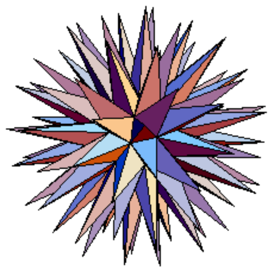
The ~~curses~~ Blessings of high Dimensionality: Explaining Ensemble Behavior

Bo Christiansen

Danish Meteorological Institute

EGU, 8 May, 2020

Christiansen, 2020, J. Clim., in review.
Christiansen, Mon. Weather Rev, 2019, doi:10.1175/MWR-D-18-0211.1
Christiansen, J. Clim., 2018, doi:10.1175/JCLI-D-17-0197.1



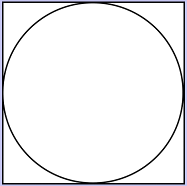
Outline

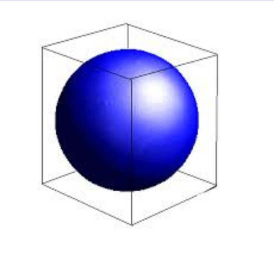
- The challenge:
  - ⋆ For a large range of ensemble forecasts/scenarios the ensemble mean is very often better than individual ensemble members.
  - ⋆ Often the relative error of the ensemble mean is 30 % better than the relative errors of the individual ensemble members.
  - ⋆ The decay as function of ensemble size, of e.g. MSE with respect to observations, is often very regular and can often be predicted.
- The explanation:
  - ⋆ Simple properties of high dimensional spaces.
- Then:
  - ⋆ There is strong evidence in the literature for that multi-model ensembles (CMIP) can be described as indistinguishable from observations.
  - ⋆ We note a transition from the indistinguishable to the truth-centered situation when spatial and temporal scales increase.
  - ⋆ We argue that this is an effect of tuning/calibration on the largest scales.

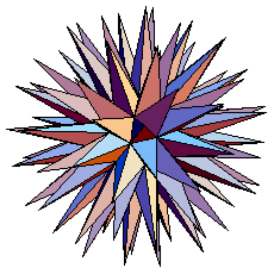
Cube in N dimensions

- Corners of a unit cube in $[-1/2, 1/2]^N$ are $[\pm 1/2, \pm 1/2, \ldots \pm 1/2]$.
- The number of vertices is 2^N.
- Length of the vertices $\sqrt{N}/2$
- Volume of inscribed sphere $\pi^{N/2} r^N / \Gamma(N/2+1)$ with $r = 1/2$.
- Fraction of volume within ϵ from edge $(r^N - (r-\epsilon)^N)/r^N$.

N	Volume	# vertices	Length of vertices	Volume of inscribed sphere	Fraction of volume within 0.05 from edge
2	1	4	0.707	0.785	0.0975
3	1	8	0.866	0.524	0.1426
5	1	32	1.118	0.164	0.2262
10	1	1024	1.581	0.00249	0.4013
25	1	$3.35\ 10^7$	2.500	$2.85\ 10^{-11}$	0.7226
50	1	$1.13\ 10^{15}$	3.535	$1.54\ 10^{-28}$	0.9231
100	1	$1.27\ 10^{30}$	5.000	$1.87\ 10^{-70}$	0.9941

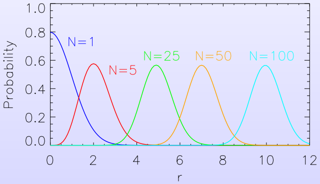






In high dimensions ...

.. all vectors from same distribution are almost always of the same length



The surface area of a hyper-sphere with radius r in N dimensions is $S_{N-1} = 2\pi^{N/2} r^{N-1} / \Gamma(\frac{N}{2})$
The standard Gaussian probability distribution is $P(\mathbf{x}) = (2\pi)^{-N/2} \exp(-\sum_{n=1}^{N} x_n^2/2)$.
So as function of r:

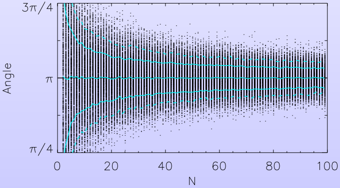
$$P(r) = S_{N-1} P(\mathbf{x}) = \frac{2^{1-N/2}}{\Gamma(\frac{N}{2})} r^{N-1} \exp(-r^2/2).$$

The maximum of $P(r)$ is reached for $r = \sqrt{N-1}$ and the width (standard deviation) of the peak is $\sqrt{2}/2$, independent of N.

Concentration of measures

.. two random vectors are almost always orthogonal

The angle between two vectors $\cos(\theta) = \mathbf{x} \bullet \mathbf{y} / ||\mathbf{x}|| / ||\mathbf{y}||$.



The angles between random pairs of points drawn on the surface of an N-dimensional hyper-sphere.

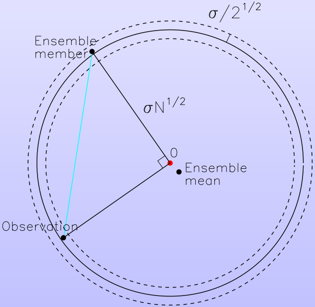
Waist concentration

Why is the relative error of the model mean almost always -0.3?

Assume models and observations drawn from same distribution, e.g. $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$

- Individual models and the observations are orthogonal and at the same distance from center.
- The model mean is at the center.
- Therefore the model mean, the observation, and an individual model forms an isosceles (two legs equal) right triangle.
- Thus, if the error of the model mean is ϵ then the error of the individual models are $\sqrt{2}\epsilon$.
- Relative error of model mean:

$$\frac{\epsilon - \sqrt{2}\epsilon}{\sqrt{2}\epsilon} = \frac{1 - \sqrt{2}}{\sqrt{2}} \approx -0.29$$



Extending the model

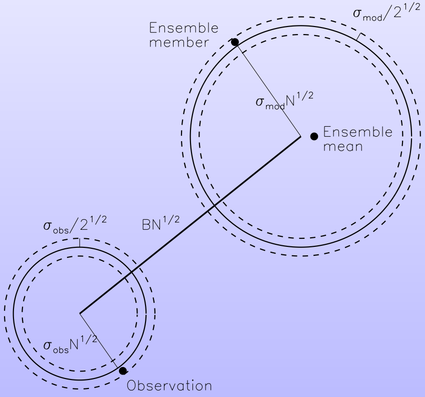
Assume that ensemble members are drawn from $\mathcal{N}(\mathbf{b}, \sigma^2_{mod}\mathbf{I})$ and that observations are drawn from $\mathcal{N}(\mathbf{0}, \sigma^2_{obs}\mathbf{I})$.

Mean square error between observations and ensemble members:

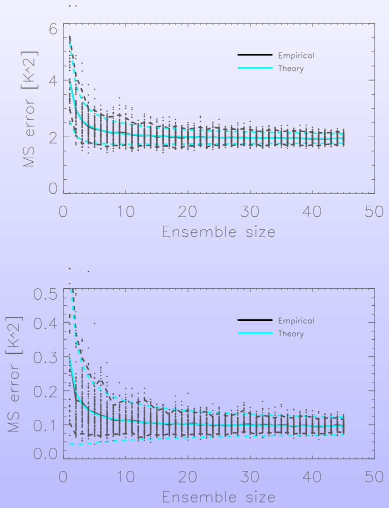
$$||\mathbf{o} - \mathbf{x}^k||^2/N \simeq B^2 + \sigma^2_{obs} + \sigma^2_{mod}.$$

The ensemble mean is special as it will – because of the law of large numbers – be situated near the center of the annulus:

$$||\mathbf{o} - \overline{\mathbf{x}}||^2/N \simeq B^2 + \sigma^2_{obs} + \sigma^2_{mod}/K.$$

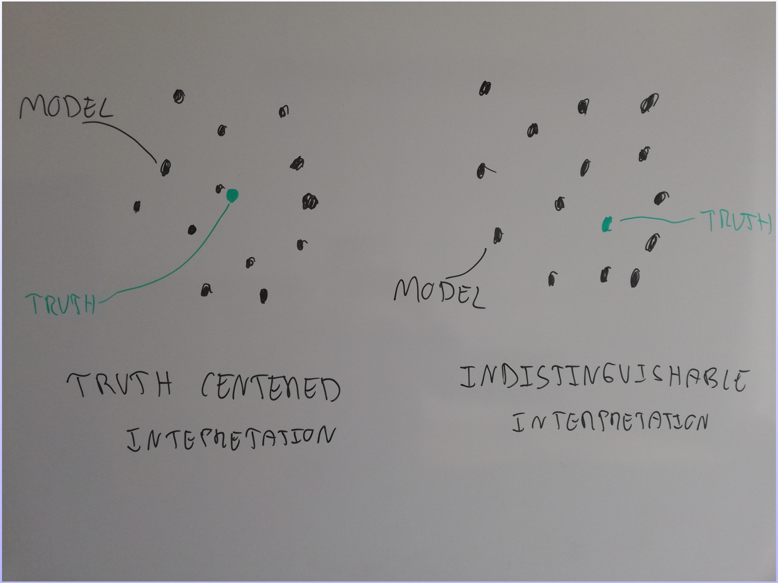


The CMIP5 ensemble: monthly climatology

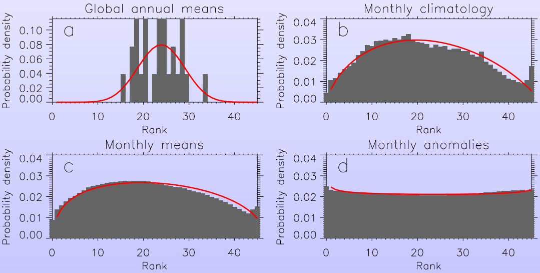


- The normalized mean-square error of the ensemble mean for the CMIP5 models as function of ensemble size. Calculated over all grid-points.
- For each ensemble size k we randomly draw k models (out of the full ensemble of size $K = 45$) with replacement and calculate the mean-square error of the ensemble mean of this sub-ensemble. This is done 100 times for each k. Black dots are the 100 individual errors while full black curves are mean of the errors over the 100 draws.
- Cyan curves are values from simple statistical model from previous slide (yes, we can also get the error-bars).
- Based on TAS for 1980-2005.

Which kind of ensemble?

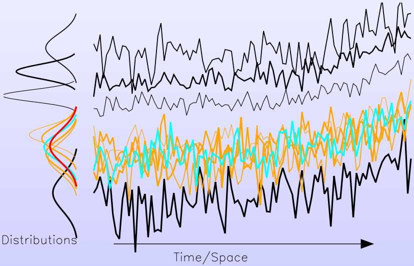


Transition from truth centered to indistinguishable situation



- Rank histograms showing the distribution of the rank of observations among the 45 CMIP5 models. Red curves are from the fitted simple statistical model. a: annual global means, b: monthly climatology, c: monthly means, and d: monthly anomalies. In b, c and d all grid-points are pooled. Based on TAS for 1980-2005.
- Note the transition from an approximate truth centered situation when global annual means are considered to an indistinguishable interpretation when monthly grid-point anomalies are considered.

Schematic view of effect of calibration on large scales



- Cyan curve is the observations, black curves are different uncalibrated models, and orange curves the calibrated models.
- The uncalibrated models have considerably different means and variances but the calibration makes ensemble means and model variances to be more closely centered around the observed values with small additive errors in both.
- The individual values (spatial or temporal) of the multi-model ensemble will after the calibration (orange) be distributed almost as the observation (cyan) but with inflated variance.

Conclusions/Talking points

- The properties of high dimensional spaces often defy our intuition based on two and three dimensions. In high dimensions random vectors drawn from the same distribution have almost always the same lengths. Independent vectors in high dimensions are almost always orthogonal.
- We work in high dimensional space when we calculate the mean square error or correlation over an extended region or period.
- These properties explain why the ensemble mean is almost always better than individual ensemble members and why the ensemble mean is often 30 % better than the mean error of individual ensemble members.
- They also explain how, e.g., the error decays with increasing ensemble members. Can be used to estimate effect of increasing number of ensemble members.
- We saw that the multi-model CMIP5 ensemble goes from the indistinguishable to the truth-centered situation when spatial and temporal scales increase. We explain this by models being calibrated/tuned/trained at the largest scales.