# *PROBABILISTIC FORECASTING*
# **WITH OR WITHOUT ENSEMBLES?**

Jie Feng[1, 2], **Zoltan Toth[1]**,
Malaquias Pena[3], and Jing Zhang[4]

[1] NOAA Global Systems Laboratory, Boulder, CO
[1,2] Fudan University, Shanghai, China
[4]Univ. Connecticut
[5]Typhoon Institute

Acknowledgements:
tba

9th NOAA Ensemble User Workshop, Aug. 23, 2023, College Park, MD

# PREAMBLE

- **Ensembles are ubiquitous** in numerical weather prediction
  - Estimated 75% of cpu, and significant developmental efforts devoted to weather & climate ensemble forecasting

- Probabilistic and other types of **products derived from ensembles**
  - Used widely
  - With clear value

# MOTIVATION

- We strongly believe that
  - **Conveying expected forecast skill is crucial**
    - Probabilistic or other formats needed

- The question we pose:
  - **What is the best way to make probabilistic forecasts?**
    - ***With or without ensembles?***

# ATTRIBUTES OF FORECAST PERFORMANCE

- **Two** (and only two) **independent attributes**
  - Resolution (Murphy & Daan 1985), or Informativeness (Krzysztofowicz 1992)
  - Reliability (Murphy & Daan 1985), or Calibration (Krzysztofowicz & Sigrest 1999)

- **Metrics** to assess attributes
  - Information measured by single common metric, irrespective of form of forecast (Toth et al. 2005) = >
    - ***Can compare skill of forecasts in any form***

  - In contrast, metrics of reliability is dependent on form of forecast
    - Reliability of single-, multi-value (ensemble), probabilistic, etc forecasts
      - Necessarily measured by different metrics

- RMSE, MAE, RPS, CRPS, and other **commonly used metrics**
  - Confound 2 independent attributes with various undetermined weights

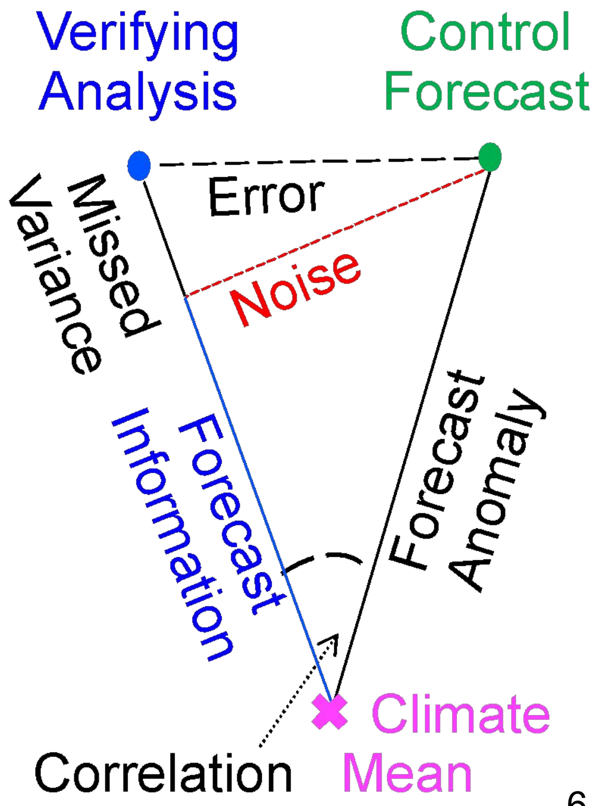
# MAIN OBJECTIVES

- Assess and **compare ensemble performance with that of an unperturbed** (”control”) **forecast**
  - In terms of 2 attributes:
    - Forecast information & statistical reliability
- Explore if ensemble may have **unique virtues?**
  - Missed by two attributes
- **Explain well-known results** w confounded metrics
  - In terms of two attributes

# INFORMATION AND NOISE IN FORECASTS

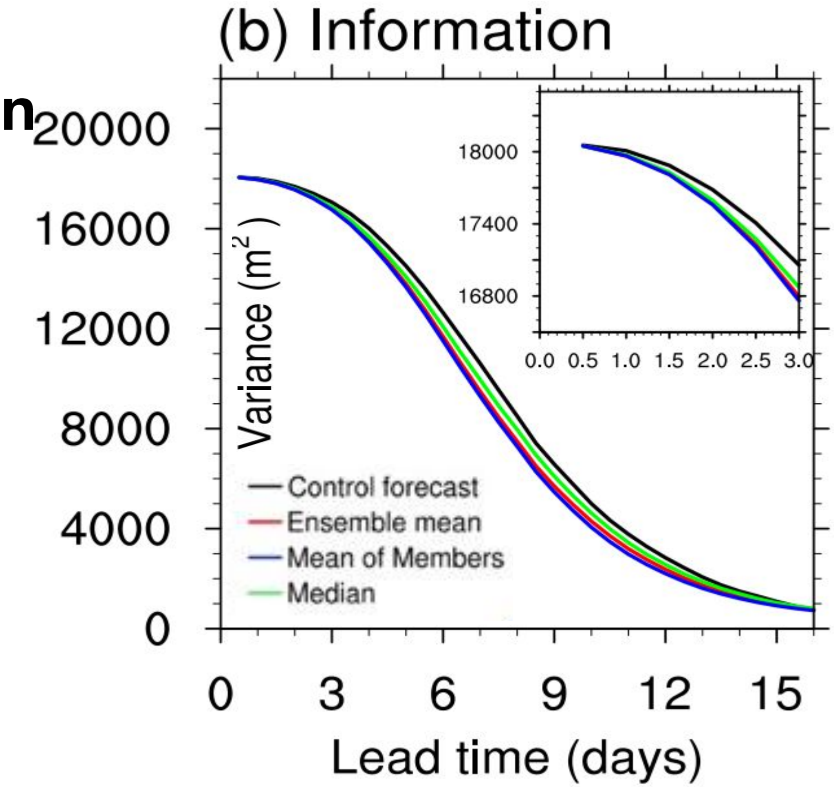
- **Decomposition of forecast anomaly** along & orthogonal to observed anomaly (Toth et al. 2023)
  - Information variance – What matches reality
  - Noise variance – What is different from reality

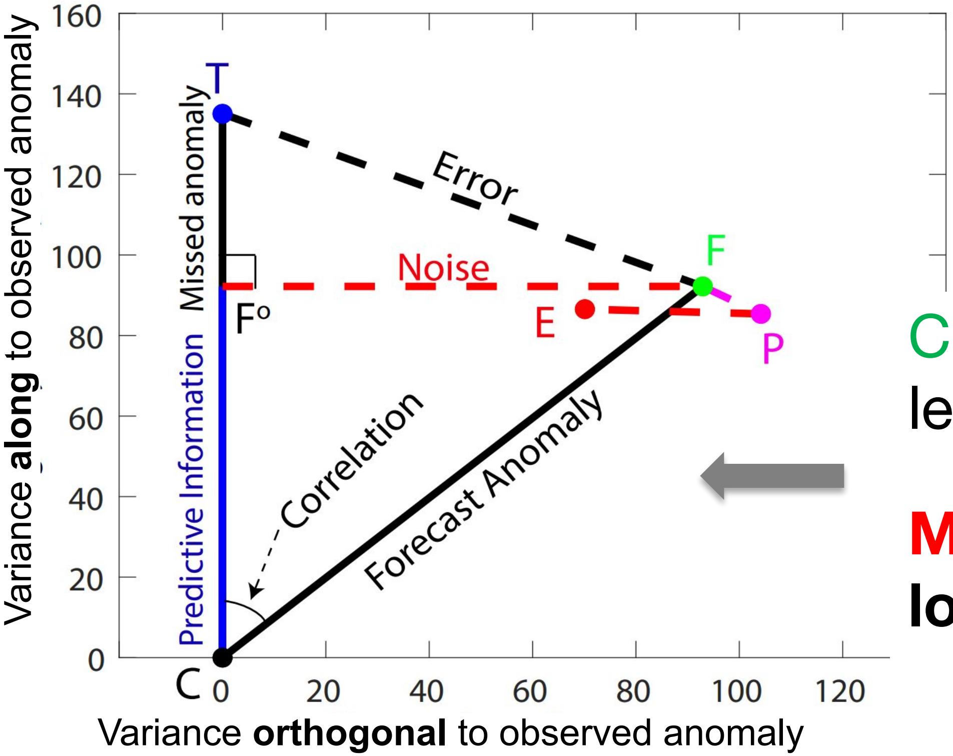
- Information is **direct metric of forecast performance**
  - Only function of what is well predicted
  - Error is also function of variance missed by forecast
- For forecast systems with realistic variance
  - Info equivalent to correlation



# INFORMATION IN CONTROL & ENSEMBLE FCSTS

- Ensemble members and mean / median have **lower information than control** at all lead times
- Forecasts are made for information about future weather
  - Disappointing result





Control has **more information** & less noise than perturbed members

**Mean filters out noise** but **inherits lower information** from members

# OTHER FORECAST ATTRIBUTE - RELIABILITY

- Strong consensus in developer and user communities
  - **Ensembles are unreliable** = >

- **Ensembles & derived products** must be
  - Statistically post-processed before their use

- **Probabilistic and other products can be** readily **generated from single control forecast** via statistical means
  - Eg, Delle Monache et al. 2013

- As **ensembles**
  - Have less information &
  - Need statistical processing anyway
    - No obvious benefit from dynamically generated ensembles?

*Are we missing any unique value from ensembles?*

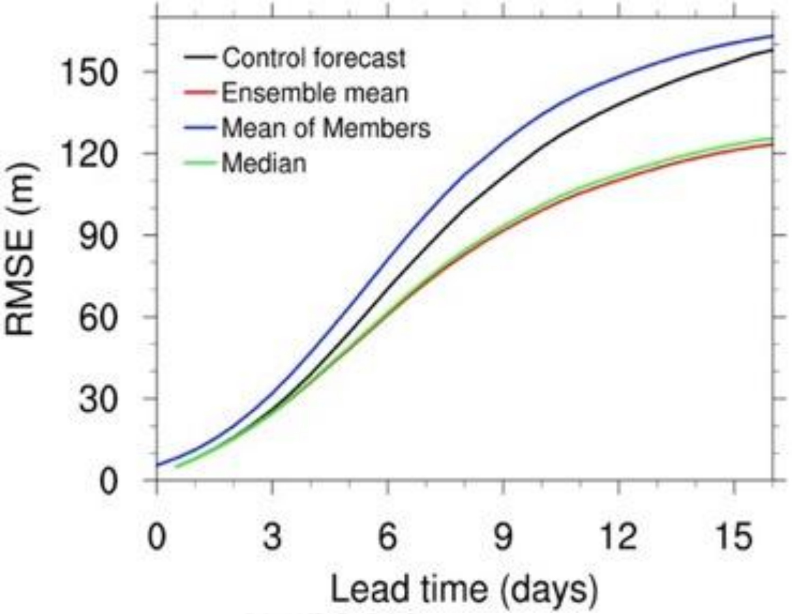
# ENSEMBLE MEAN “MORE ACCURATE”

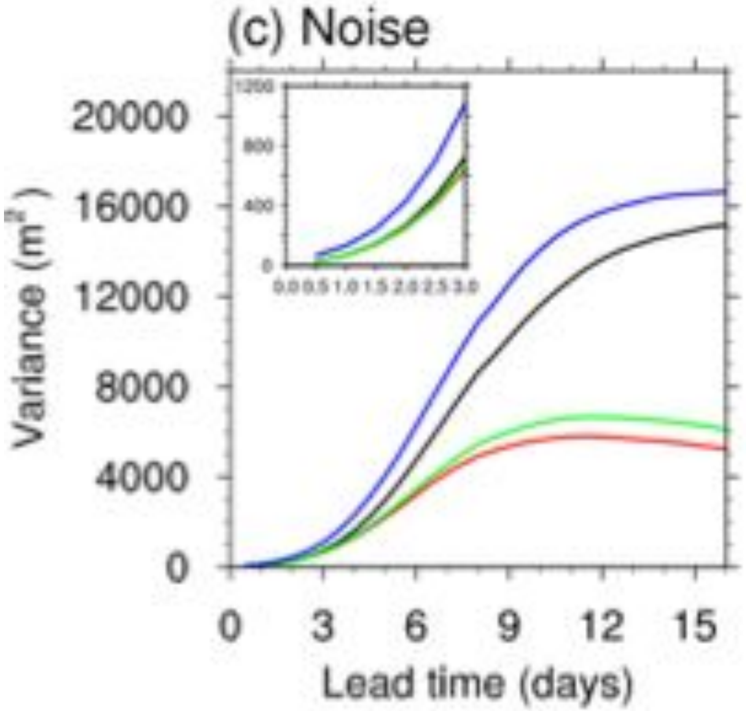
- Yes, **ensemble mean or median have much lower rms error**, MAE, etc

- Error variance =
  1 – Information + Noise variance
  - Mean has more info or less noise or both?

- Somewhat **lower information compensated by** efficient **removal of** nonlinearly saturated perturbations (**noise**) in mean

- We posit that **noise removal** in mean is **conditioned not on “cases” but on scales?**
  - Hence may be reproducible by statistical methods





# CASE-DEPENDENT FLUCTUATIONS

MYTH

- **Widely held belief**
  - It is case-dependent fluctuations in the distribution that make
    - CRPS lower in ensemble- than in control-derived probabilistic forecasts (Roulston and Smith 2003, etc)

- **CRPS is analogous to MAE** – a measure of “accuracy” for probabilistic forecasts
  - As error in mean, CRPS is lower
    - Due to reduced noise in median of distribution
    - And not because, but despite lower forecast information

REALITY

- **CRPS is not** even **affected by**
  - Case dependent fluctuations in shape of distribution
    - It depends only on average of spread over the sample (Hersbach 2000)

# SPREAD – ERROR CORRELATION

- **Correlation low** – Only ~10% variance in error explained (e.g., Hopson 2014)

- **No record of its use** by anyone in literature?

- May be **related to multinormal nature of distribution** of natural and forecast states
  - Toth et al. 1991a,b; Kleeman 2011

# TAKE-HOME MESSAGE

- **Ensembles recreate the same information**, albeit at a somewhat reduced level, already present in the control **N times**, while

- With painstaking accuracy, **dynamically generate N new realizations of noise** different from that in the control

*A dubious enterprise, as far as we can tell*

# WAY OUT?

- **Redirect** computational and developmental **resources to**
  - Increase resolution of unperturbed forecast for
    - Improved performance

- **Use statistical methods to derive** reliable **probabilistic** and any other **products** users need

# BACKGROUND

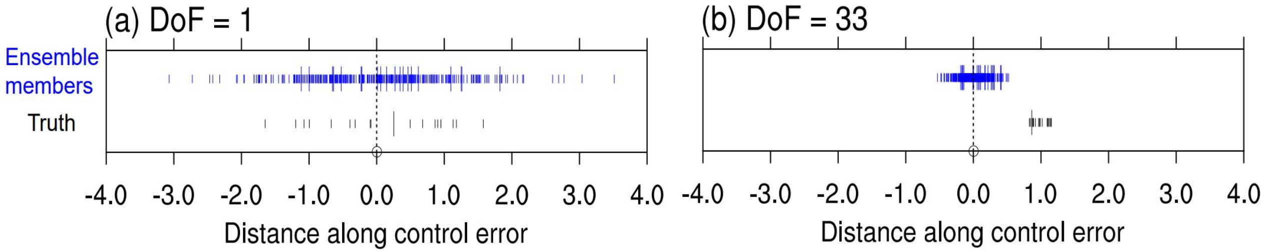
# REFERENCES

- Delle Monache, L., F. Anthony Eckel, Daran L. Rife, Badrinath Nagarajan, and Keith Searight, 2013: Probabilistic weather prediction with an analog ensemble. Mon. Weather Rev., 141 (10), 3498-3516.
- Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather and Forecasting, 15(5), 559–570.
- Hopson, T. M., 2014: Assessing the Ensemble Spread–Error Relationship, Monthly Weather Review, 142(3), 1125-1142. Retrieved Mar 14, 2022, from https://journals.ametsoc.org/view/journals/mwre/142/3/mwr-d-12-00111.1.xml
- Kleeman, R., 2011: Information Theory and Dynamical System Predictability. Entropy 13, no. 3, 612-649. https://doi.org/10.3390/e13030612
- Krzysztofowicz, R., 1992: Bayesian Correlation Score: A Utilitarian Measure of Forecast Skill. Mon. Wea. Rev., 120, 208–220, https://doi.org/10.1175/1520-0493(1992)120<0208:BCSAUM>2.0.CO;2.
- Krzysztofowicz, R., and A. A. Sigrest, 1999: Comparative Verification of Guidance and Local Quantitative Precipitation Forecasts: Calibration Analyses. Wea. Forecasting, 14, 443–454, https://doi.org/10.1175/1520-0434(1999)014<0443:CVOGAL>2.0.CO;2.
- Murphy, A. and H. Daan, 1985: Forecast Evaluation. In: Probability, Statistics, and Decision Making in the Atmospheric Sciences, A. H. Murphy and R. W. Katz, Eds., Westview Press, 379-437.
- Roulston, M. S., & L. A. Smith, 2003: Combining dynamical and statistical ensembles. Tellus, Series A: Dynamic Meteorology and Oceanography, 55(1), 16–30. https://doi.org/10.1034/j.1600-0870.2003.201378.x
- Toth, Z., 1991a: Estimation of Atmospheric Predictability by Circulation Analogs, Monthly Weather Review, 119(1), 65-72. Retrieved Mar 14, 2022, from https://journals.ametsoc.org/view/journals/mwre/119/1/1520-0493_1991_119_0065_eoapbc_2_0_co_2.xml
- Toth, Z., 1991b: Circulation Patterns in Phase Space: A Multinormal Distribution?, Monthly Weather Review, 119(7), 1501-1511. Retrieved Mar 14, 2022, from https://journals.ametsoc.org/view/journals/mwre/119/7/1520-0493_1991_119_1501_cpipsa_2_0_co_2.xml
- Toth, Z., O. Talagrand, and Y. Zhu, 2005: The attributes of forecast systems: A framework for the evaluation and calibration of weather forecasts. Predictability of Weather and Climate, T. N. Palmer and R. Hagedorn, Eds., Cambridge University Press, 584—595.
- Toth, Z., J. Feng, M. Pena, and J. Zhang, 2023: A Foray of Dynamics into the Realm of Statistics: A Review of Ensemble Forecasting. Under review at QJRMS.

# WHAT IS BEHIND?

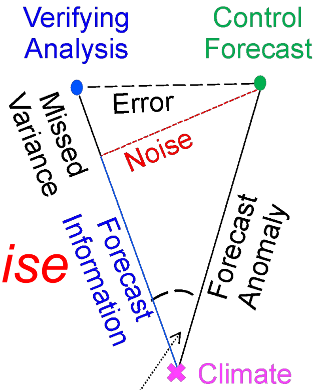
(Expectations about) ensembles based on experience in 1D

- All variability in same, single direction
- Reliable perturbations around state estimate
  - Perturbation variance = Error variance
- Mean of error variance in perturbed states double that in control
  - Yet many (48%) members have error lower that that in control = >
- Reality encompassed by ensemble

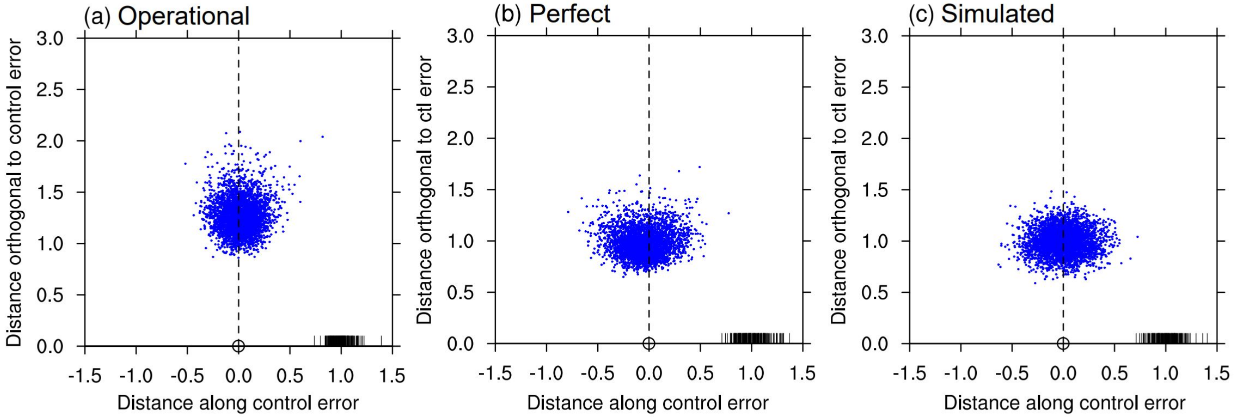


Things work very differently in high dimensional space of atmospheric dynamics

- Variability equally distributed among all directions
  - Unknown error in control lies in a single direction
- Random perturbations have negligible projection on error
  - They lie in directions orthogonal to error = >
- **Perturbations are dynamically irrelevant**, *add only random noise*
  - ***Reality is missed by cloud of ensemble***



# REAL, PERFECT, SIMULATED ENSEMBLES



NCEP Operational 500 hPa height 12-hr ensemble

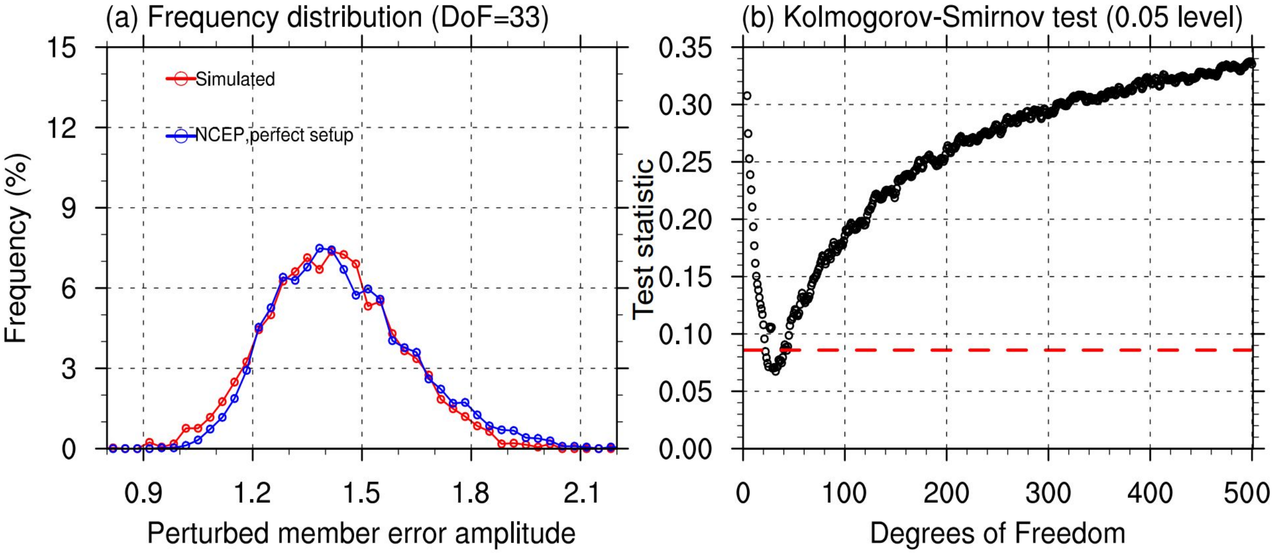
Perfect model/ensemble: 1 member = Truth

Random draws from multinormal distribution

- Ensemble members
  - Blue dots
- Proxy for reality
  - Black bars (180 cases)
- Plotted along (X) & orthogonal to error in control (Y)
  - Both standardized by error in control

- Perturbations are statistically reliable
  - Perturbation variance = Error variance
- Small / large parts of perturbation variance are
  - Dynamically relevant / noise, respectively = >
- Ensemble cloud completely misses reality
  - They constitute randomly reproducible noise

# HOW MANY DEGREES OF FREEDOM (dof)?

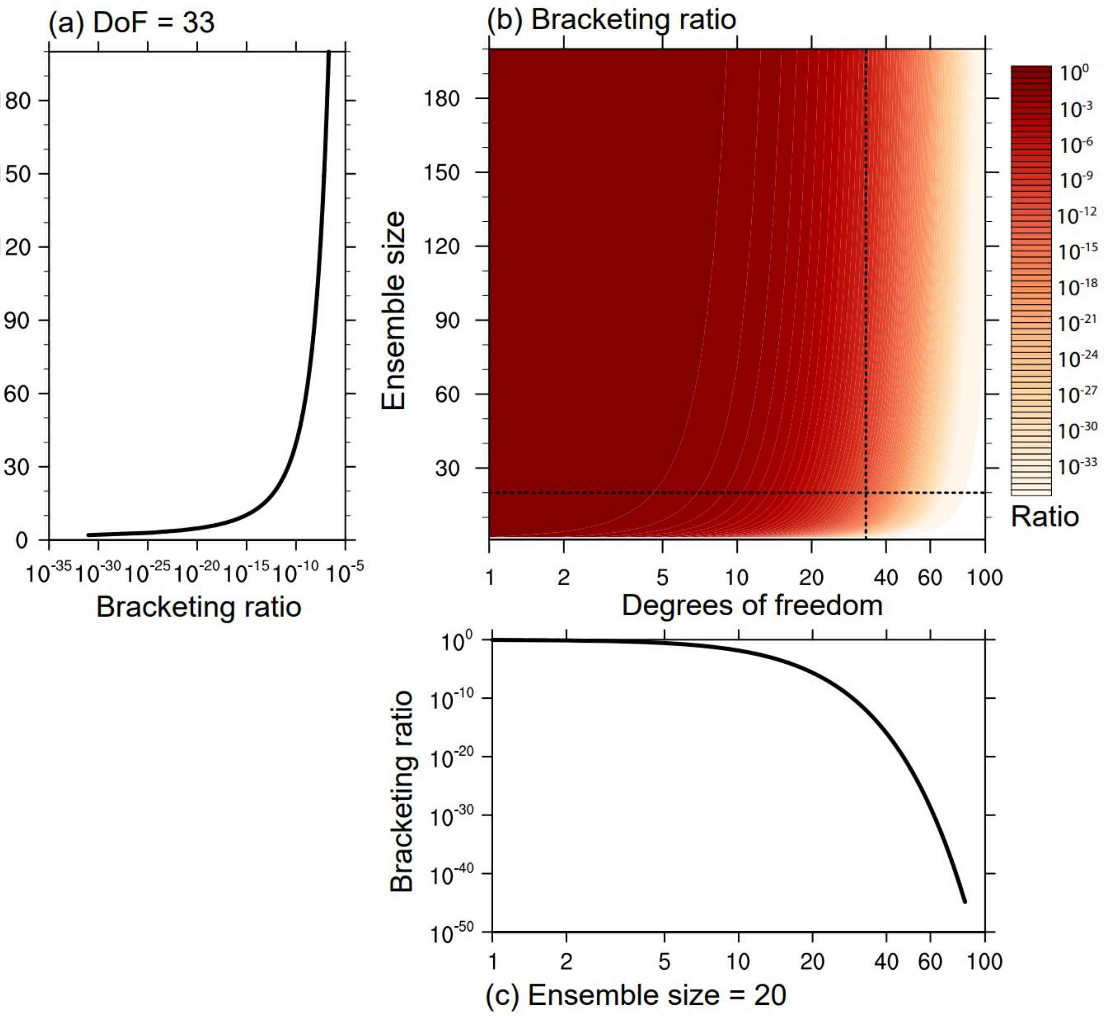


Winter of 2017/18

- Error distribution from perfect 500 hPa height ensemble indistinguishable from
  - Simulation with dof in 28-38 range
    - Dof = 33 gives best fit for NH
    - Dof = 50 for global domain
- Considering variance across all levels & variables
  - Independent dof estimated to be in range of 150-200

# HOW IMPOSSIBLE BRACKETING IS?

Less than $10^{-40}$ chance with realistic-size ensembles



# PROLOGUE

- Ensembles as we know them today have been around for about ~35 yrs

- With Jie Feng, we have worked on this review for ~6 yrs

- I got to summarize in ~12 mins / slides
  - 20 secs for each year of ensembles
  - 2 mins / slides for each year of our study…

# BACKGROUND

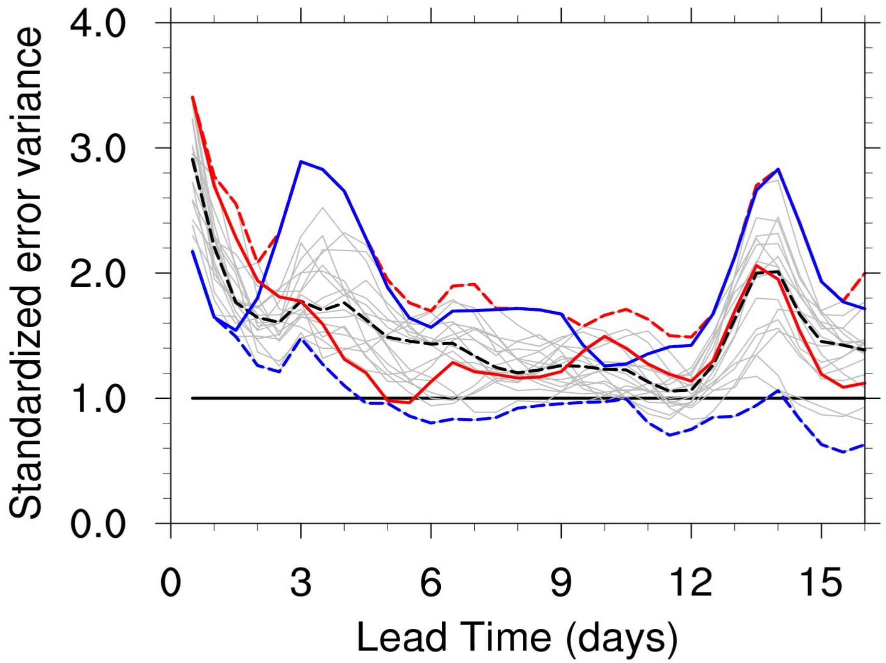


Figure. 11. Same as Fig. 8a, except error variance of individual forecasts against the verifying analysis for the single case initialized at 12 UTC, 30 Dec 2017. The three dashed curves represent the error in the best (bottom, blue), median (middle, black), and worst member (top, red) at each lead time separately. The blue and red solid curves show the error variance in the members best and worst at the 12-hr lead time, respectively. Light grey curves show the error variance of individual members.

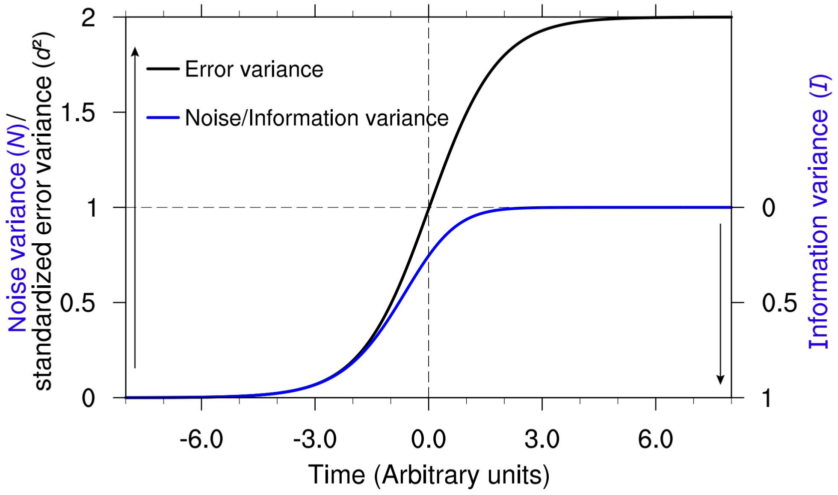
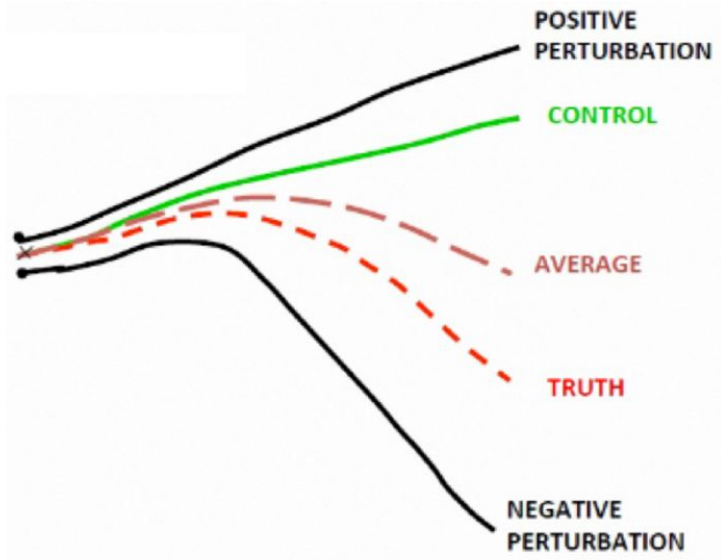


Figure A. Schematic depicting the growth of noise (blue line, left axis) and the decrease of information variance (blue line, right axis) in a forecast characterized by logistically growing standardized error (black line). For further details, see text.



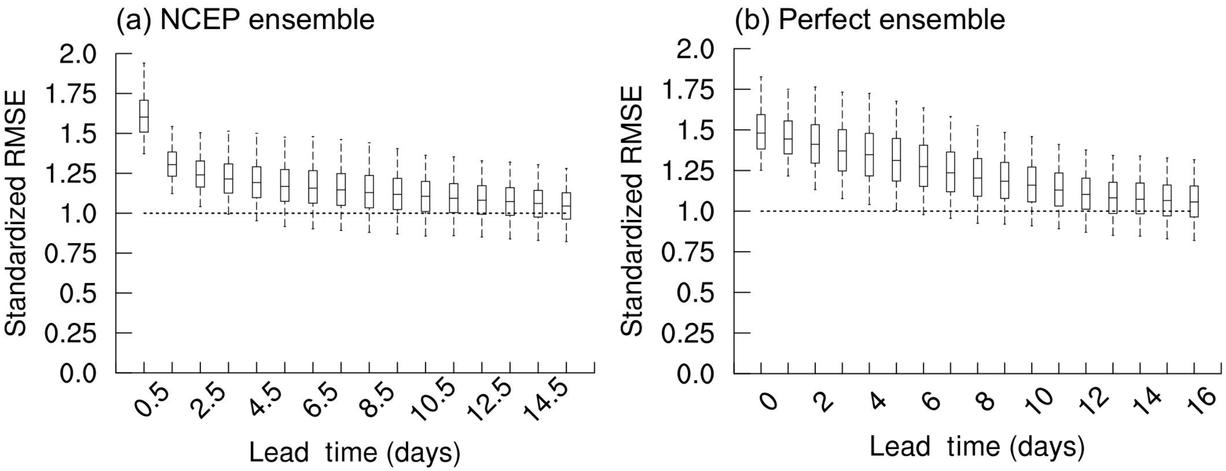
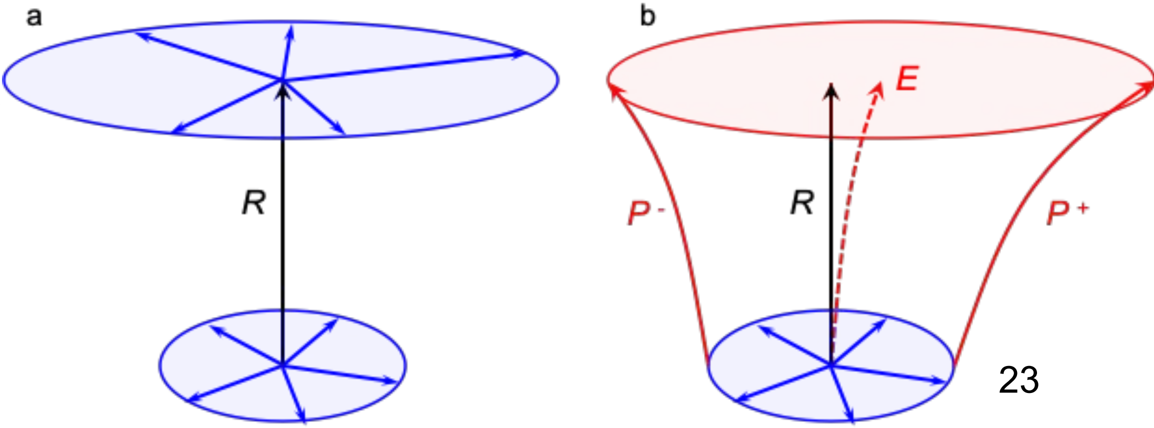


Figure 8. NH (30º -65ºN) 500 hPa height perturbed forecast rms error evaluated against the verifying analysis (a) and a randomly selected member (b), standardized by the error in 0.5 - 15.5 (panel a) and 0 - 16 day (panel b) control forecasts, ranked from lowest to highest, and averaged over all 180 cases. The top and bottom of whiskers and boxes represent the average of the extreme sample point and 25 / 75% quantile values of the 20 and 19 ranked perturbed forecast error values in panels (a) and (b), respectively.

Figure 1. Schematic of statistical (a) vs. dynamical (b) forecast perturbation generation. In either case, initial perturbations (bottom ellipsoids) are centered on a reference initial condition (R, that can be either the truth in an ideal, or the control analysis and forecast in a realistic ensemble, vertical black line). Forecast perturbations (top ellipsoids) are either statistically added and centered on R (a, blue arrows), or generated via the numerical integration of a dynamical model from perturbed initial conditions (b, red arrows). $P^-$, $P^+$(red solid), and E (red dashed) represent two perturbations initially symmetric around, but later off-center of R, and the mean of the ensemble, respectively. For further explanation, see text.



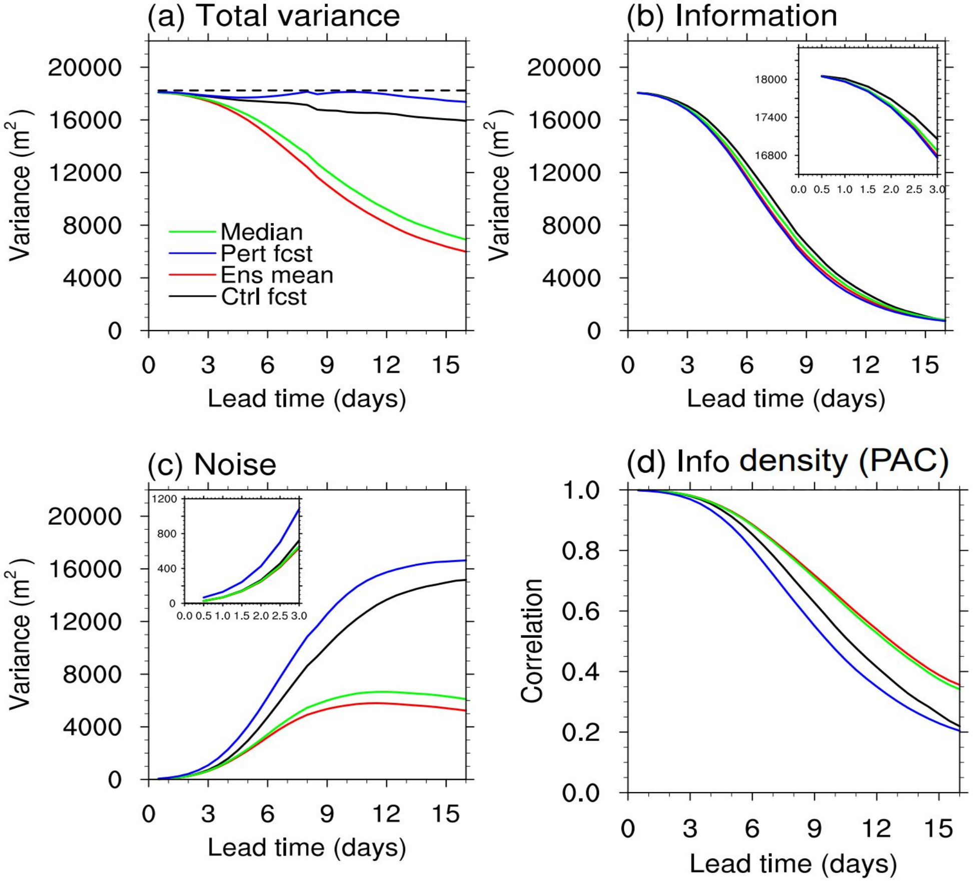
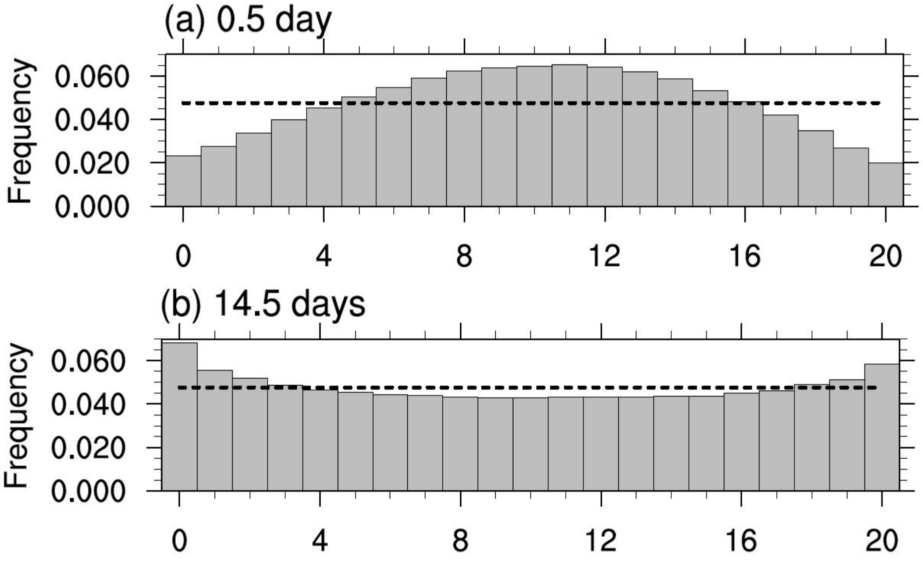


Figure. 3. Sample mean non-standardized (a) total variance, (b) information variance, (c) noise variance, and (d) information density (or pattern anomaly correlation) of 500-hPa geopotential height forecasts in the NH extratropics (30º - 65ºN). The dashed line in panel (a) indicates the climatic variance present in the analysis.

Figure 5. Talagrand (or analysis rank) diagram indicating the frequency of the verifying analysis falling into the intervals defined by the 20 ranked values of 500 hPa geopotential height ensemble member forecasts at individual grid-points, aggregated over the NH extratropics (30º -65ºN) over the 3-month experimental period, at 0.5 (a) and 14.5 days lead times (b). A flat distribution (dashed horizontal lines) indicates a perfectly reliable ensemble (where forecast probabilities of events exactly match their observed frequencies).


(a) 0.5 day
Frequency
(b) 14.5 days
Frequency