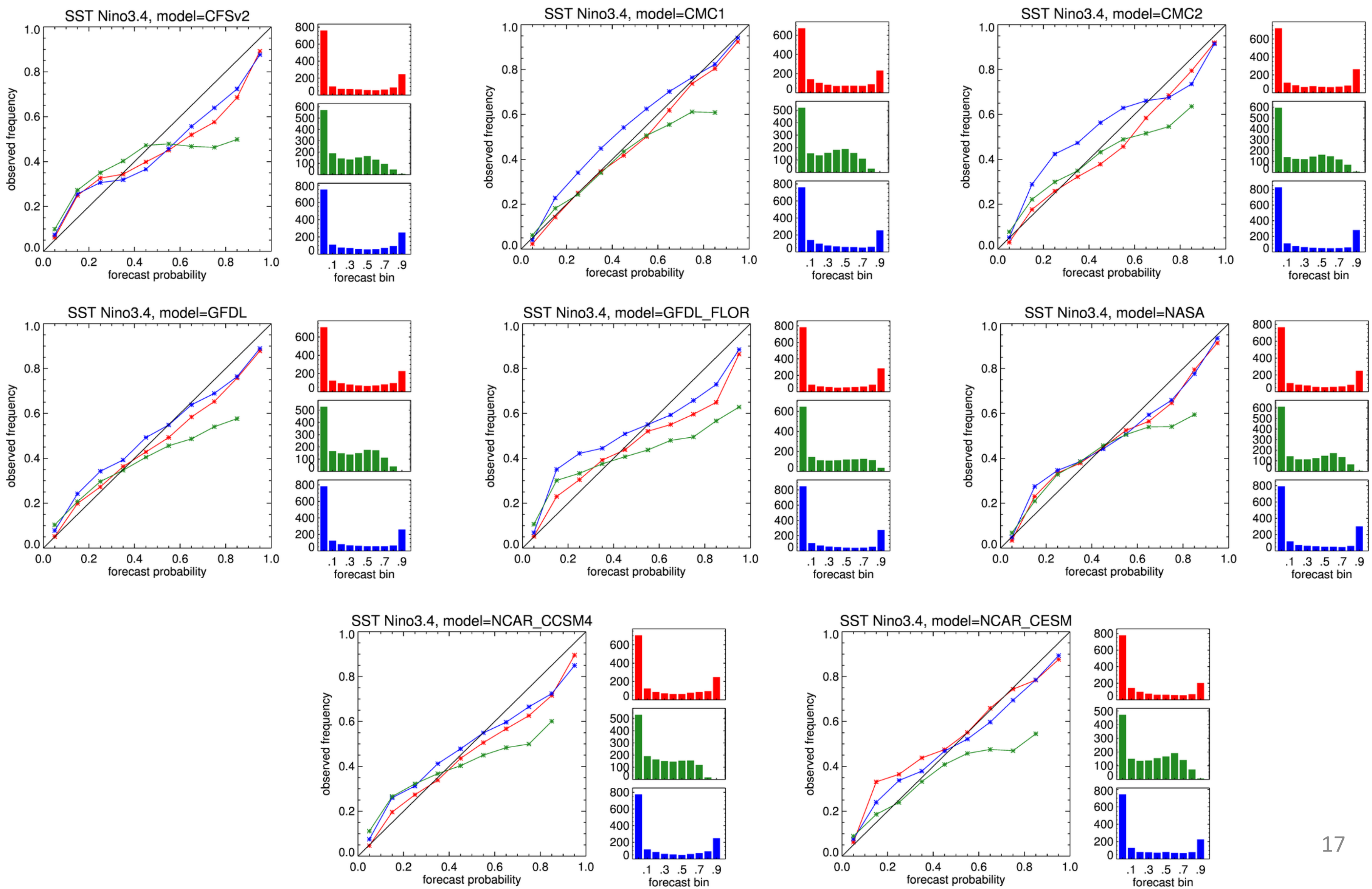
Line of no skill



lead-1 seasonal forecasts, 1982-2010
SST is aggregated over the Nino3.4 region

Reliability Diagram: **Nino3.4**, 3 category

Courtesy Emily Becker NOAA/CPC

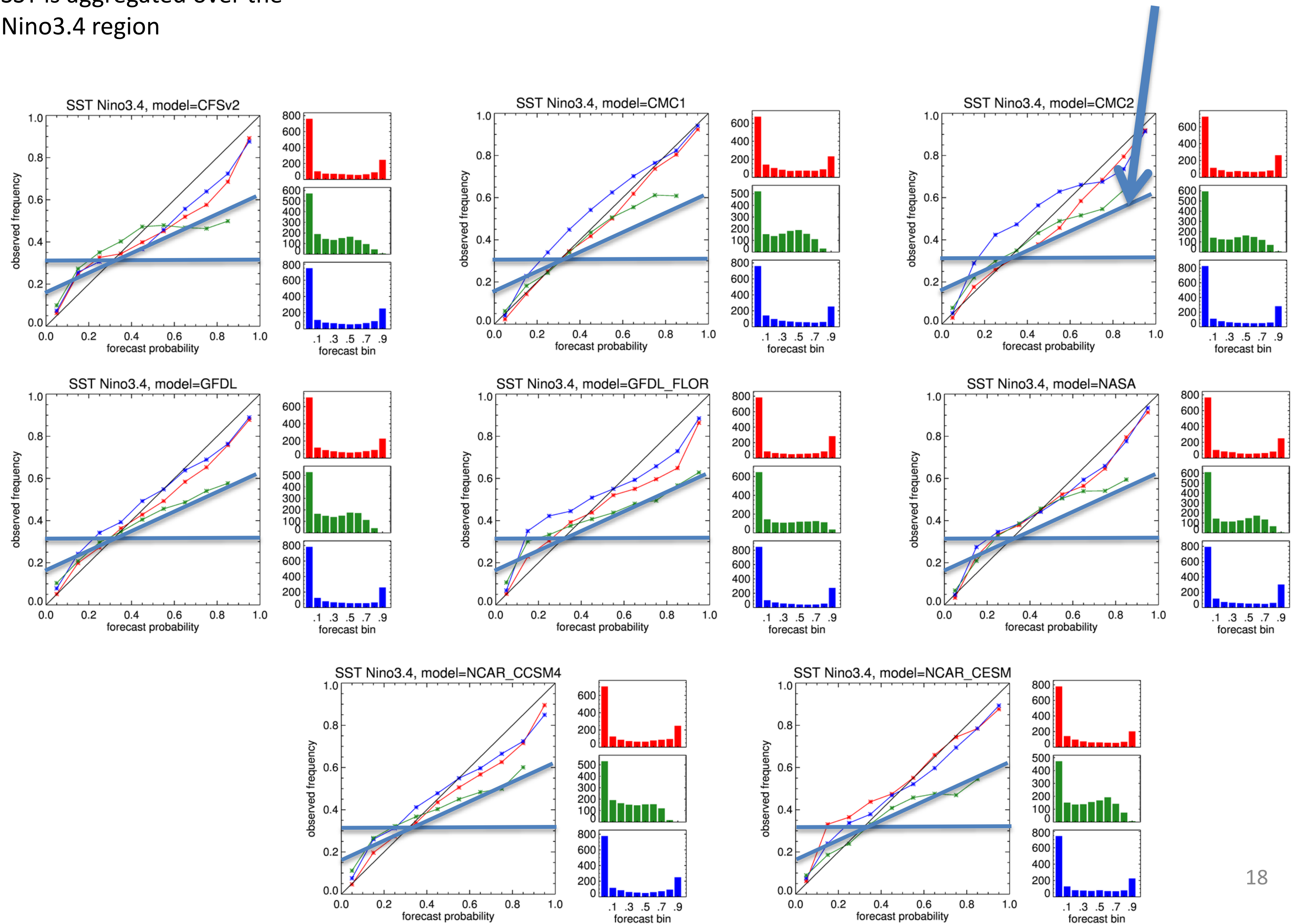


lead-1 seasonal forecasts, 1982-2010
SST is aggregated over the Nino3.4 region

Reliability Diagram: **Nino3.4,** 3 category

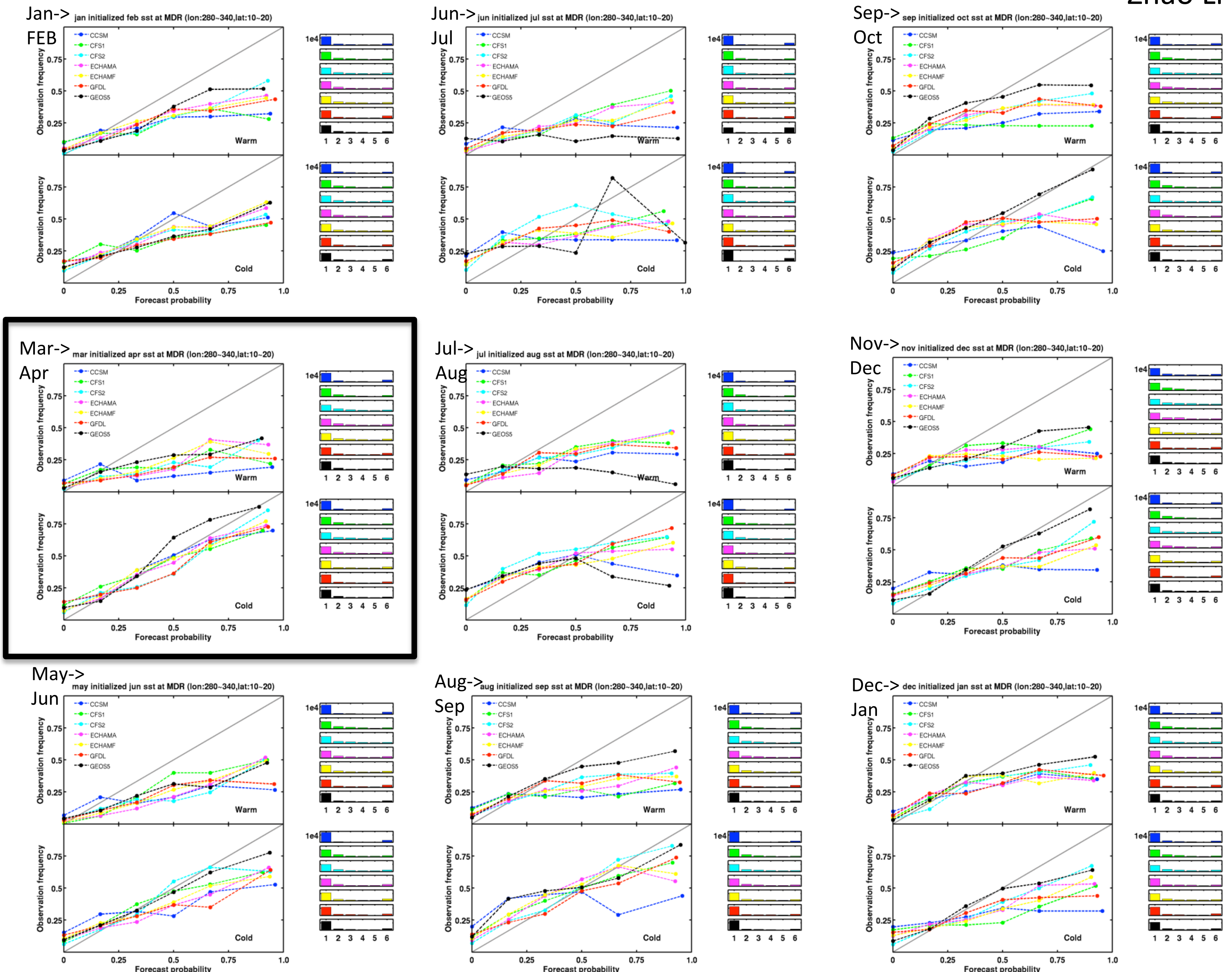
Courtesy Emily Becker NOAA/CPC

Line of no skill



SST in Main Development Region (1-Month Lead)

Zhao Li



Consistency

- do observations lie within the forecast ensemble spread?

Malaquias Peña
IMSG at EMC/NCEP/NOAA

Rank Histogram (Talagrand Diagram)

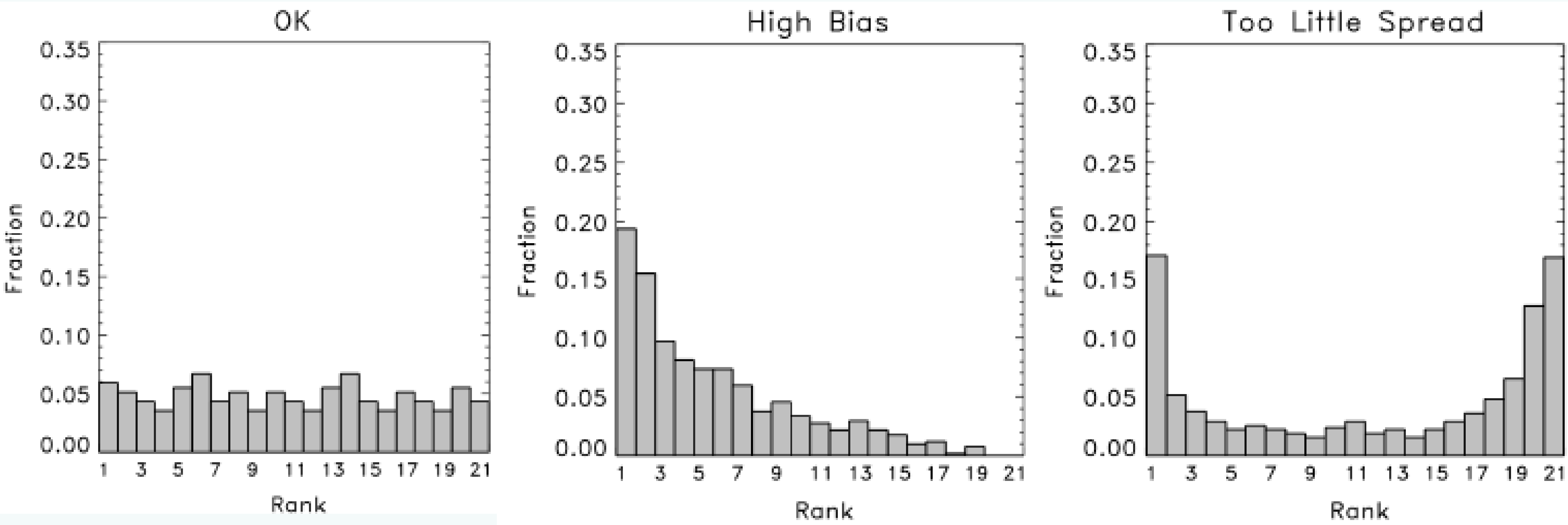
- Rank Histograms asses whether the ensemble spread is consistent with the assumption that the observations are statistically just another member of the forecast distribution
  - Check whether observations are equally distributed among predicted ensemble
  - Sort ensemble members in increasing order and determine where the observation lies with respect to the ensemble members

Rank 1 case

Rank 4 case

Malaquias Peña
IMSG at EMC/NCEP/NOAA

Rank Histograms



OBS is indistinguishable from any other ensemble member

OBS is too often below the ensemble members (biased forecast)

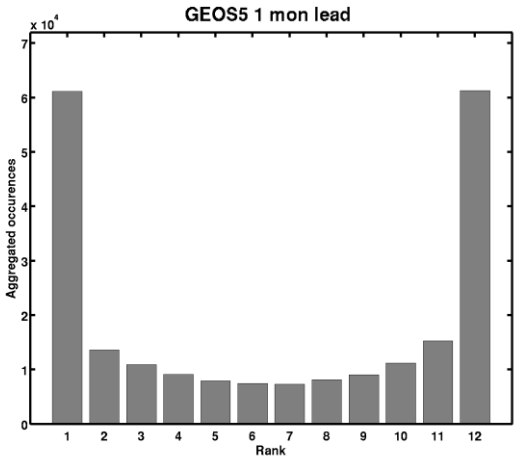
OBS is too often outside the ensemble spread

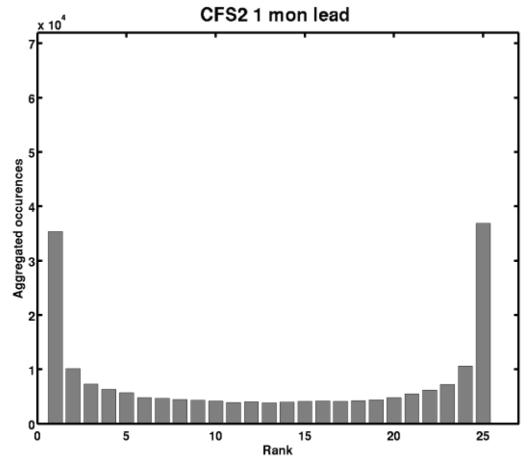
A uniform rank histogram is a necessary but not sufficient criterion for determining that the ensemble is reliable (see also: T. Hamill, 2001, MWR)

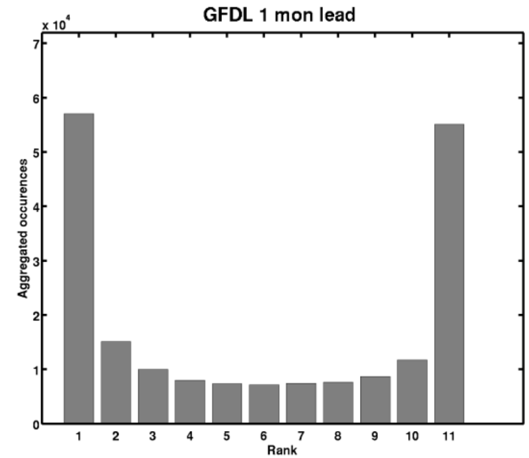
Rank Histograms:
1 month lead (seasonal mean)

GEOS-5 | CFS2 | GFDL

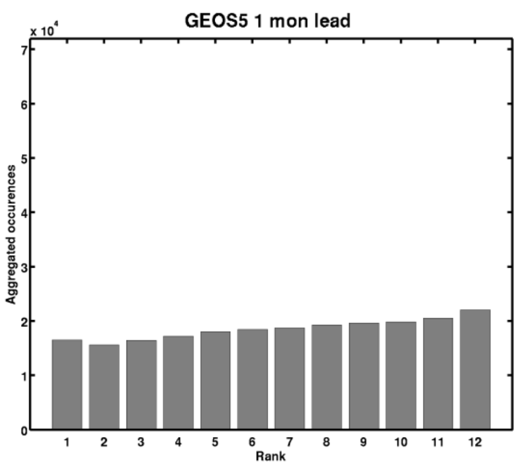
NINO34 SST

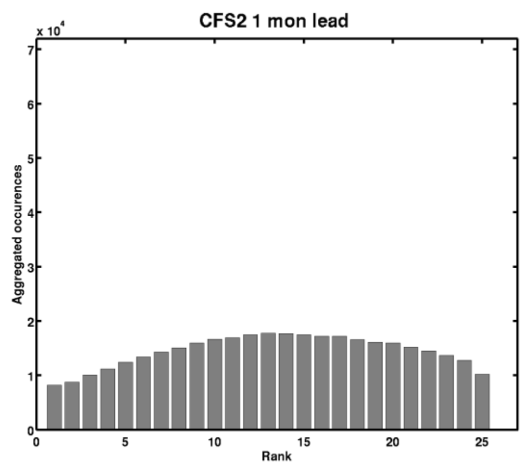


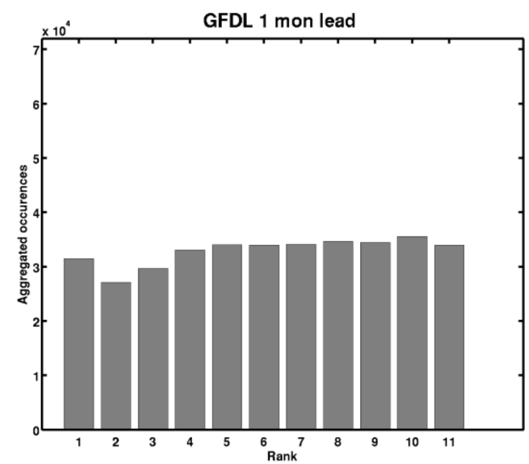




US Precip



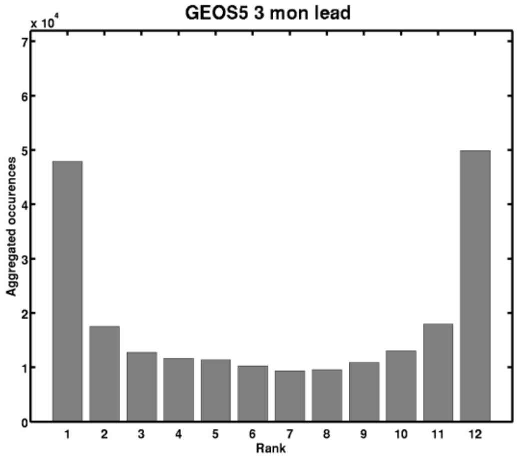


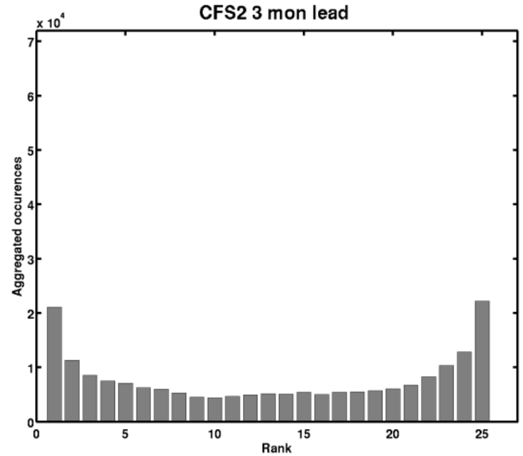


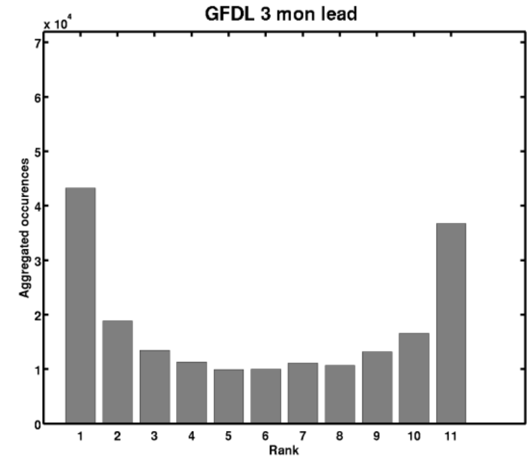
Rank Histograms:
3 month lead (seasonal mean)

GEOS-5 | CFS2 | GFDL

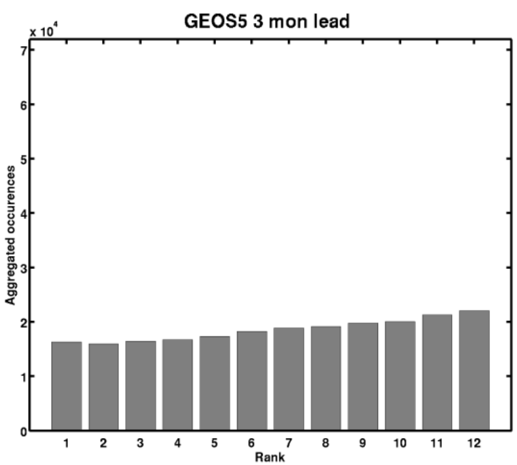
NINO3 SST

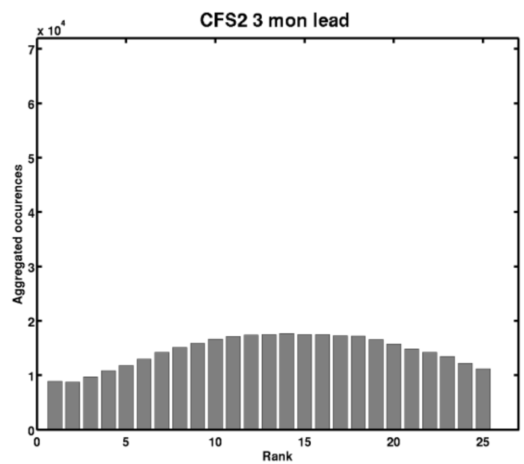


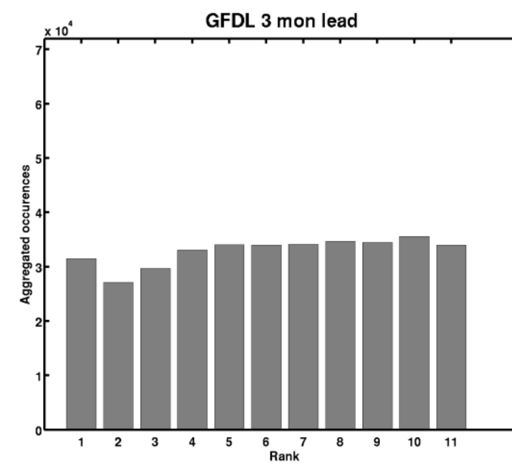




US Precip







Nino3.4: ratio of ensemble spread to standard error of estimate

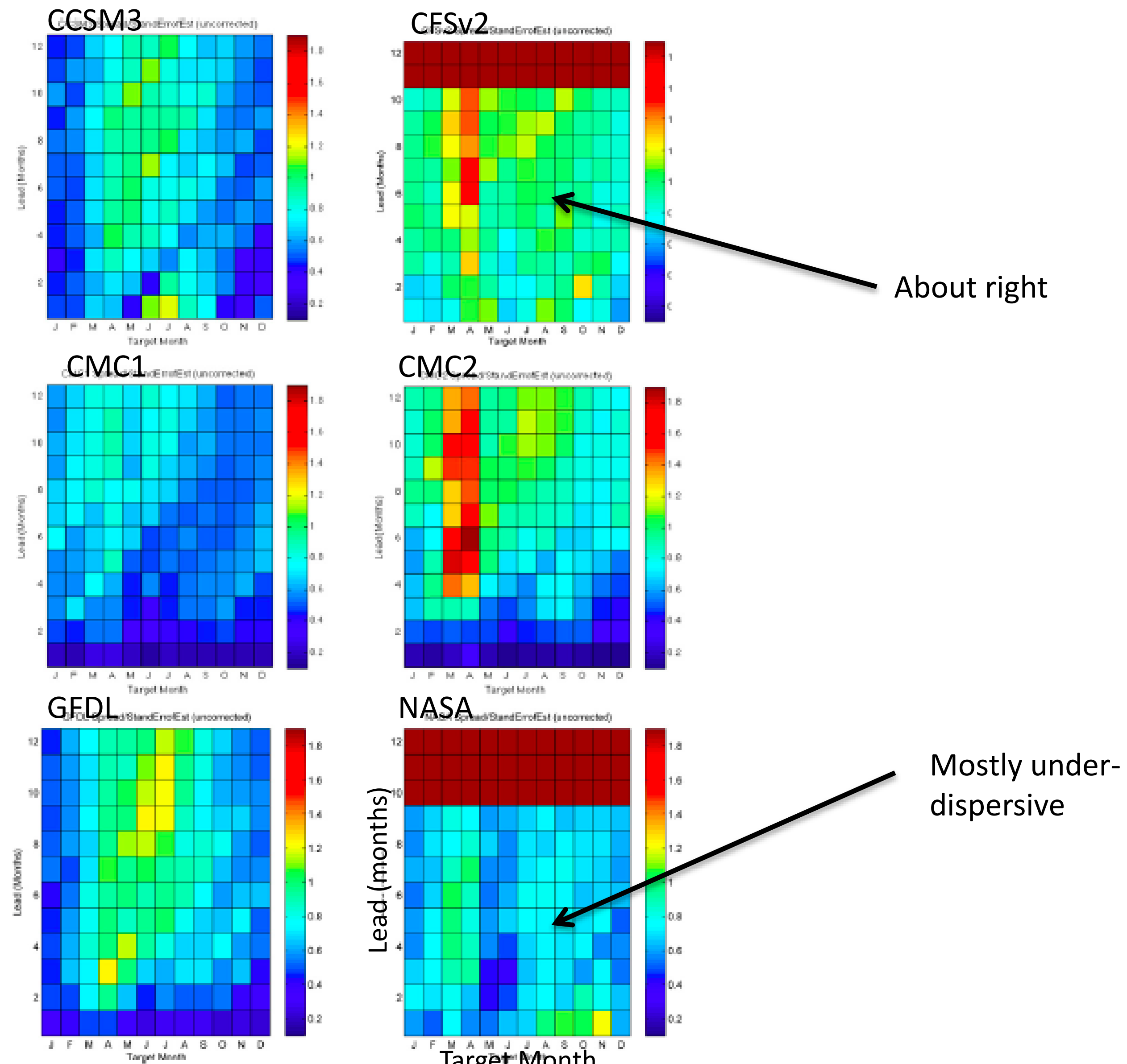


FIG. 6. Ratio of ensemble spread (as standard deviation of ensemble members about the ensemble mean) to the SEE based on the hindcast correlation skill, as a function of forecast target month (x axis) and lead time (y axis). Results are shown for the (top left) CCSM3, (top right) CFSv2, (middle left) CMC1, (middle right) CMC2, (bottom left) GFDL, and (bottom right) NASA models.

How to Improve Consistency?

- Calibrate the spread after the fact
  - probably something NMME/CPC should do (they only want the raw output)

- Improve initial ensemble to better sample the uncertainties
  - more ensemble members
  - choose the fastest growing (e.g., breeding)

- Perturb the model (much recent work done focusing on ECMWF system)[1]
  - addresses uncertainties due to model error
    - structural: missing physics
    - uncertainties in existing physics (e.g., perturbed parameter)
  - focus of research has been broader – seen as a way of improving skill, reducing bias, improving consistency, reliability, etc

[1]*Note: No perturbed physics of any kind is used in CFSv2: good ensemble spread may be the result of a forecast model that has good variability in the tropics on intraseasonal time scales. Initial conditions may also play a role in that they come from a weakly coupled data assimilation system (*Suranjana Saha *: EMC/NOAA)*

Perturbed Atmospheric Model- focus on convection

(Christensen *et al. 2015;* Weisheimer et al 2011)

- **perturbed parameter**: values of a selected set of parameters are sampled from a distribution representing the uncertainty in their values, with each ensemble member assigned a different set of parameters
    - Used by MetOffice in GLoSea5 - random parameter perturbation scheme (Bowler et al, 2008)

- **stochastic physics** schemes: e.g., a stochastic perturbation to the input to a deterministic scheme, such as by using a stochastic representation of convective available potential energy or convective inhibition

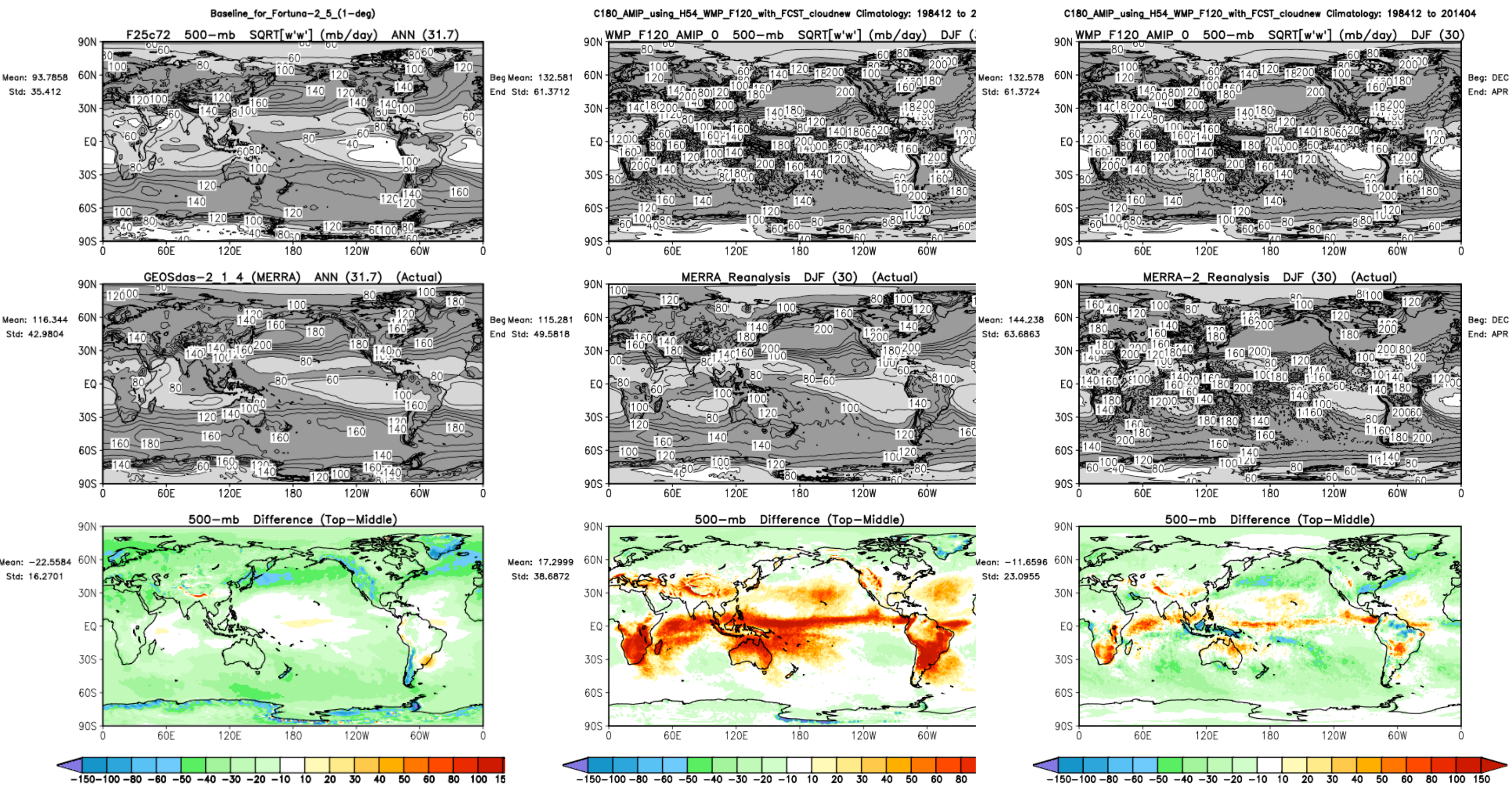
- perturbed parameter ensembles provide an attractive way to include stochasticity into a parameterization scheme in a physically motivated way. E.g., the **stochastically perturbed parameterization tendencies** (SPPT) scheme, used operationally at ECMWF[1].

$$\frac{\partial X}{\partial t} = D + K + (1+e)\sum_{i=1}^{5} P_i,$$

e is a random variable with specified spatial and temporal structure, P_i are the 5 physics terms in IFS, D is dynamics, K is horizontal diffusion

[1] Actually two stochastic parameterization schemes are used at ECMWF. SPPT (Palmer et al. 2009) uses multiplicative noise to perturb the total parameterized tendencies about the average value that a deterministic scheme represents, thus addressing model uncertainty due to the physical parameterization schemes. The second scheme, stochastic kinetic energy backscatter (SKEB) (Berner et al. 2009), represents a physical process absent from the IFS deterministic parameterization schemes. It uses random streamfunction perturbations to represent upscale kinetic energy transfer, counteracting the kinetic energy loss from excessive dissipation in the numerical integration schemes

SQRT(W'W') 500mb Annual Mean



Fortuna 2.5 versus MERRA (used incurrent seasonal forecast system)

WMP_F120 versus MERRA

WMP_F120 versus MERRA-2

Perturbed Ocean Model

Andrejczuk et al. 2016 : Oceanic Stochastic Parameterizations in a Seasonal Forecast System

"**The oceanic variability induced by the atmospheric forcing of the coupled system is, in most regions, the major source of ensemble spread.**"

Some impact was found "… in regions of strong eddy activity, such as along western boundary currents in the Gulf Stream and Kuroshio regions, in the North Atlantic subpolar region, and also in parts of the Southern Ocean."

"Similar results were found by Juricke et al. (2014) in the context of applying stochastic perturbations to the sea ice strength in seasonal sea ice modeling. They showed that **after a few weeks the atmospheric variability is the largest contributor to sea ice ensemble spread**."

Perturbed Land Model

MacLeod et al. Perturbing hydrology parameters in seasonal forecasts

Conclusions: mixed depending on method

We show here that **perturbing parameters in CY41R1 gives large improvements in terms of soil moisture reliability**.

The **model spread/error ratio is increased with perturbation**. For soil moisture the SP experiments give the largest improvement, however the PP experiment gives an unusually large increase in spread of soil temperature despite only perturbing soil hydrology parameters.

Work at ECMWF now focuses on perturbing the land-atmosphere coupling parameter

MacLeod et al. (2015) Improved seasonal prediction of the 2003 European heatwave through better uncertainty representation in the land surface, QJRMS 142:694 pp 79-90

Conclusions/Summary Concerning Current GEOS-5 Seasonal Forecasting System

- Lead 1 seasonal predictions (1982-2010) of ENSO SST (both cold and warm) are skillful, reliable, have good resolution, and sharpness
  - there is however evidence that the ensemble spread is under-dispersive

- Lead 1 seasonal predictions of T2m (averaged over NH land, 1982-2010), lack sharpness, are overconfident, and are marginally skillful (true for all NMME models)
  - there is some suggestion that increasing the number of ensemble members may be beneficial (CFSv2 and FLOR have 24 members)
  - Near normal category has no resolution, no skill (true for all models)
  - Reinforces need to examine conditional skill (e.g., linked to ENSO, land, etc)

- Lead 1 seasonal predictions of precipitation (averaged over NH land, 1982-2010), lack sharpness, are overconfident, and have no skill (true for all NMME models)
  - Reinforces need to examine conditional skill (e.g., linked to ENSO, land, etc)

- There appear to be substantial differences in the quality of predictions for warmer than normal SST (poor) versus colder than normal SST (better) in the MDR (true for all NMME models – need to understand why)

- Most success to date in dealing with under-dispersive ensembles focused on atmospheric convection schemes (“**oceanic variability induced by the atmospheric forcing of the coupled system is, in most regions, the major source of ensemble spread”**) – though research on perturbing ocean and land appears to be in its infancy.

- Recent versions of our AGCM have much great tropical variability than the AGCM (Fortuna 2.5) used in our current seasonal forecast system, suggesting the next seasonal system should benefit from that to produce greater (more realistic) ensemble spread

216 / Viatcheslav Kharin et al.

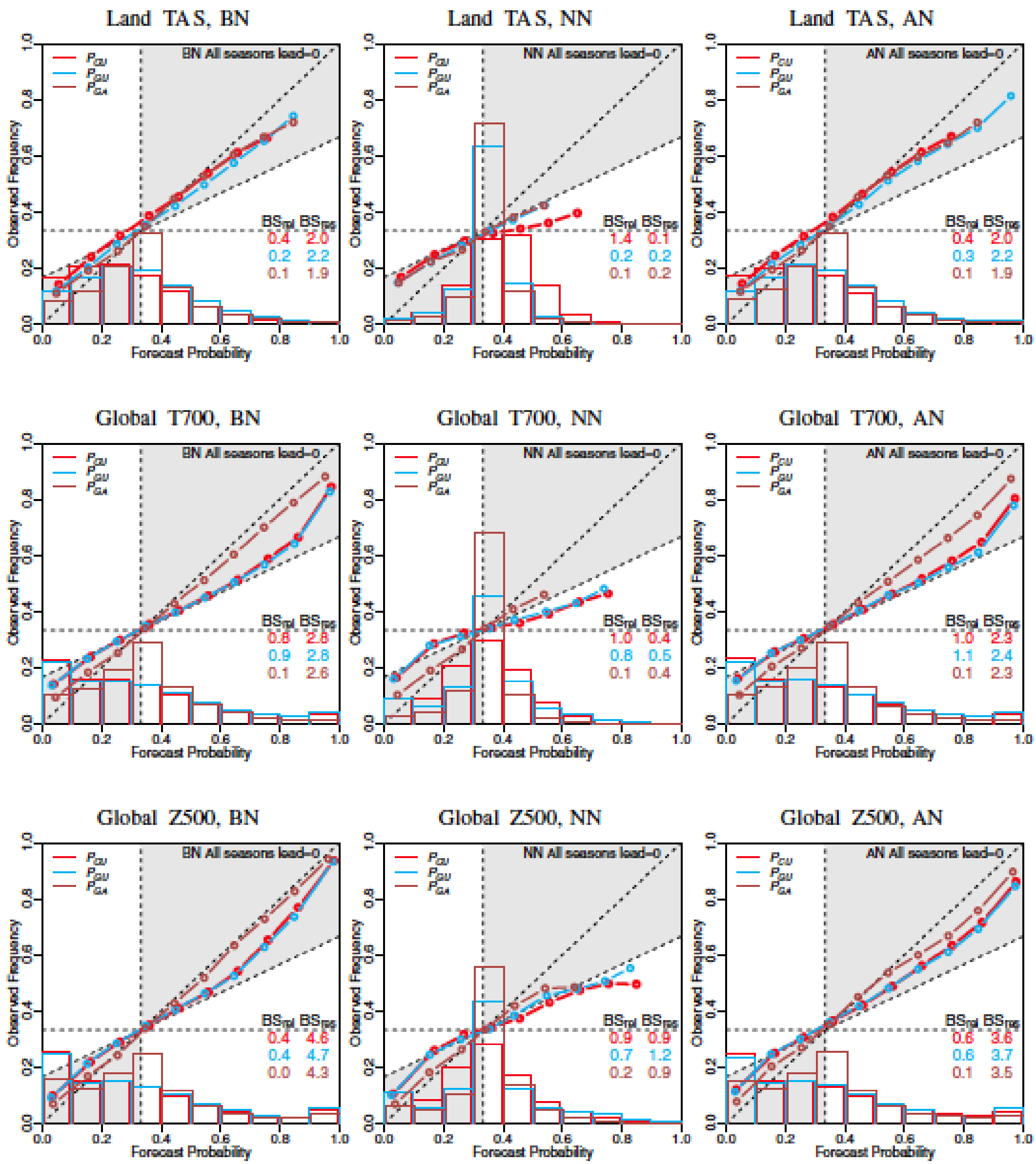


Fig. 7 Reliability diagrams for zero-month lead probabilistic categorical forecasts of global TAS over land (top row), T700 (middle row) and Z500 (bottom row) obtained by the unweighted count method (P_{CU}, red lines), unweighted Gaussian method (P_{GU}, blue lines), and statistically adjusted Gaussian method (P_{GA}, brown lines). All statistics are averaged over all 12 seasons. The columns from left to right are for the below normal, near normal and above normal categories, respectively. The averaged values of the reliability and resolution terms BS_{rel} and BS_{res} (multiplied by 100) are also indicated.

Suggest We Follow This Compact Style of Reliability Diagrams

Includes information on: reliability, resolution, sharpness, and skill including contributions to BSS from each (reliability and resolution) term

Here - compares three different NMME forecasts in one panel

Barbara G. Brown UCAR

Useful references

- **Good overall references** for forecast verification:
  - (1) Wilks, D.S., 2006: *Statistical Methods in the Atmospheric Sciences (2nd Ed).* Academic Press, 627 pp.
  - (2) WMO Verification working group forecast verification web page, http://www.cawcr.gov.au/projects/verification/
  - (3) Jolliffe and Stephenson Book: Jolliffe, I.T., and D.B. Stephenson, 2012: *Forecast Verification. A Practitioner's Guide in Atmospheric Science.*, 2nd Edition, Wiley and Sons Ltd.
- **Verification tutorial – Eumetcal** (http://www.eumetcal.org/-learning-modules-)
- **Rank histograms**: Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550-560.
- **Spread-skill relationships**: Whitaker, J.S., and A. F. Loughe, 1998: The relationship between ensemble spread and ensemble mean skill. *Mon. Wea. Rev.*, **126**, 3292-3302.
- **Brier score, continuous ranked probability score, reliability diagrams**: Wilks text again.
- **Relative operating characteristic**: Harvey, L. O., Jr, and others, 1992: The application of signal detection theory to weather forecasting behavior. *Mon. Wea. Rev.*, **120**, 863-883.
- **Economic value diagrams**:
  - (1)Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Royal Meteor. Soc.*, **126**, 649-667.
  - (2) Zhu, Y, and others, 2002: The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteor. Soc.*, **83**, 73-83.
- **Overestimating skill**: Hamill, T. M., and J. Juras, 2006: Measuring forecast skill: is it real skill or is it the varying climatology? *Quart. J. Royal Meteor. Soc.*, Jan 2007 issue. http://tinyurl.com/kxtct

Useful references continued:

Verification of Ensembles: Barbara G. Brown UCAR, 2015.

Verification of probability and ensemble forecasts: Laurence J. Wilson Atmospheric Science and Technology Branch Environment Canada

Ensemble Forecasting and their Verification: Malaquías Peña, Environmental Modeling Center, NCEP/NOAA , 2014.

Bowler, N.E., Arribas, A., Mylne, K.R., Robertson, K.B. and Beare, S.E., 2008: The MOGREPS short-range ensemble prediction system. QJR Meteorol Soc, 134, 703-722, doi:10.1002/qj.234.

Internet sites with more information:

http://wwwt.emc.ncep.noaa.gov/gmb/ens/index.html

http://www.cawcr.gov.au/projects/verification/#Methods_for_probabilistic_forecasts

http://www.ecmwf.int/newsevents/training/meteorological_presentations/MET_PR.html