Data, Inference & Applied Machine Learning

Course: 18-785

Patrick McSharry
patrick@mcsharry.net
www.mcsharry.net
Twitter: @patrickmcsharry

Fall 2022

ICT Center of Excellence
Carnegie Mellon University

Course outline

Week	Description
1	Regression
2	Linear models
3	Nonlinear models
4	Unsupervised learning
5	Supervised learning
6	Ensemble approaches

Applied Machine Learning

WEEK 12A

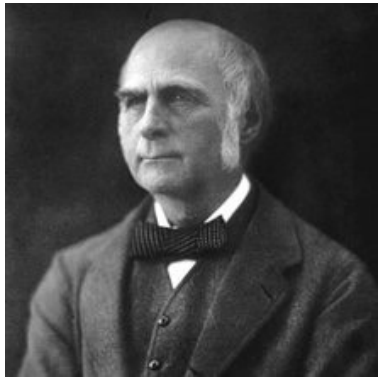
Today's Lecture

No.	Activity	Description	Time
1	Challenge	Coping with uncertainty	10
2	Discussion	Wisdom of the crowd	10
3	Case study	Ensemble forecasts	10
4	Analysis	Bootstrap	20
5	Demo	Bootstrap techniques	20
6	Q&A	Questions and feedback	10

All models are wrong!

- “All models are wrong, but some models are useful” - George P. E. Box.
- “Whatever can go wrong, will go wrong” - Murphy’s Law (Sod’s Law).
- All real-world systems are nonlinear and non-stationary: structural breaks, external shocks.
- Any honest modeller should account for the fact that there is substantial uncertainty.

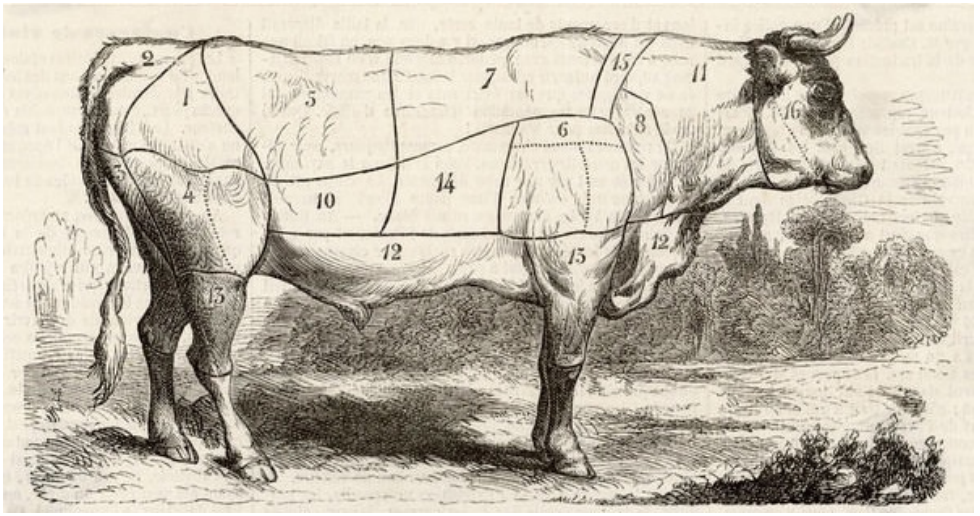
Francis Galton (1822-1911)



- Sir Francis Galton was an English polymath who made important contributions in many fields of science.
- These included:
  - meteorology (first popular weather map in 01-Apr-1975);
  - statistics (regression and correlation);
  - biology (the nature and mechanism of heredity); and
  - criminology (fingerprints).
- Much of this was influenced by his penchant for counting and measuring.
- In 1906, Sir Francis Galton was at a country fair and was interested in the “Guess the weight of the ox” competition.
- He analysed the 787 guesses.

Guess the weight of the ox

- Galton found that the average of the 787 guesses was within how many percent of the true answer:
  a) 10%
  b) 1%
  c) 0.1%
  d) 0.01%
- **slido.com**
- **#19099**



Crowd sourcing

- Francis Galton was a convinced eugenicist and he was searching for evidence of the populace's stupidity.
- Much to his disappointment, he found the average estimate to be off by only 0.1 percent (actual weight was 1,198 pounds and average of 787 guesses was 1,197 pounds).
- "The result," he wrote ruefully, "seems more creditable to the trustworthiness of a democratic judgement than might have been expected."

F. Galton (1907). Vox populi. Nature 75: 450–451.

Wisdom of the crowd

- Many opinions versus expert opinion
- Perhaps wisdom of the crowd is better than an expert?
- Who Wants to Be a Millionaire? Ask the audience (correct answer 91% of the time).
- Wikipedia – crowdsourcing knowledge.
- The wisdom of crowds - James Surowiecki.

Independence

- Crowds are wise only under very special conditions.
- It is important that its members act in complete independence of one another.
- The moment individual decisions interact, we witness the stampedes and crazes usually associated with crowd behaviour.
- The stock market provides a classic illustration. An individual's decision to buy or sell stock depends not only on his estimate of the company's value, but also on his/her estimate of everyone else's estimate.

Mutually reinforcing decisions

- In Keynes's famous analogy, investment is like a beauty contest in which the prize is awarded "to the competitor whose choice most nearly corresponds to the average preferences of the competitors as a whole".
- This fact accounts for the vulnerability of financial markets to boom and bust.
- Other examples of crowd irrationality, such as rioting or lynching, can similarly be explained as the outcome of a series of mutually reinforcing decisions.

Forecast combination

- Suppose we have N forecasts, f_n with n=1,…,N, from separate sources.
- We can construct a forecast combination which is a hybrid given by pooling the information from the individual forecasts
- The forecast combination, f_c, is given as a weighted sum of the individual forecasts
- DeMenezes et al. (2000) describe seven methods for forecast combination.

Simple Average

- The simple average forecast combination is given by

  $f_c = (f_1 + \ldots f_N)/N$

- This approach makes sense intuitively if one has no other information to inform about the relative merits of the individual forecasts.
- It is impartial, robust and has a good ``track-record'' (see Clemen's review of 1989).

Outperformance

- The forecast combination is $f_c = p'f$ where p is a simplex of probabilities which can be assessed and revised in a Bayesian manner.
- Each individual weight is interpreted as the probability that its respective forecast will perform the best (in the smallest absolute error sense) on the next occasion. Each probability is estimated as the fraction of occurrences in which its respective forecasting model has performed the best in the past.
- It is a robust nonparametric method of achieving differential weights with intuitive meaning which performs well when there is relatively little past data and/or when the decision maker wishes to incorporate expert judgement into the combining weights (Bunn, 1985)

Optimal

- Bates and Granger (1969) proposed estimating linear weights to minimise the error variance of the combination (assuming unbiasedness for each individual forecast).
- The weights, w, are determined according to the formula

  $w = S^{-1}e/e'S^{-1}e,$
- where e is the (N x 1) unit vector and S is the (N x N) covariance matrix of forecast errors.

Optimal (independence)

- Optimal approach where the individual forecasts are assumed to be independent
- The estimate of the covariance matrix S in is restricted to be diagonal, comprising just the individual forecast error variances.
- Therefore the off-diagonal covariances are assumed to be equal to zero

Optimal – restricted weights

- Optimal approach where weights are estimated adaptively
- In this case the optimal formula has the additional restriction that no individual weight can be outside the interval [0,1].

Regression

- The regression forecast combination is

  $f_c = w_0 + w_1 f_1 + \ldots w_N f_N$

- In this method the constituent forecasts are used as explanatory variables in an ordinary least squares (OLS) regression with the inclusion of a constant.
- Granger and Ramanathan (1984) argued that this has the advantage over the popular optimal method that an unbiased combined forecast is produced regardless of whether the constituent forecasts are biased.

Regression (restricted weights)

- Regression with restricted weights:
- Ordinary least squares regression is performed with the inclusion of a constant but the weights are constrained to sum to one.
- The constraint maintains the sense of an average.
- The constant removes any existing bias.

Simple robust averages

- Jose and Winkler (2008) suggest using simple robust averages as a means of combining forecasts
  - Trimmed means
  - Winsorized means
- Moderate trimming of 10–30% or Winsorizing of 15–45% improves results
- More trimming or Winsorizing being indicated when there is more variability between individual forecasts

Uncertainty in modelling

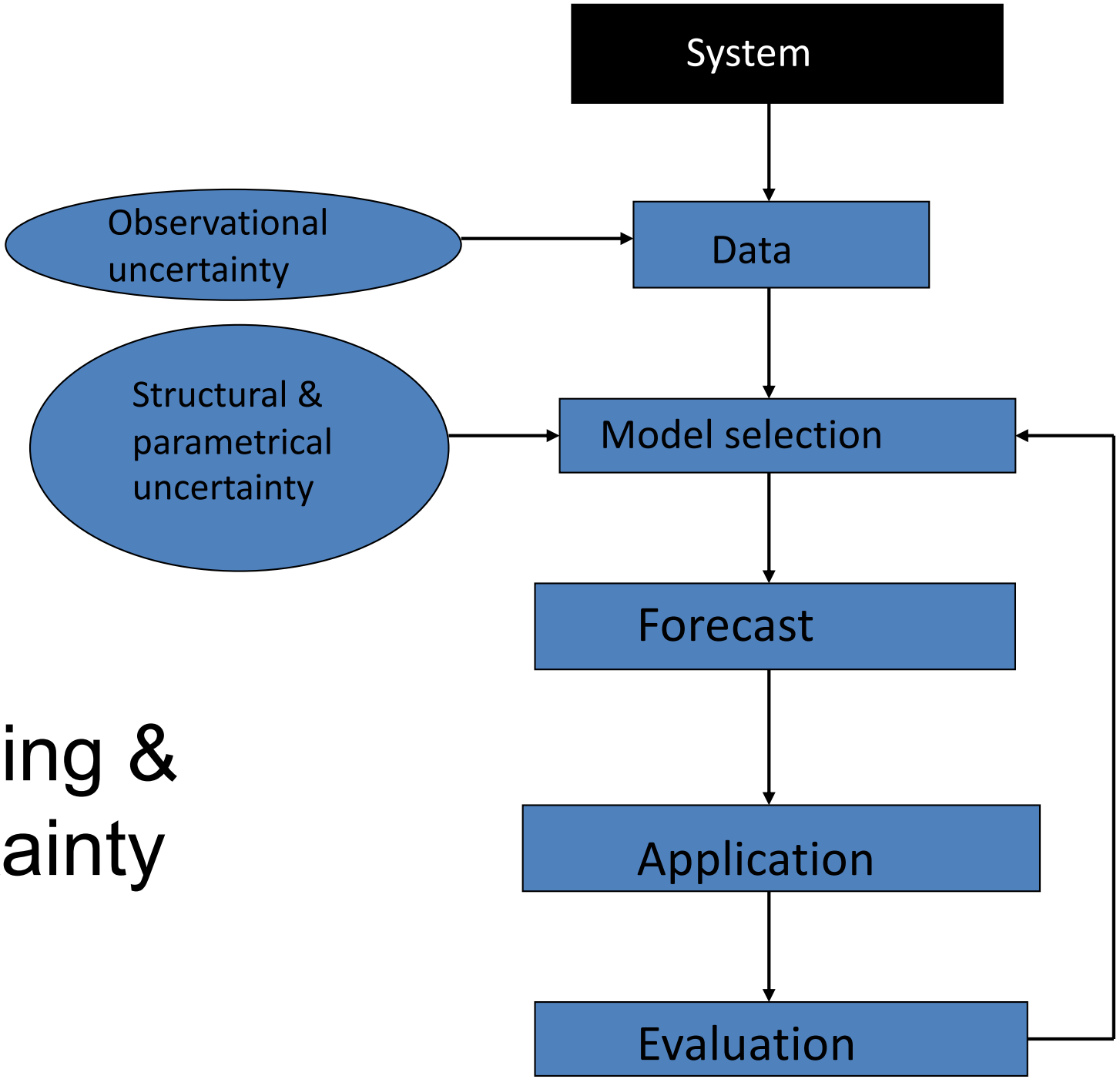
- Where might uncertainty come from when constructing a model?

- **slido.com**
- **#19099**

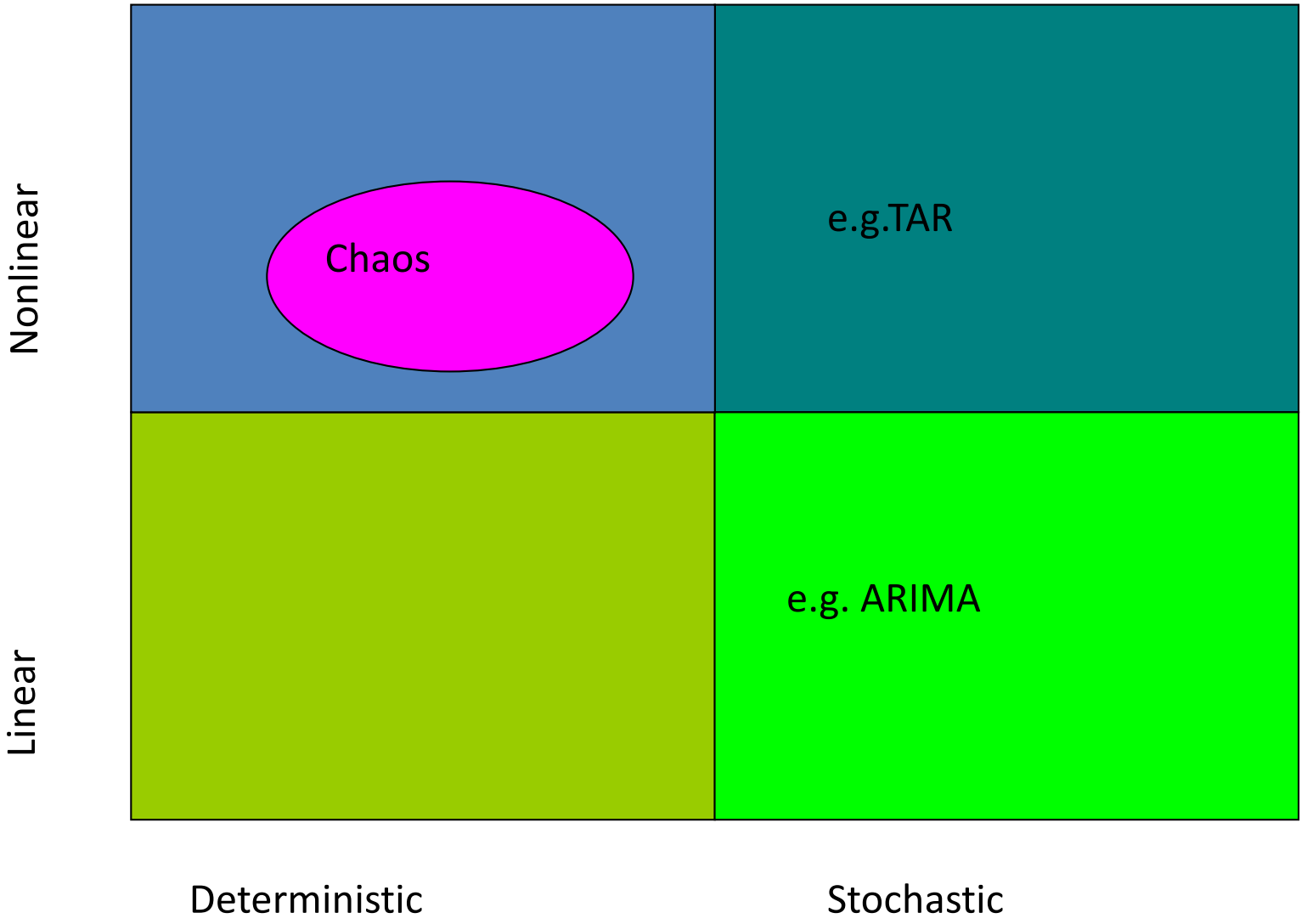
Uncertainty

- Sources of uncertainty:
  - Observational uncertainty
  - Parametrical uncertainty
  - Structural uncertainty
- Ensembles provide a way of accounting for uncertainty
- Multiple initial conditions
- Multiple models

Modelling & uncertainty

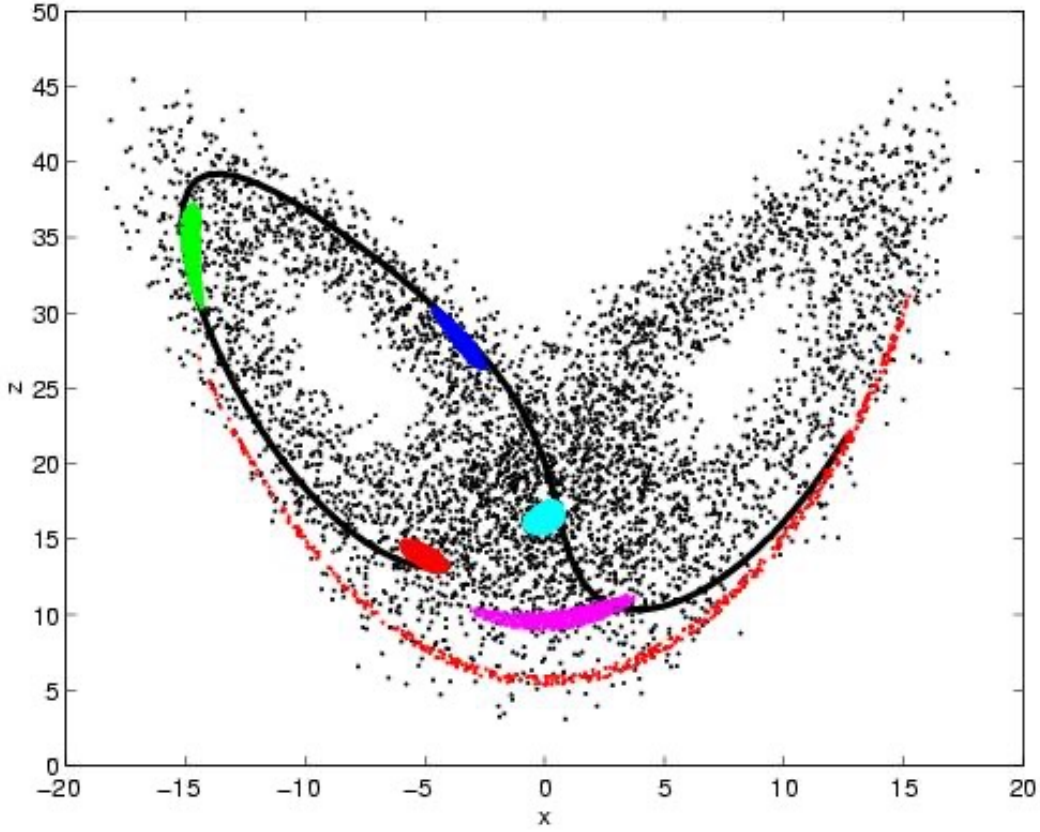


Model class and complexity

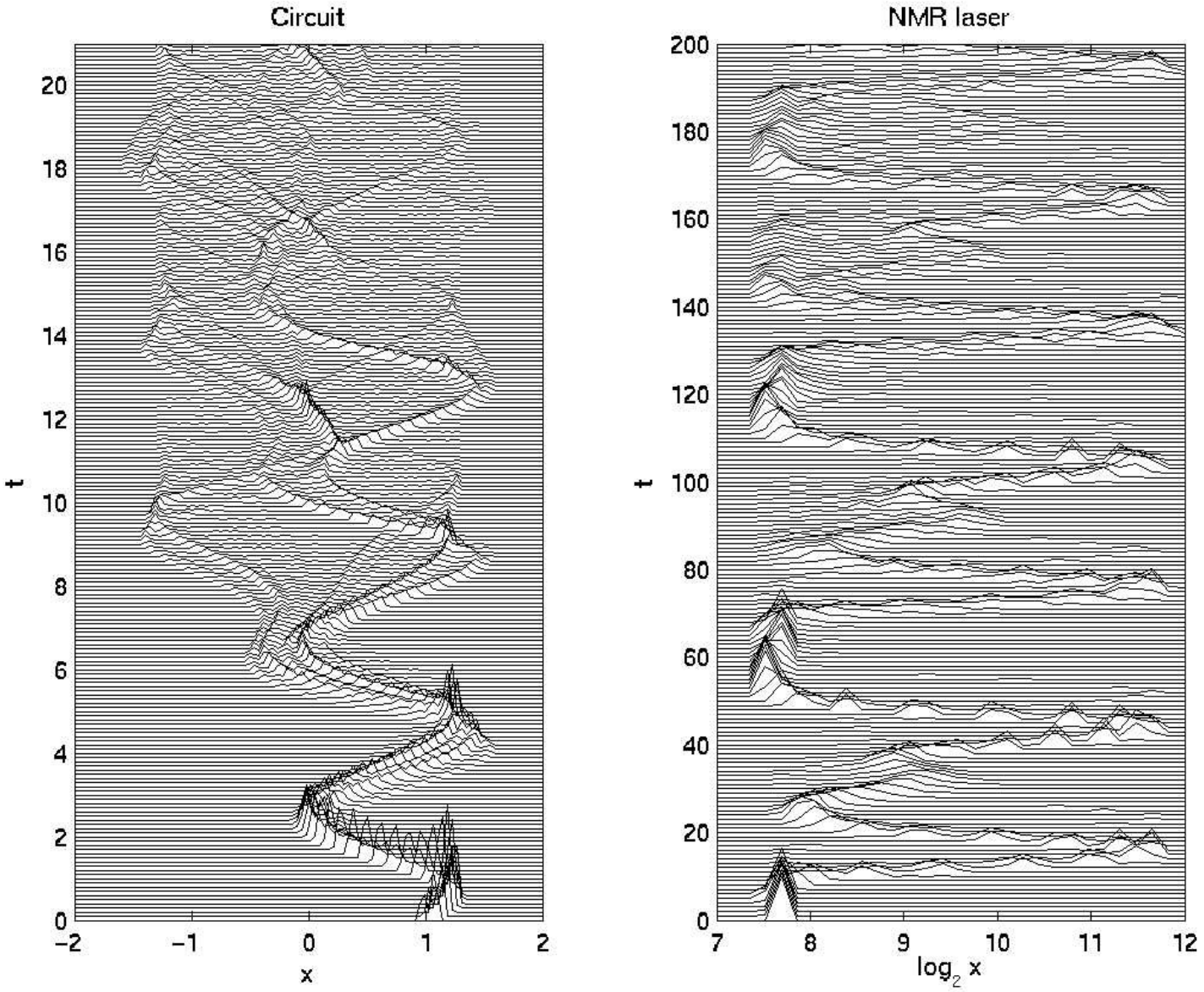


Chaos in the Lorenz system

- Butterfly effect
- Sensitivity to initial condition
- Initial uncertainty grows with time
- Predictability varies with position

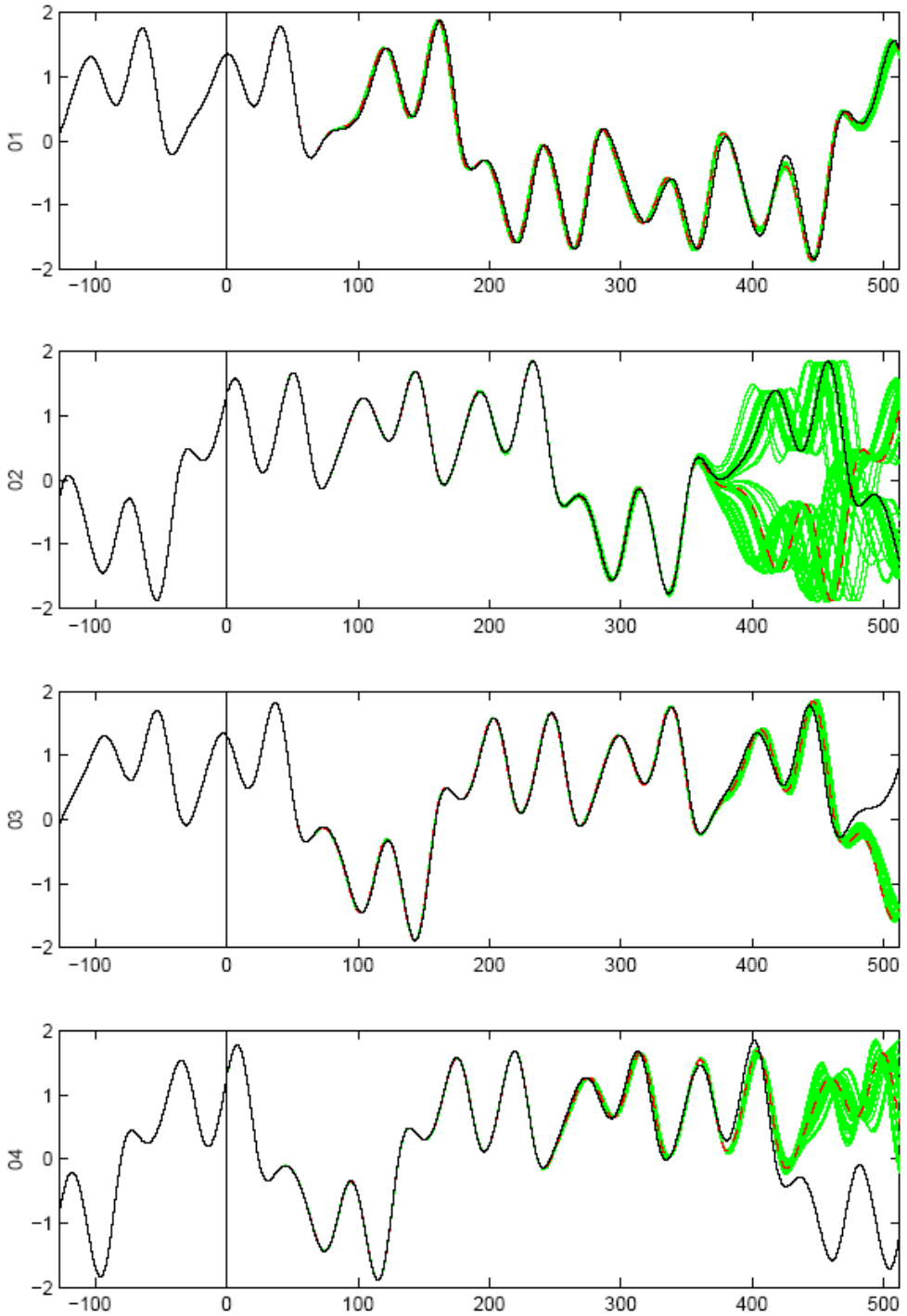


Uncertainty evolution



- Ensemble forecasts
- 128 points for each initial condition
- Forecast lead time of 512 steps
- Observations (black)
- Point forecast (red)
- Ensemble (green)
- Forecast quality depends on the initial condition
- Model inadequacy test

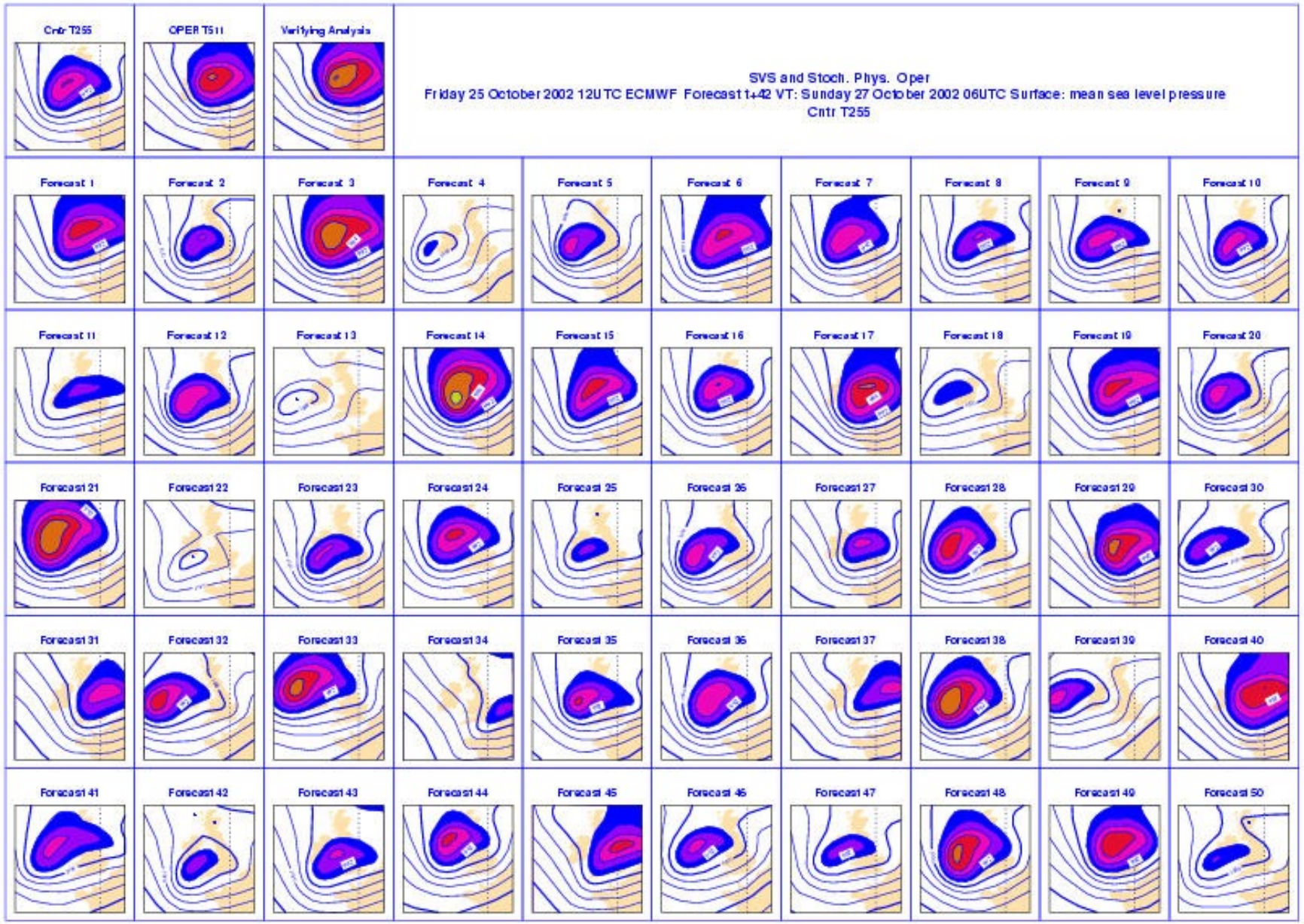
  McSharry & Smith. Physica D (2004)



Quiz

- The European Centre for Medium-Range Weather Forecasts (ECMWF) Ensemble Prediction System (EPS) is run twice a day.
- ECMWF's forecasts provide an ensemble of how many individual members?
  a) 17
  b) 27
  c) 52
  d) 102

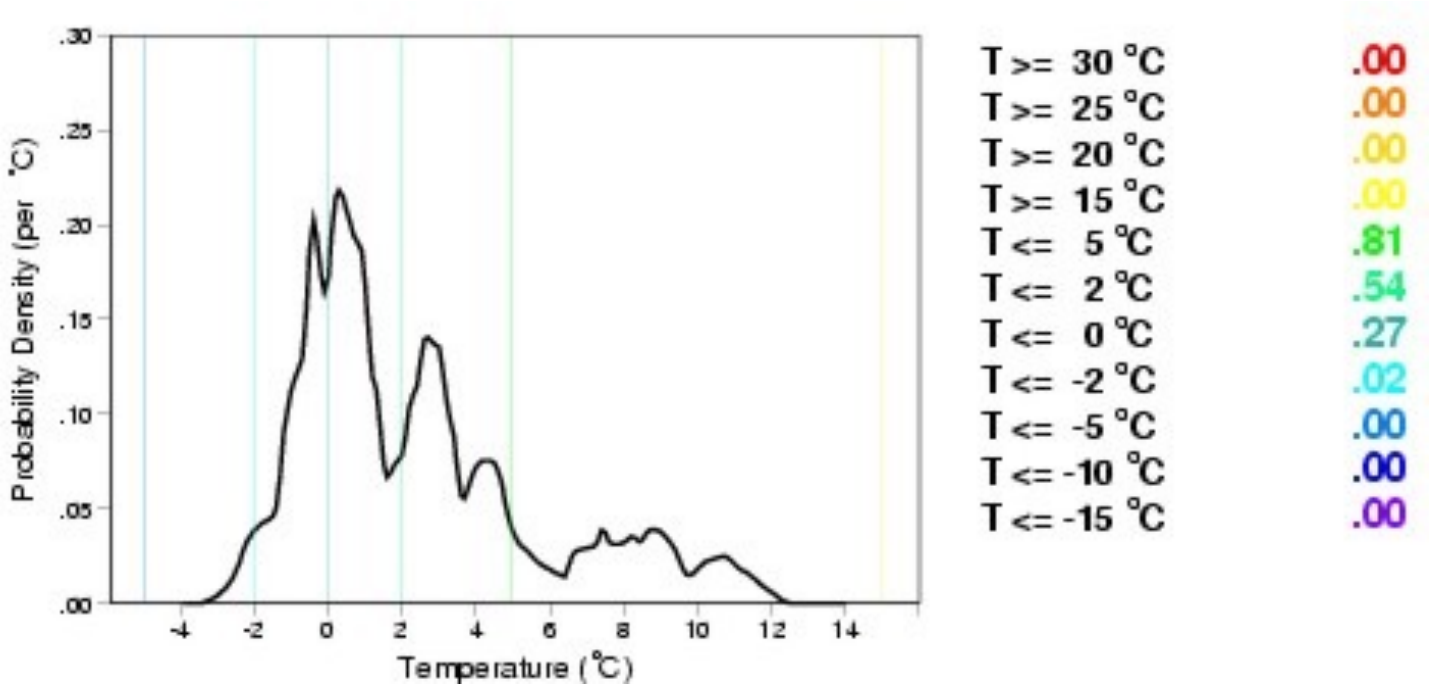
ECMWF 42 hour forecast for the October Storm of 2002



Each simulation looks physically reasonable, and most have storms…

Source: Lenny Smith

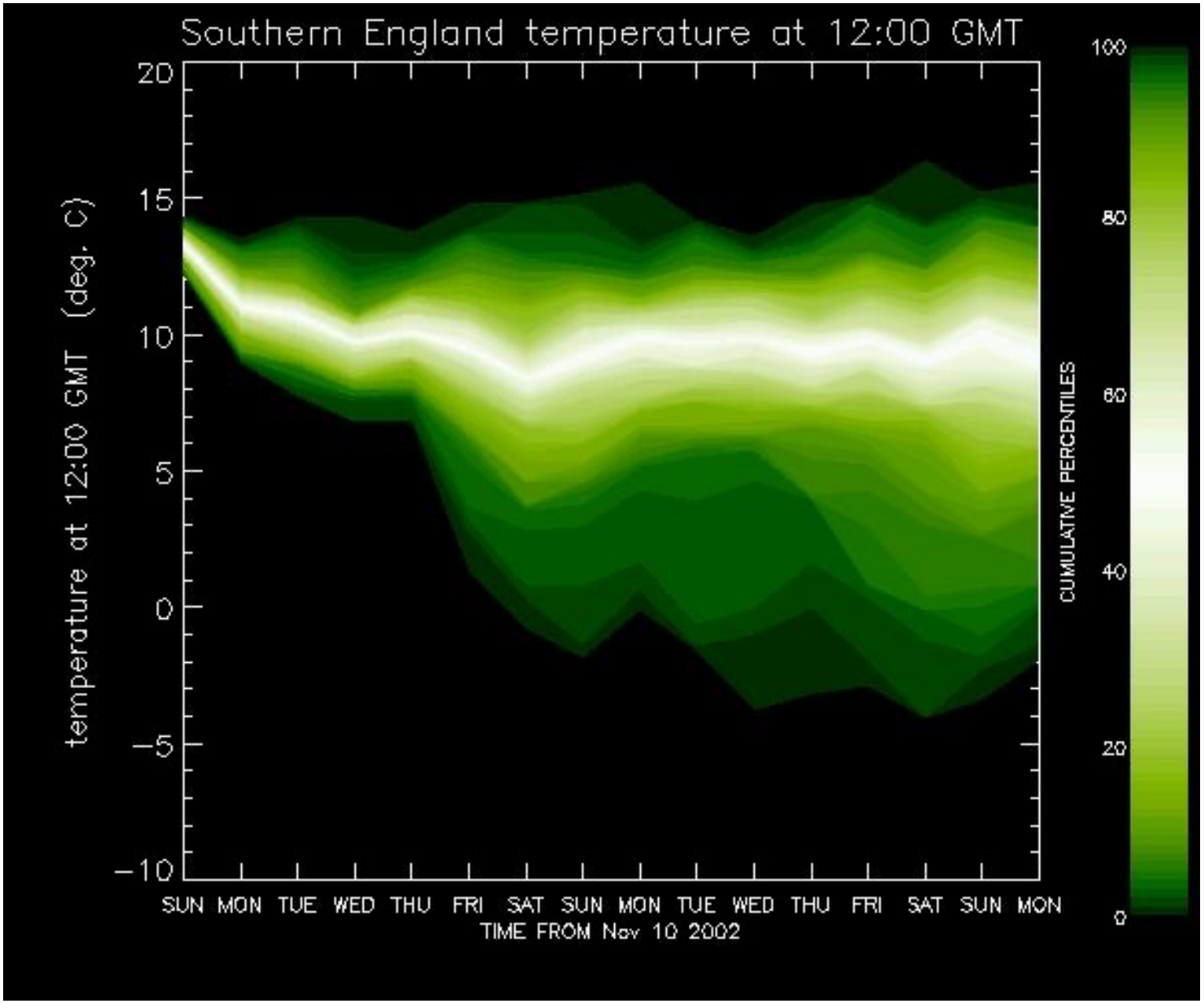
Temperature density forecast



- A calibrated 5-day forecast of the relative probabilities of different temperatures at Heathrow Airport for midday on 28th February 2004

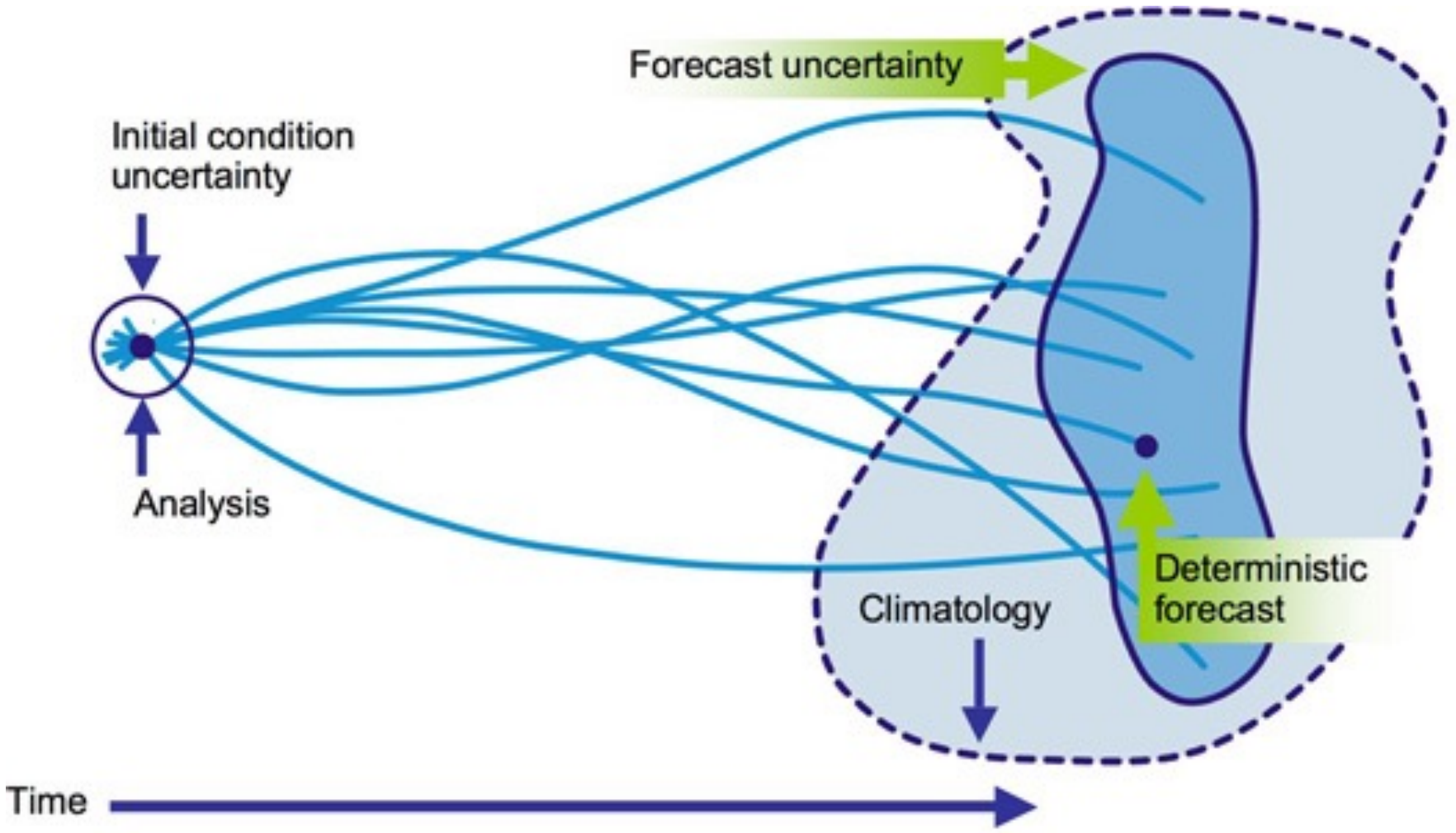
Source: UK Met Office

This is a DIME probability forecast for temperature:



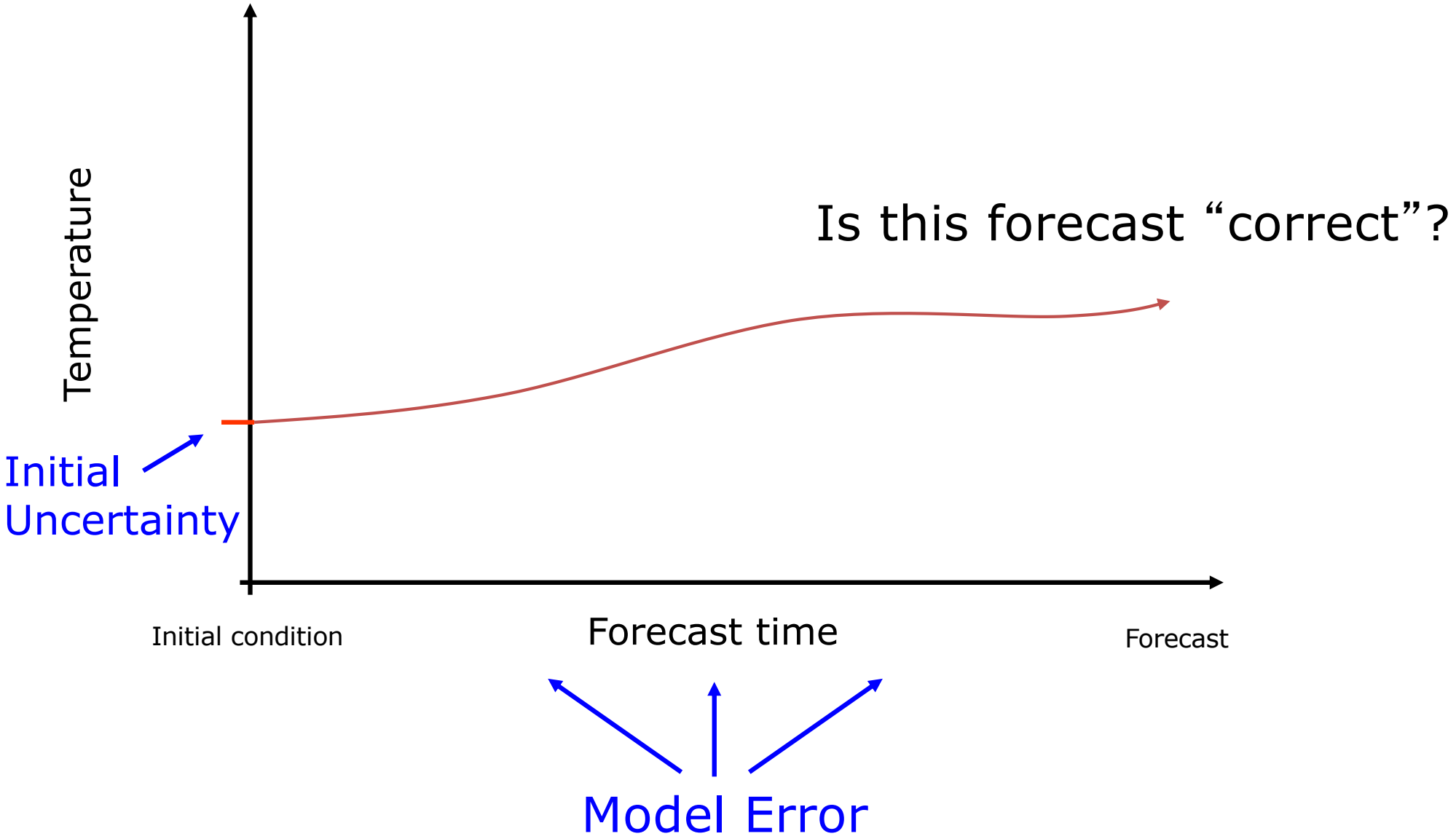
www.dime.lse.ac.uk

Ensemble creation



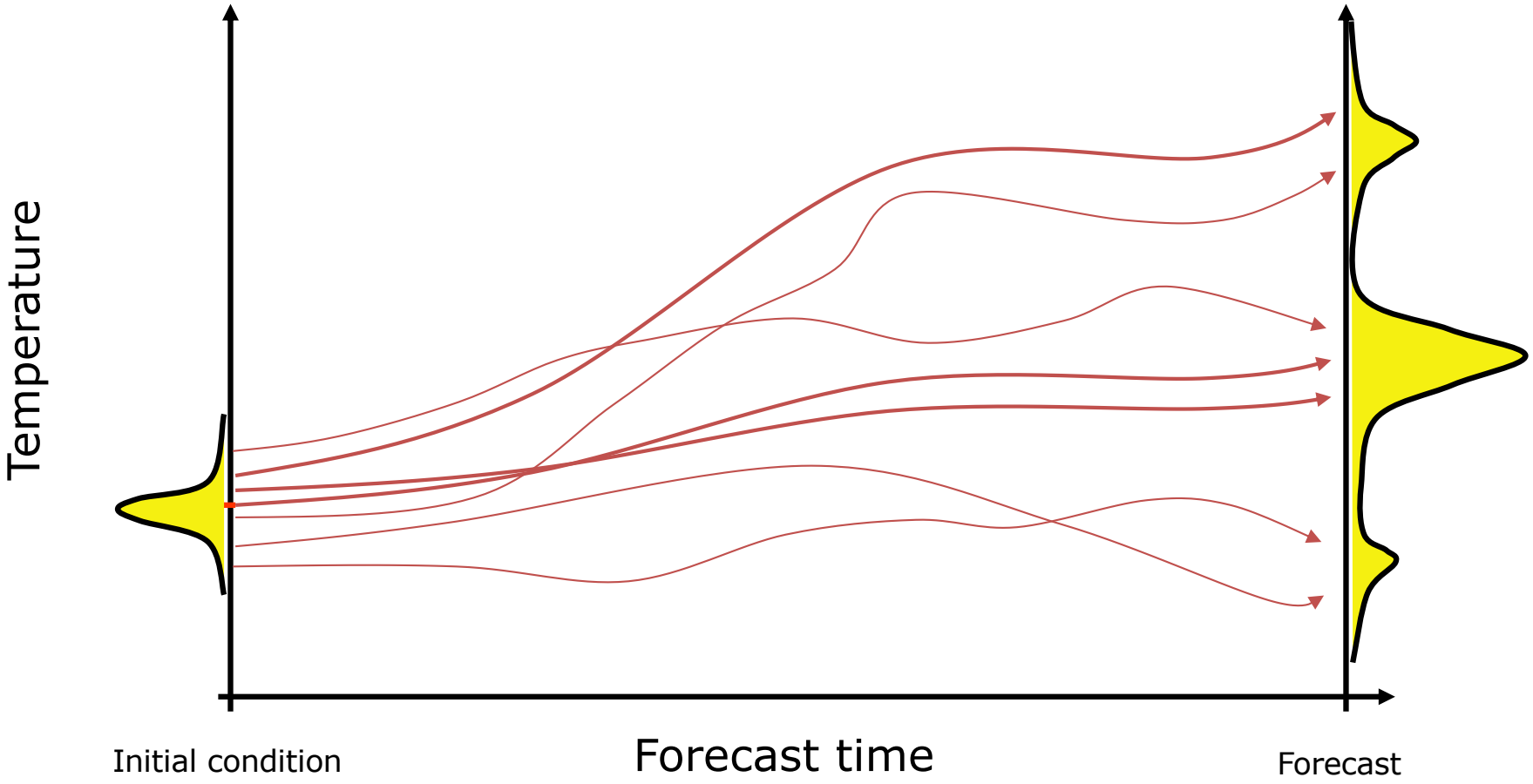
Source: www.easterbrook.ca

Deterministic Forecasting



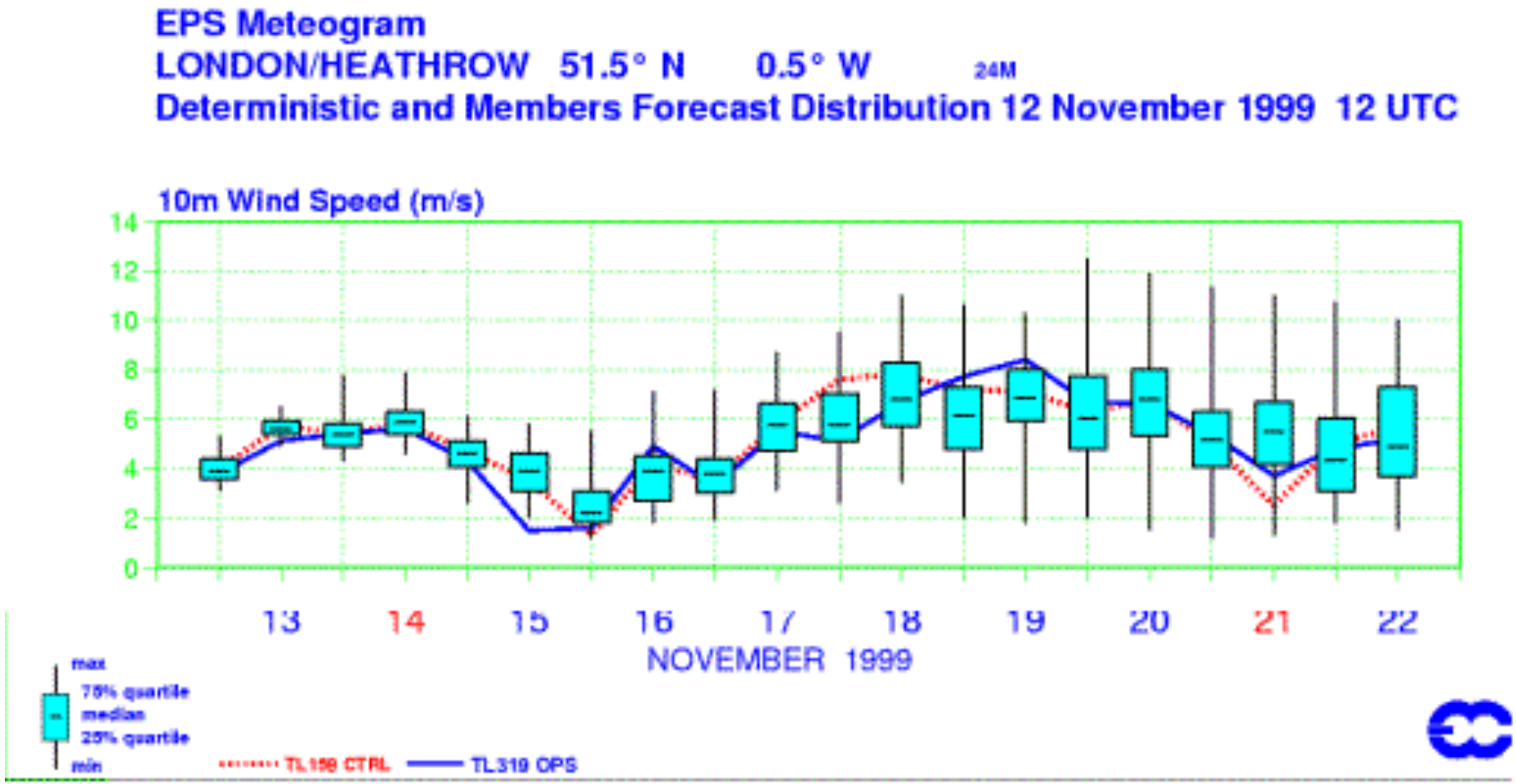
Source: Renate Hagedorn, ECMWF

Ensemble Forecasting



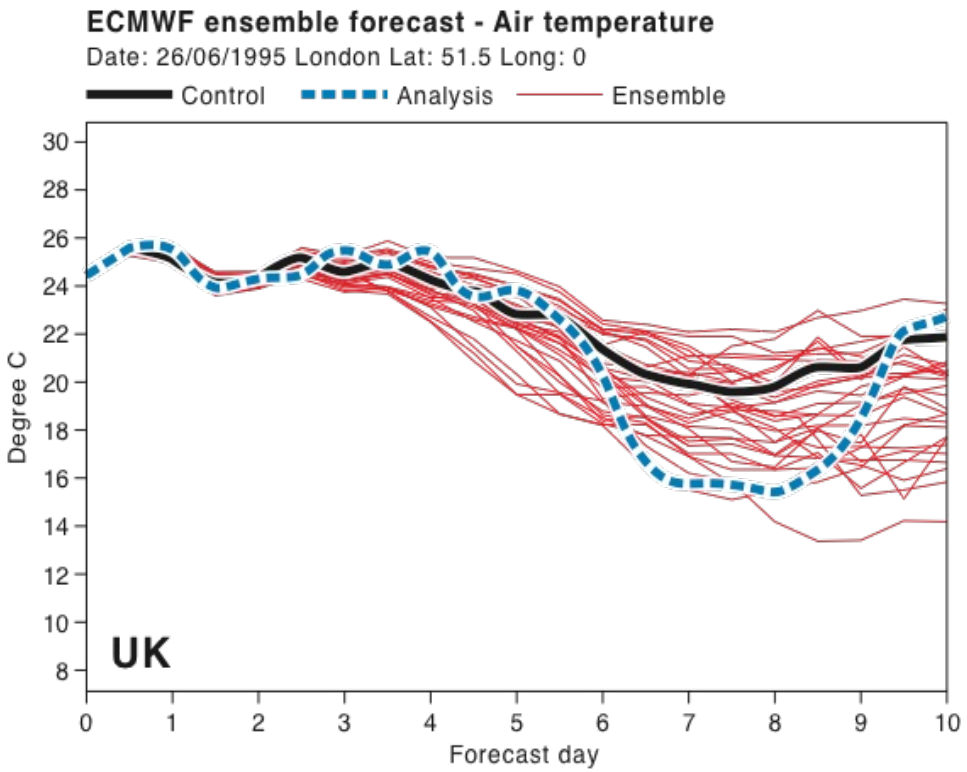
Description of weather prediction in terms of a Probability Density Function (PDF)

EPS Meteogram

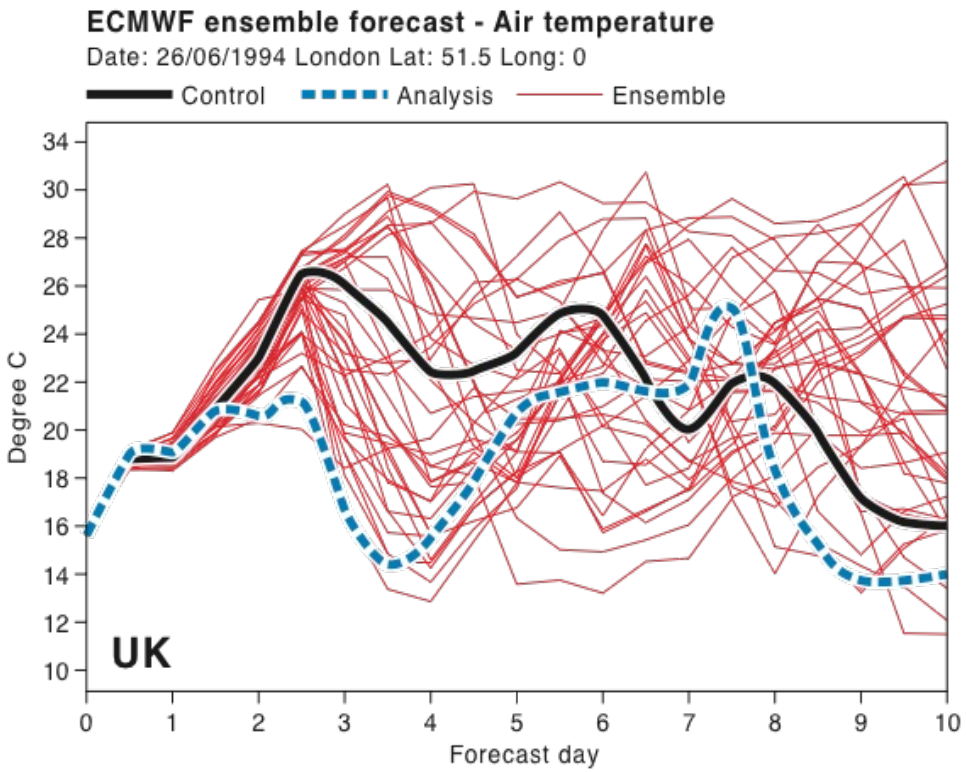


ECMWF Ensemble prediction

26th June 1995

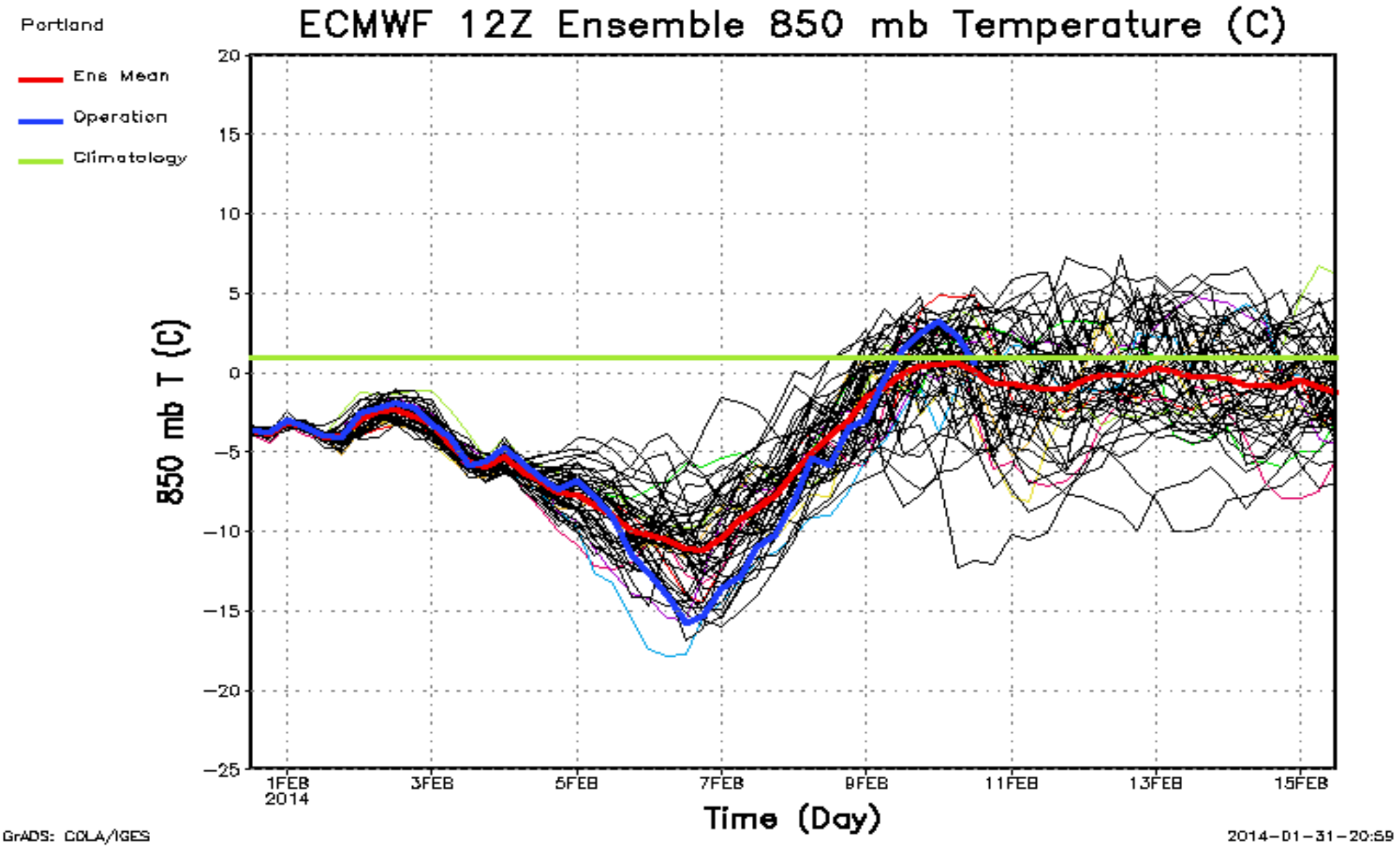


26th June 1994

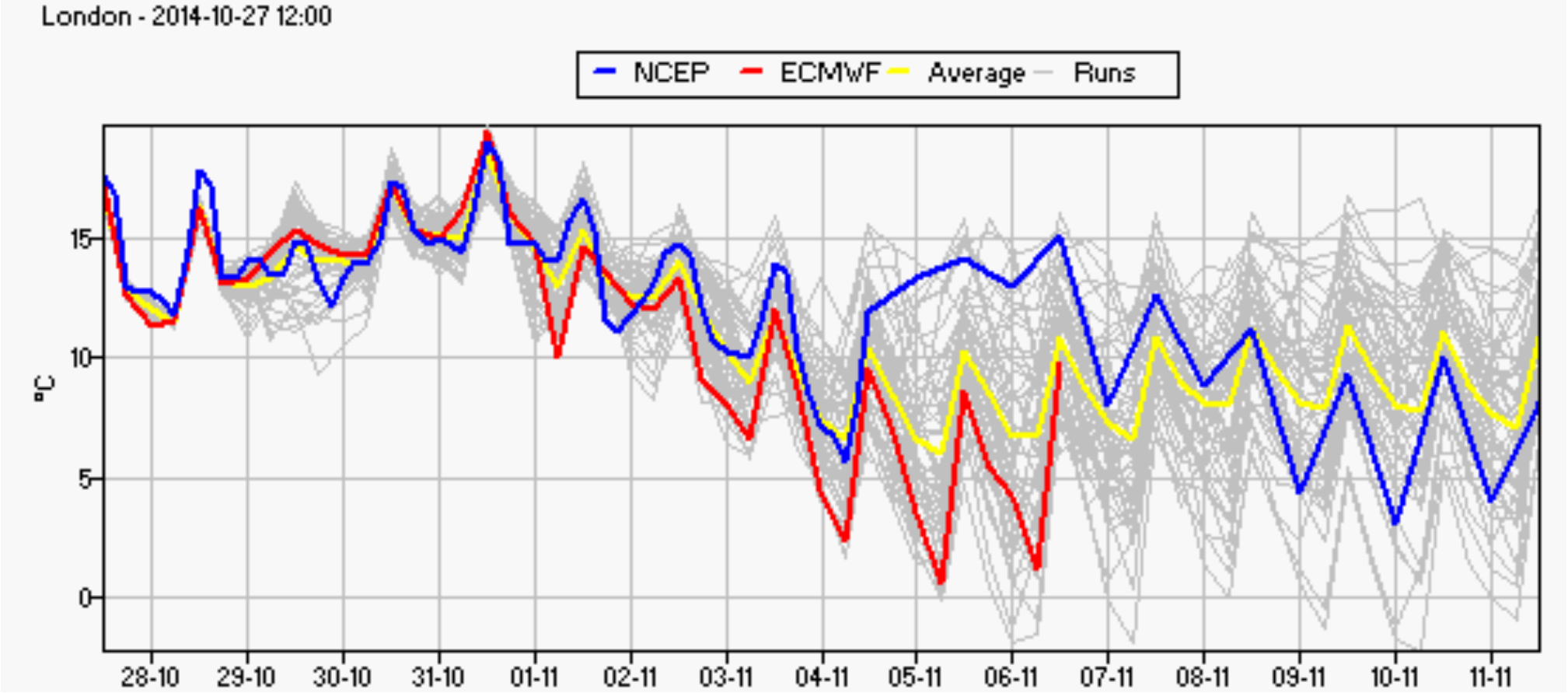


If the forecasts are coherent (small spread) the atmosphere is in a more predictable state than if the forecasts diverge (large spread)

ECMWF ensemble

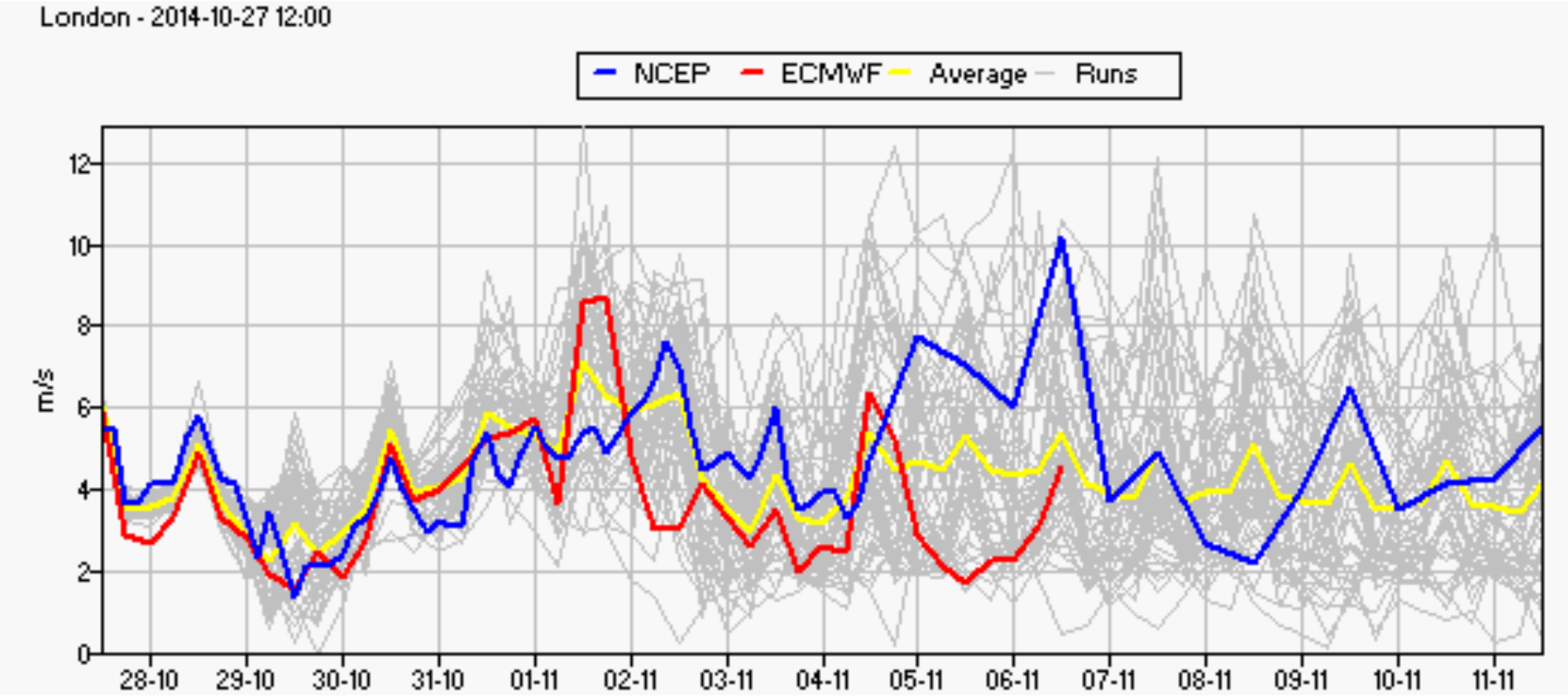


Multiple Temperature Ensembles



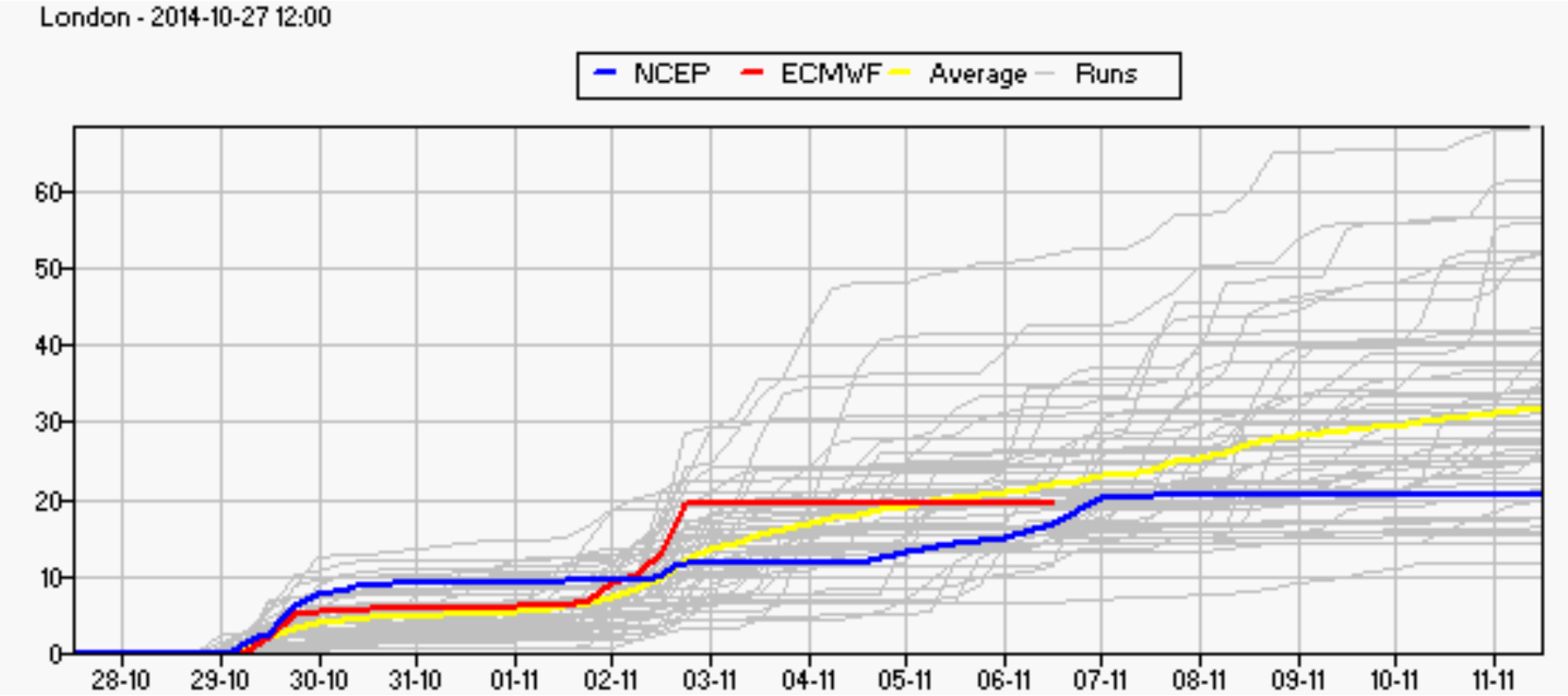
Source: www.weathercast.co.uk

Multiple Wind Ensembles



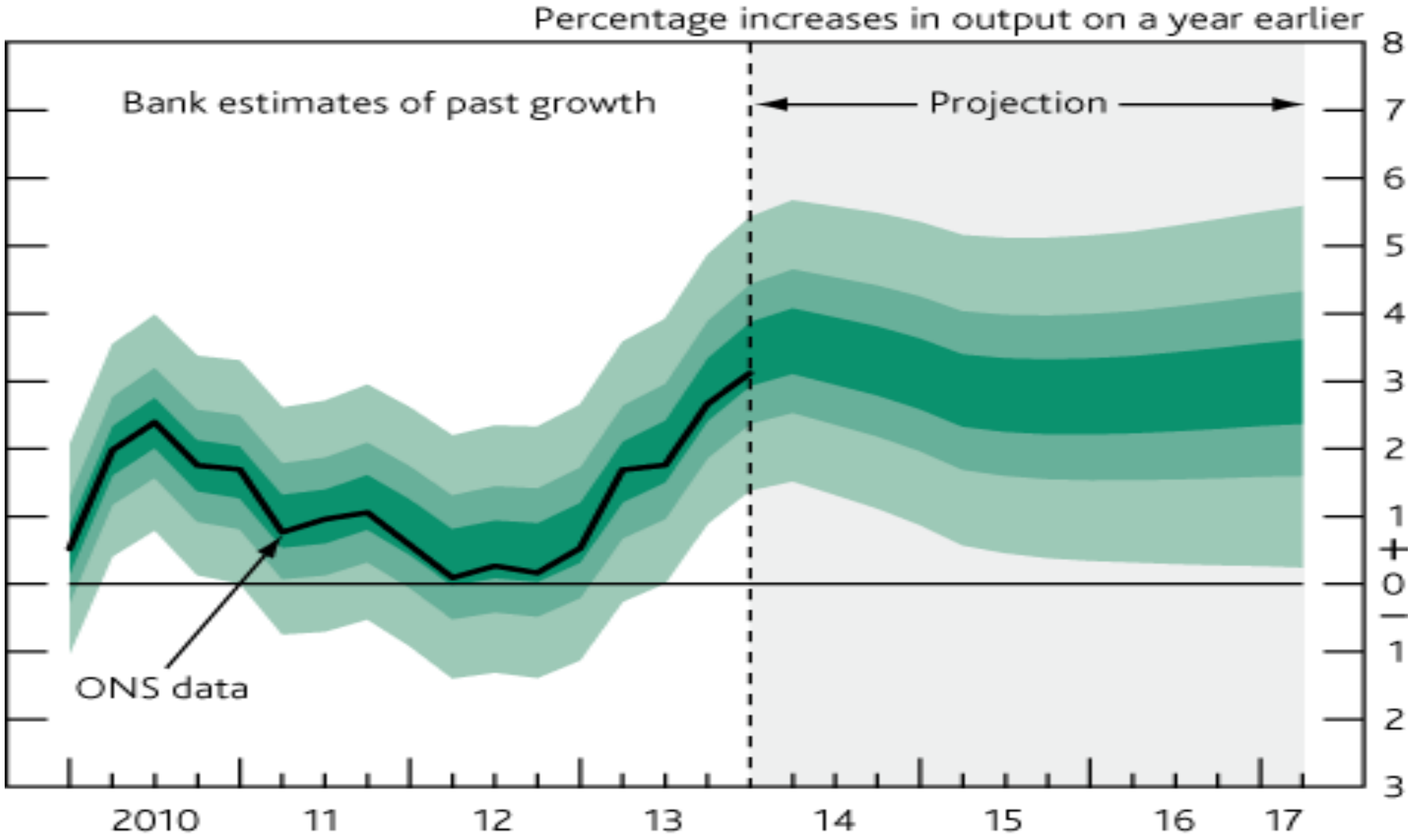
Source: www.weathercast.co.uk

Multiple Rainfall Ensembles



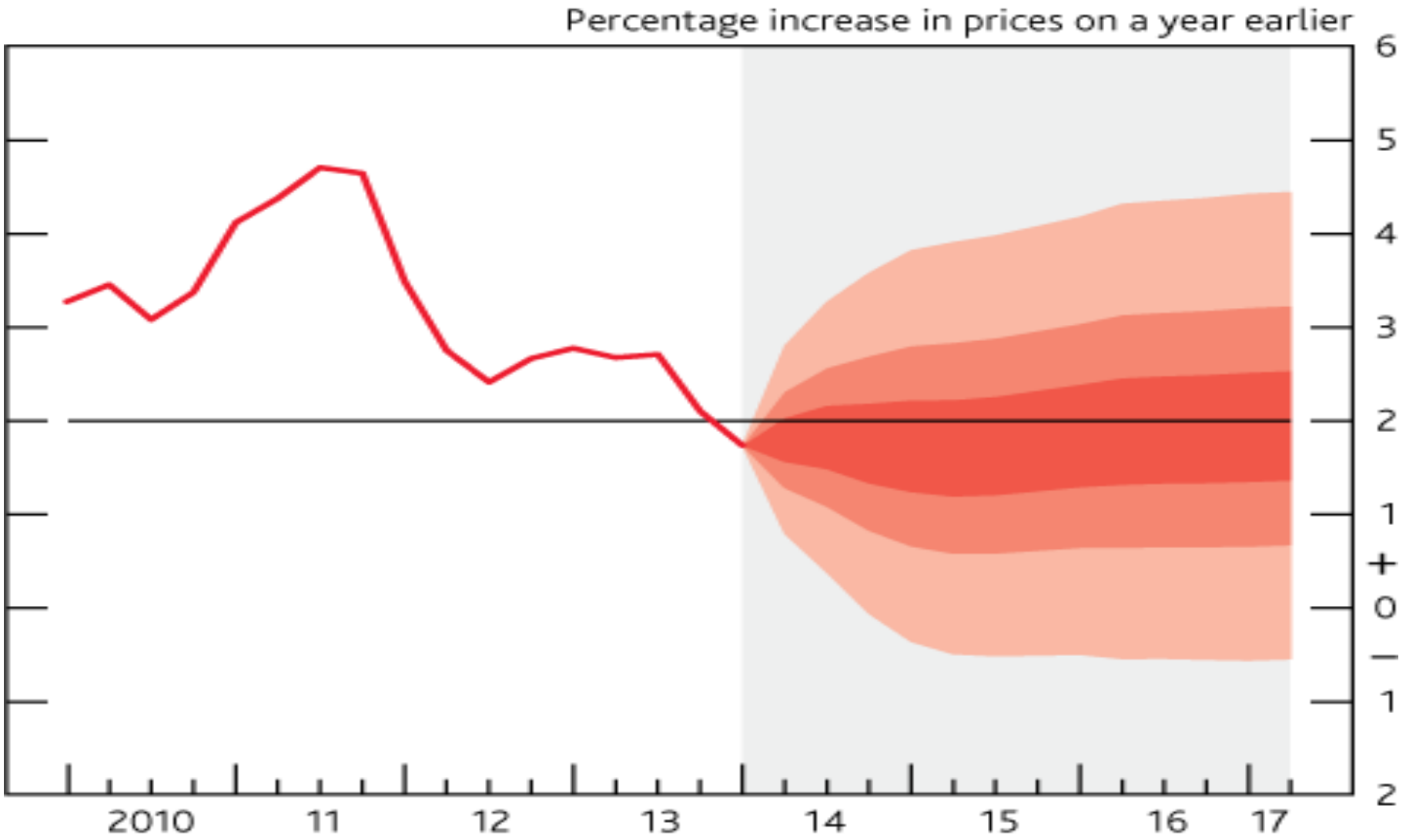
Source: www.weathercast.co.uk

BoE GDP



Visualisation of probability of various outcomes for GDP growth in the future.
Source: Inflation Report 2014, Bank of England.

BoE Inflation



Visualisation of probability of various outcomes for CPI inflation in the future.
Source: Inflation Report May 2014, Bank of England.

Jackknife

- The jackknife uses resampling to estimate the bias and standard error of a sample statistic.
- The jackknife resamples systematically (rather than at random as the bootstrap does).
- For a sample with n points, the jackknife computes sample statistics on n separate samples of size n-1.
- Each sample is the original data with a single observation omitted.

Bootstrap

- Bootstrap refers to any test or metric that relies on random sampling with replacement.
- Bootstrapping provides measures of accuracy (defined in terms of bias, variance, confidence intervals, prediction error or some other such measure) to sample estimates.
- This technique allows estimation of the sampling distribution of almost any statistic using random sampling methods.

Bootstrap for height

- Suppose we are interested in the average height of people worldwide.
- We cannot measure all the people in the global population, so instead we sample only a tiny part of it, and measure that.
- Assume the sample is of size N; that is, we measure the heights of N individuals.
- From that single sample, only one estimate of the mean can be obtained.
- In order to reason about the population, we need some sense of the variability of the mean that we have computed.

Bootstrap for height

- The simplest bootstrap method involves taking the original data set of N heights, and sampling from it to form a new bootstrap sample that is also of size N.
- The bootstrap sample is taken from the original using sampling with replacement so, assuming N is sufficiently large, for all practical purposes there is virtually zero probability that it will be identical to the original "real" sample.
- This process is repeated many times, and for each of these bootstrap samples we compute its mean (a bootstrap estimate).
- The collection of bootstrap means provides an estimate of the shape of the distribution of the mean from which we can answer questions about how much the mean varies.

Final Assignment A7

- PCA – 30 stocks
- Dendrogram for 30 stock to visualise relationships and cluster stocks
- Ensembles for classification
  - Titanic (LR, Tree, KNN, RF models)
  - ROC analysis
- Ensembles for regression
  - Wine quality feature selection (correlation, LASSO)
  - Compare LR, KNN and RF models
- Kaggle challenge bonus

Q&A

Applied Machine Learning

WEEK 12B

Today's Lecture

No.	Activity	Description	Time
1	Challenge	Model averaging	10
2	Discussion	Many versus the best	10
3	Case study	Netflix competition	10
4	Analysis	Bagging and Random Forests	20
5	Matlab Demo	Ensemble techniques	20
6	Q&A	Matlab questions and feedback	10

Poll

- Given a collection of models for prediction, can we asssume that the best model in-sample will always be the best model out-of-sample?
  - Yes
  - No

- **Slido.com**
- **#59055**

Model averaging

- Model Averaging is a technique designed to help account for the uncertainty inherent in the model selection process.
- This uncertainty in model structure is often neglected in traditional statistical analysis.
- By averaging over many different competing models, model uncertainty can be incorporated into conclusions about parameters and predictions.
- Model averaging has been applied successfully to many statistical model classes.
- In many cases, model averaging is successful at improving predictive performance.

Bayesian Model Averaging

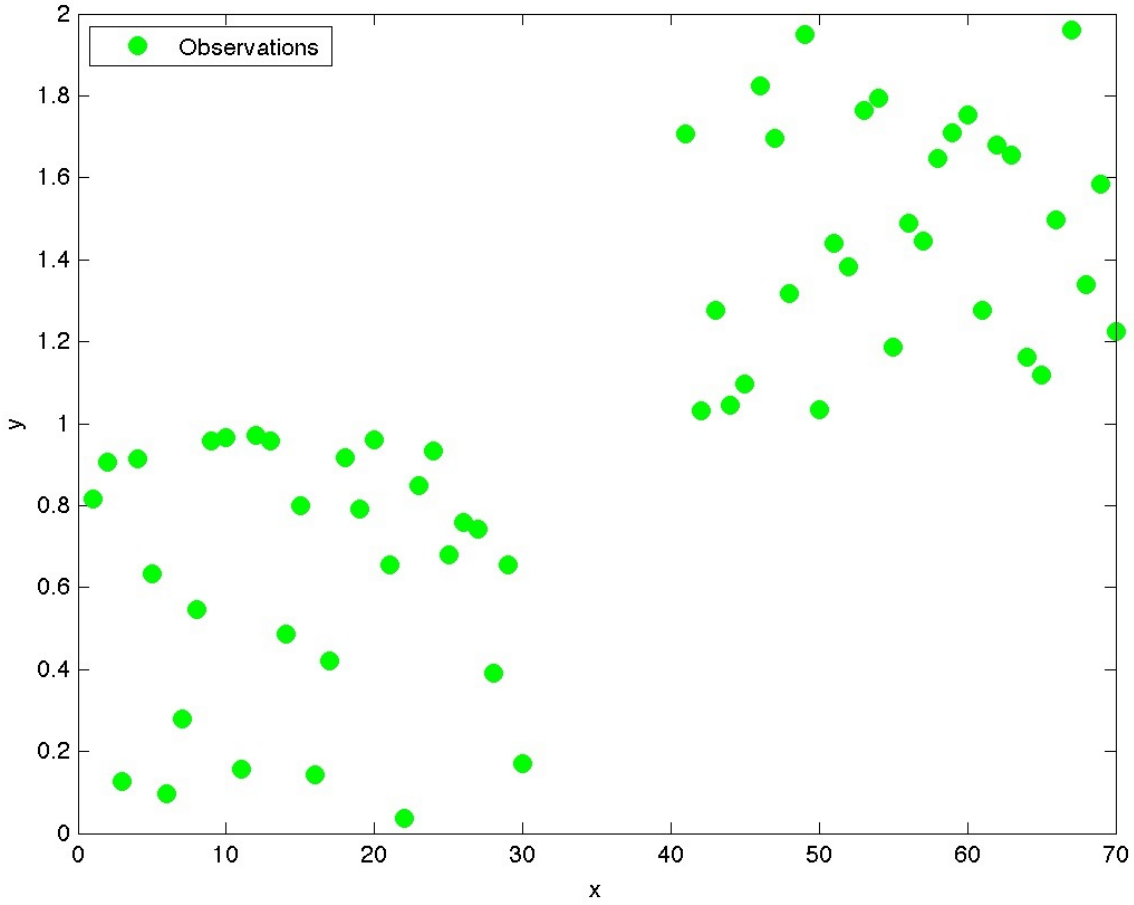
- Bayesian Model Averaging (BMA) models parameter uncertainty through the prior distribution and model uncertainty by obtaining posterior parameter and model posteriors using Bayes' theorem.
- BMA provides direct model selection, combined estimation and prediction.
- For a given quantity of interest x, observed data D and K models M_k (k=1,…,K), the marginal posterior distribution across all models p(x|D) is given by the average of all posterior distributions weighted by each posterior model probability:

$$p(x|D) = \sum_k p(x|D,M_k)p(M_k|D)$$

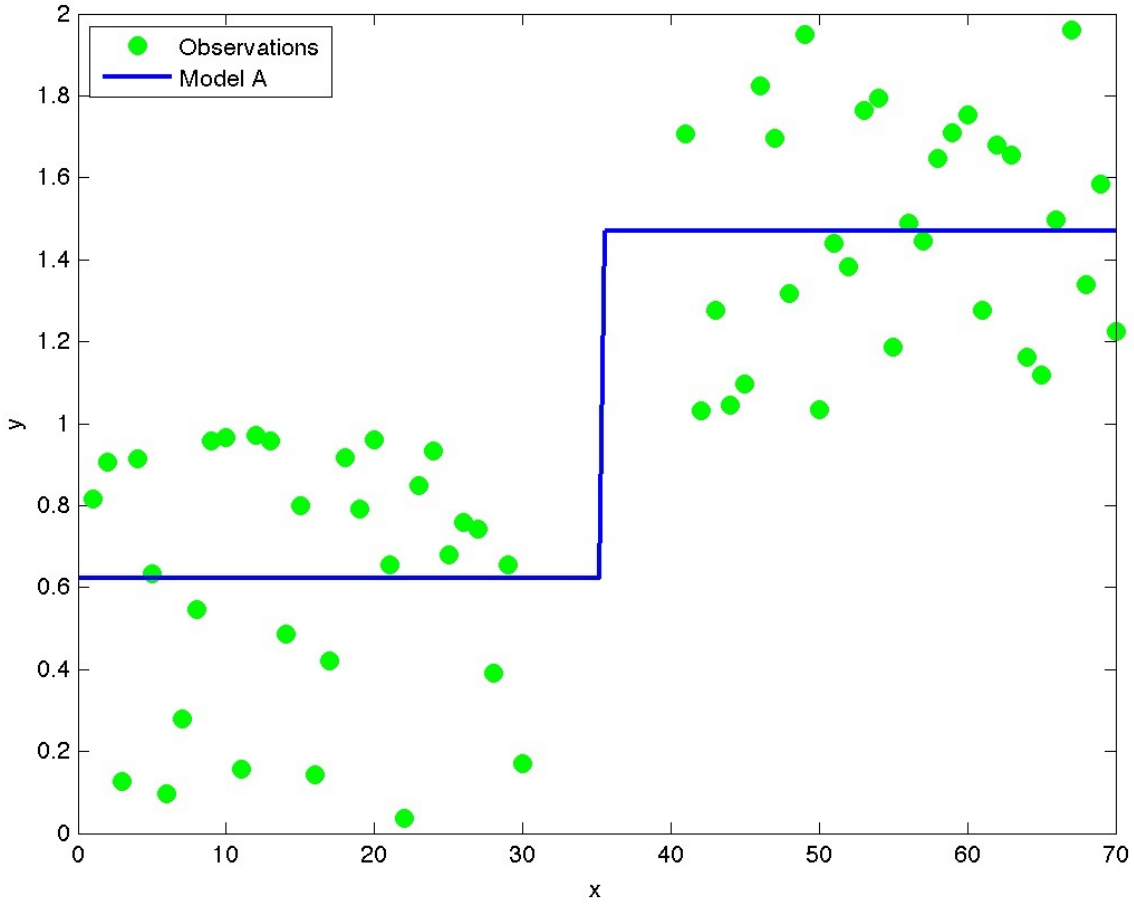
Bagging

- Bootstrap aggregating, known as bagging, aims to improve the stability and accuracy of machine learning approaches.
- Reduces variance and helps to avoid overfitting.
- Improves estimates from unstable procedures.
- Given a standard training set X of size *N*, bagging generates *M* new training sets (bootstraps), each of size *N′*, by sampling from X uniformly and with replacement.
- The *M* models are fitted using these *M* bootstrap samples and combined by averaging the output (for regression) or voting (for classification).

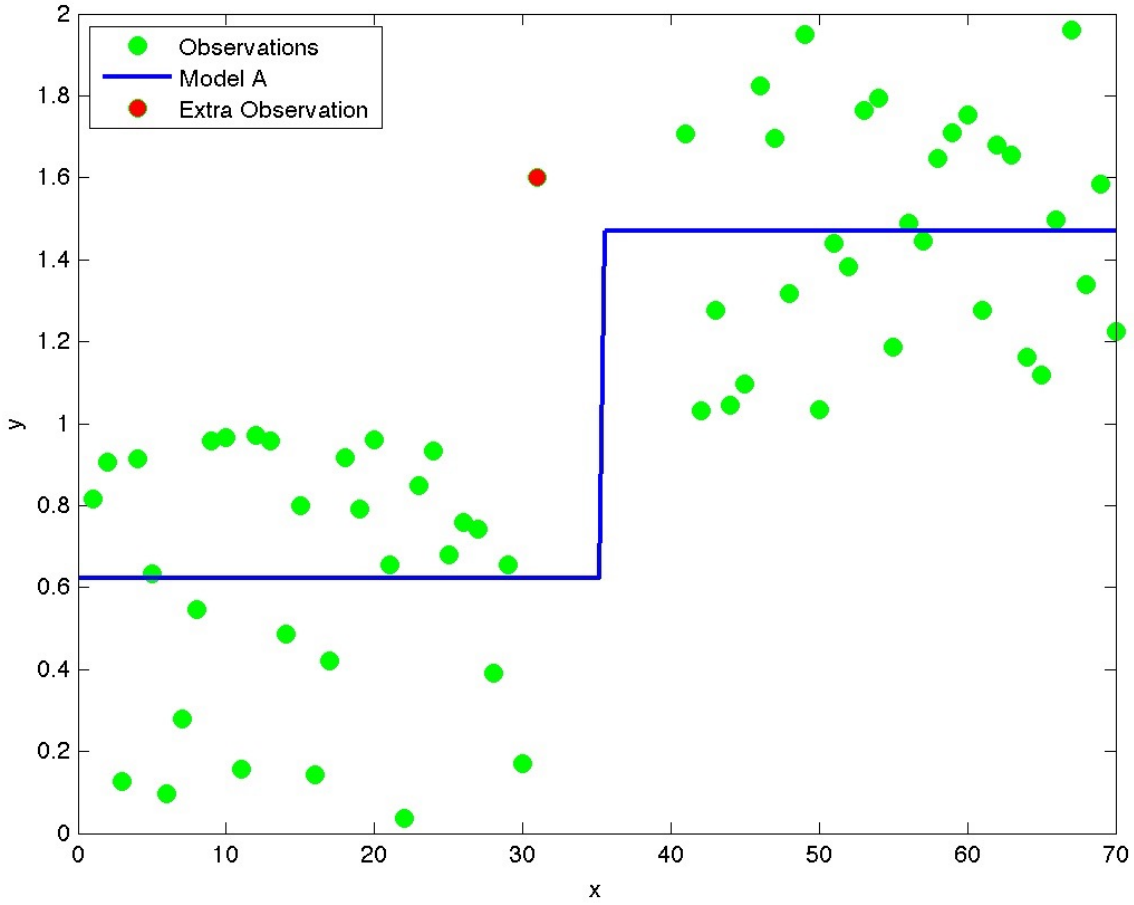
Bagging Demonstration



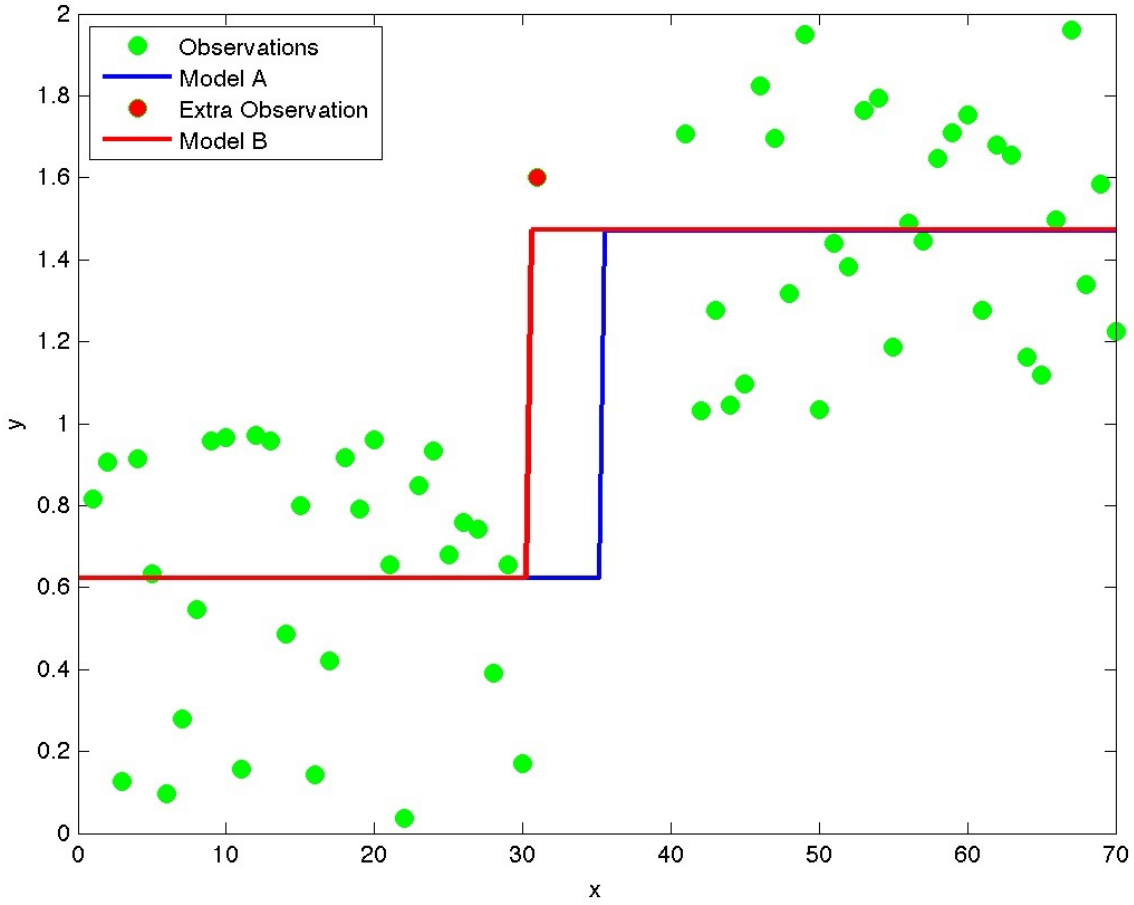
Bagging Demonstration



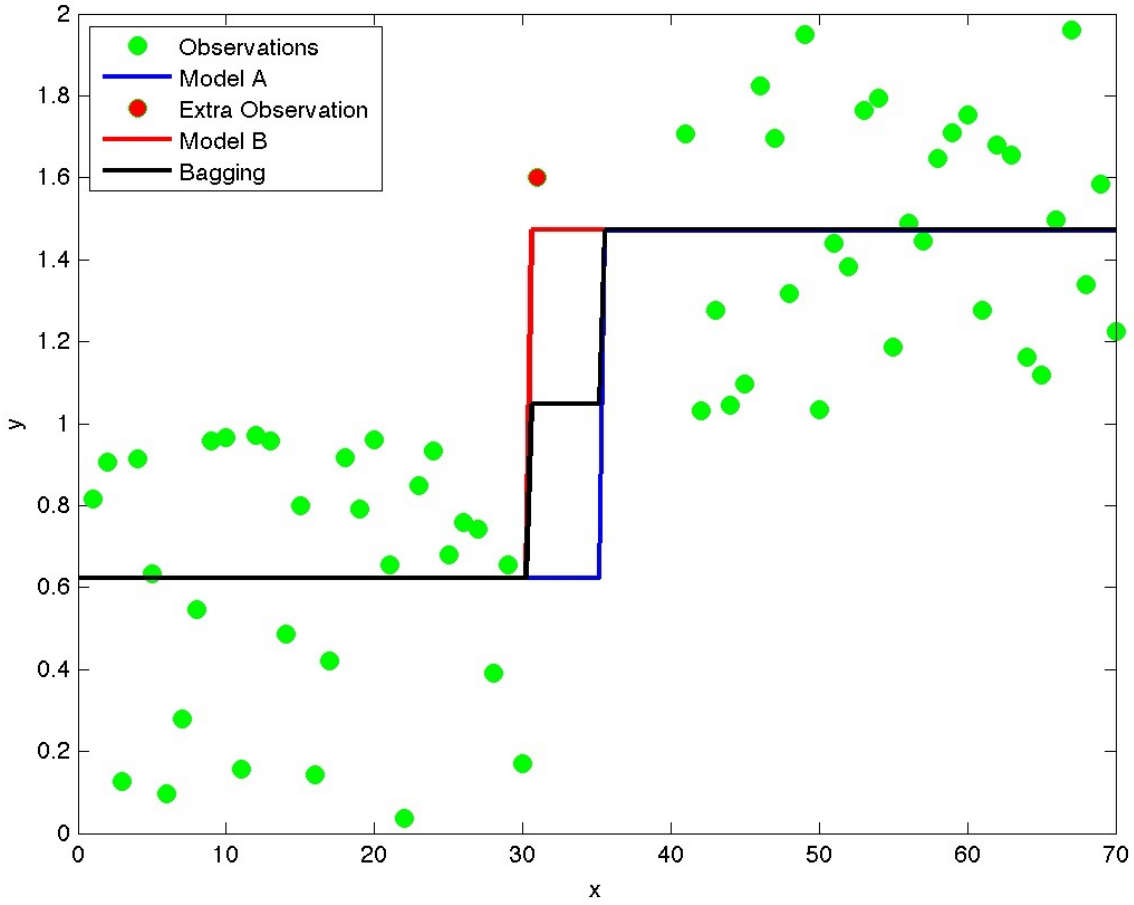
Bagging Demonstration



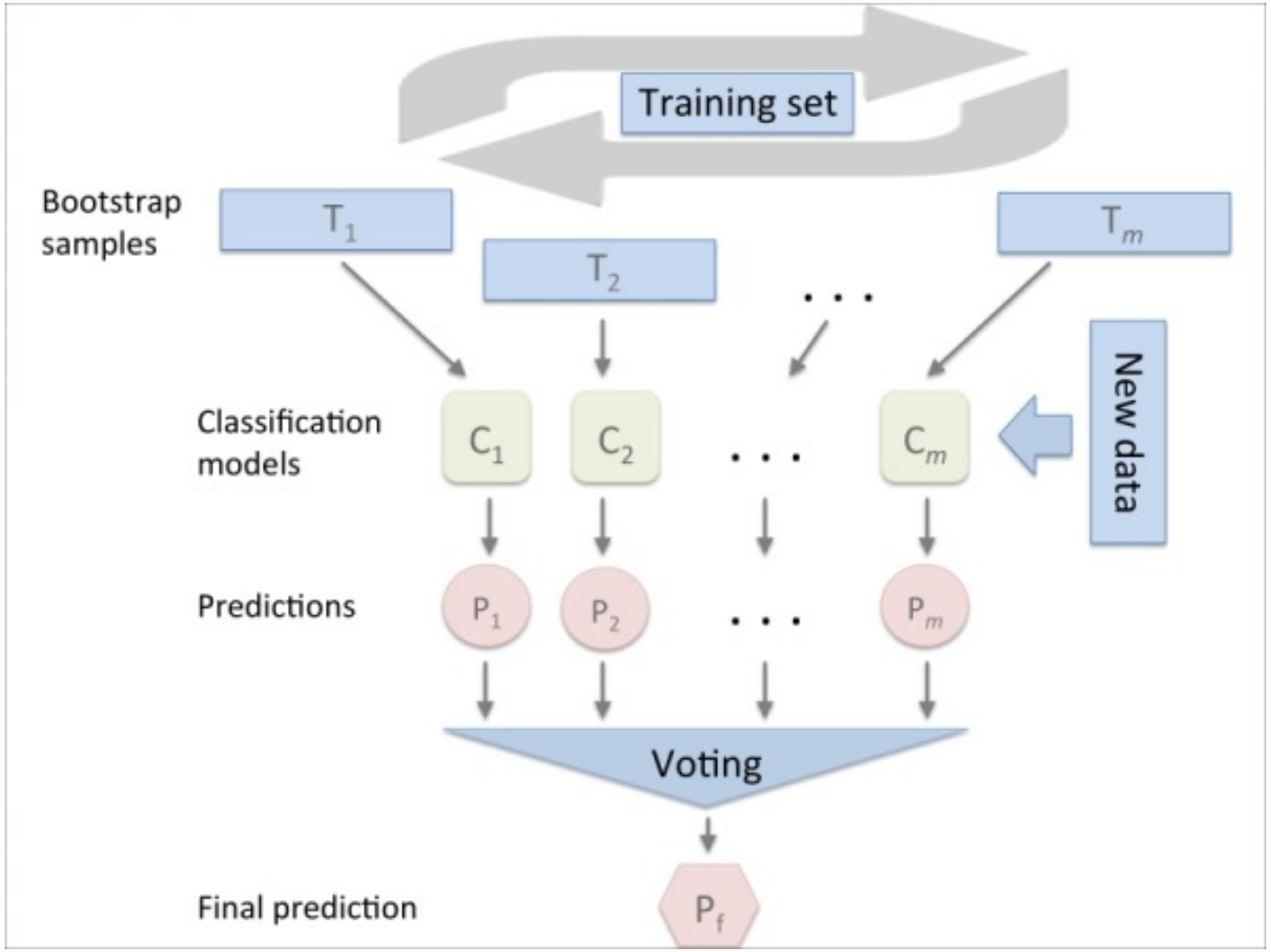
Bagging Demonstration



Bagging Demonstration

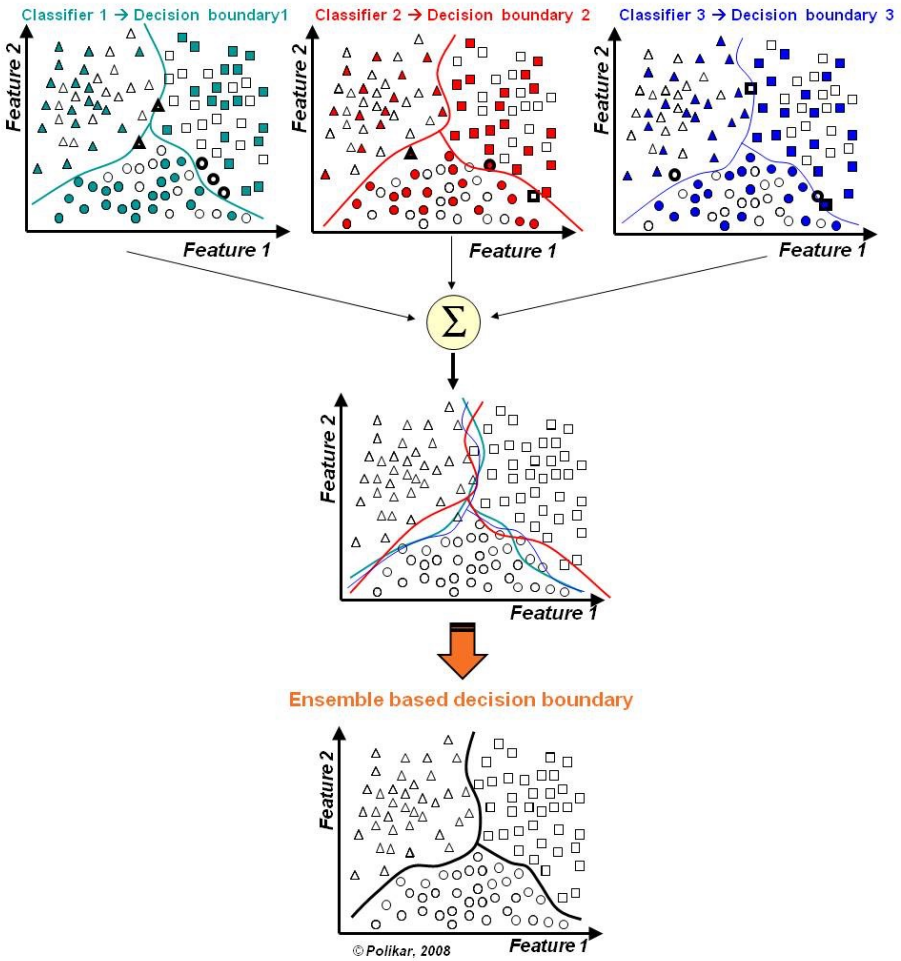


Bagging flowchart

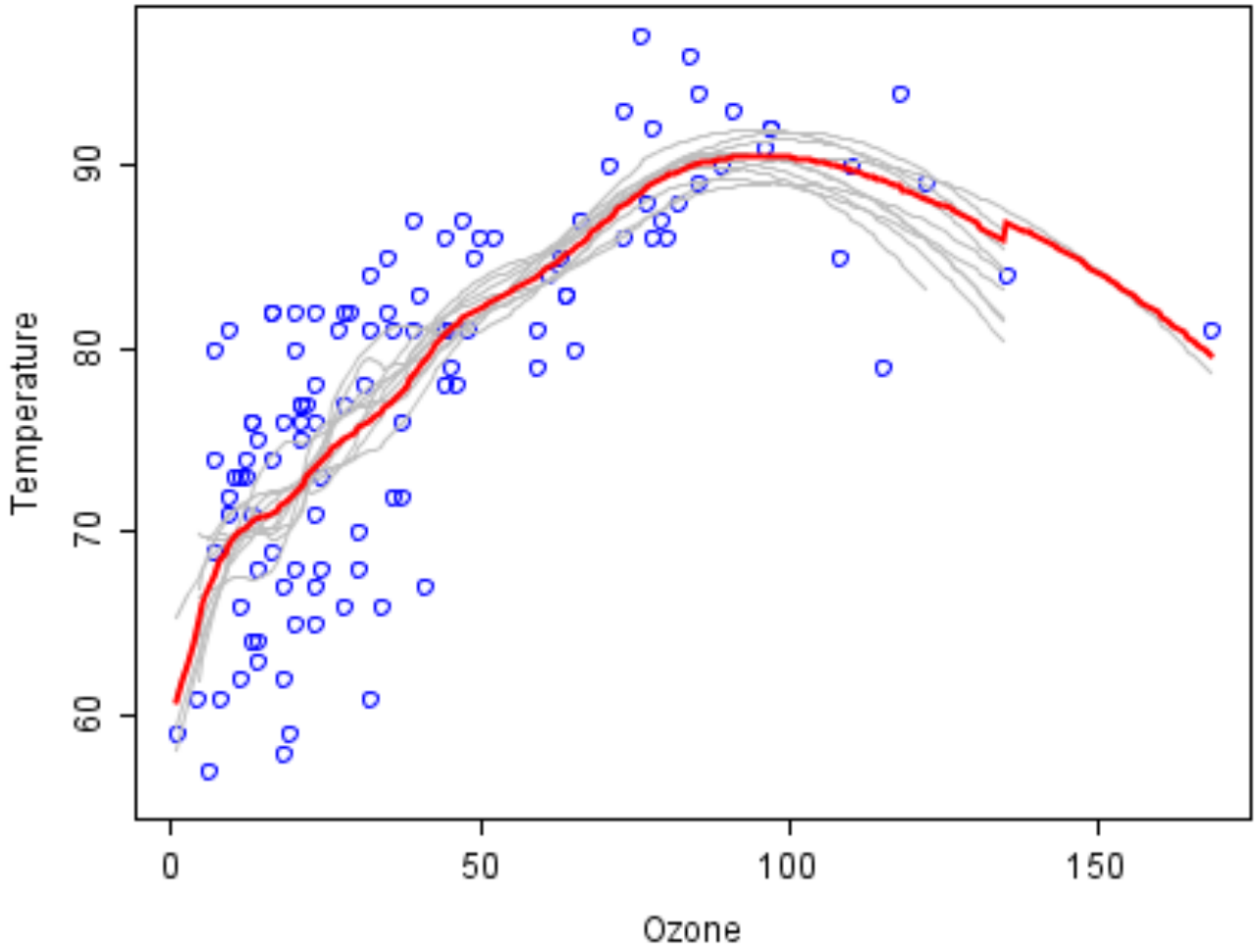


Source: Python Machine Learning

Bagging demonstration



Temperature and ozone



Source: Wikipedia and data from Rousseeuw and Leroy (1986)

Quiz

- Which of the following is likely to result from bagging:
  a) Decreased accuracy
  b) Increased variance
  c) Enhanced likelihood of overfitting
  d) Reduced interpretability

- **Slido.com**
- **#59055**

Ensemble modelling

- Combining knowledge and data-driven approaches
- No single perfect model exists
- Model selection often depends on particular realisation of time series or database available
- Identification of multiple predictive signals
- Ensembles provide a means of pooling predictive information

Ensemble learning

- Ensemble meta-algorithms exist to improve individual machine learning techniques.
- Concept: use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms.
- Combine many *weak learners* in an attempt to produce a *strong learner.*

Netflix competition

- Netflix offered a $1 million prize to anyone who could significantly improve its movie recommendation system Cinematch (with an RMSE of 0.9525) by 10%
- The winning team, "BellKor's Pragmatic Chaos", a group of 7 individuals, achieved 10.06%
- The runners-up, "Ensemble", formed from a collection of 28 teams, achieved 10.06%
- A 50/50 blend of the two would have achieved 10.19%
- The competition inspired kaggle.com and zindi.africa

Poll

- One of the main objectives for sponsoring a prize on a prediction competition platform is:
  a) Providing access to data
  b) Posing a well defined challenge
  c) Evaluating independent solutions
  d) Encouraging community collaboration

Boosting

- Boosting involves incrementally building an ensemble by training each new model instance to emphasize the training instances that previous models mis-classified.
- In some cases, boosting has been shown to yield better accuracy than bagging, but it also tends to be more likely to over-fit the training data.
- Adaboost or “Adaptive Boosting” is one such technique.

Decision trees

- Decision trees have many advantages:
  - invariant under scaling and various other transformations of feature values;
  - robust to inclusion of irrelevant features; and
  - produces transparent and understandable models.
- Unfortunately, decision trees are rarely accurate.
- Deep trees tend to over-fit their training sets, because they have low bias, but very high variance.

Ensemble of trees

- The predictions of a single tree are highly sensitive to noise in its training set.
- Fortunately the average of many trees is less sensitive, as long as the trees are not correlated.
- The average output is defined as the mode of the classes for classification and the mean prediction for regression.
- Training many trees on a single training set would give strongly correlated trees.
- If the training algorithm is deterministic, it would yield the same tree many times.
- Bootstrap sampling offers a way of de-correlating the trees by showing them different training sets.

Correlated trees

- An ordinary bootstrap sample is problematic in situations where a few features are very strong predictors for the response variable.
- In this case, the same features will be selected in many of the trees in the ensemble, implying that these trees will be strongly correlated.
- Bagging works only when the individual trees are not correlated.
- A diverse set of trees is required for bagging to lead to a reduction in variance.

Feature bagging

- A modified tree learning algorithm that selects, at each candidate split in the learning process, a random subset of the features offers a means of obtaining trees that are not so highly correlated.
- This process is sometimes called "feature bagging".
- In practice, for a dataset with m features, $\sqrt{m}$ features are used in each split.

Random Forests

- Breiman & Cutler (2001) introduced Random Forests (RF) as an approach which combines bagging of trees with the random selection of features.
- RF averages across multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance.
- This approach comes at the expense of a small increase in the bias and some loss of interpretability, but generally greatly improves the performance of the final model.

How many trees?

- The number of trees used is a key parameter.
- In practice a few thousand trees are used. Computational time is an issue and depends on the size and nature of the training set.
- An optimal number of trees can be found using cross-validation.
- An alternative is to use the out-of-bag error: the mean prediction error on each training sample x_i, using only the trees that did not have x_i in their bootstrap sample.
- The training and test error tend to level off after some number of trees have been fit.

Feature importance

- During the RF fitting process, the out-of-bag error for each data point is recorded and averaged over the forest.
- To measure the importance of the mth feature after training, the values of the mth feature are permuted among the training data and the out-of-bag error is again computed on this perturbed data set.
- The importance score for the mth feature is computed by averaging the difference in out-of-bag error before and after the permutation over all trees.
- The score is normalized by the standard deviation of these differences.
- Features which produce large values for this score are ranked as more important than features which produce small values.

Random subspace method

- Also known as attribute bagging, this is a generalization of the random forest method.
- Rather than use decision trees, a random subspace classifier can be composed from any type of underlying classifiers such as linear classifiers, SVMs and KNNs.
- Tends to work well when the number of features is much larger than the number of training objects.

Genetic Algorithms

- A genetic algorithm (GA) is a search heuristic that mimics the process of natural selection.
- Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution.
- Key ingredients include inheritance, mutation, selection, and crossover.

Implementing a GA

- **Initialization**: random selection of solutions
- **Selection**: for each successive generation, a proportion of the existing population is selected using a fitness function to breed a new generation.
- **Breeding**: a pair of "parent" solutions is selected to produce a "child" solution using the methods of crossover and mutation.
- **Crossover**: combination of parents solutions.
- **Mutation**: additional of random solution.
- **Termination**: minimum criteria satisfied or number of generations exceeded.

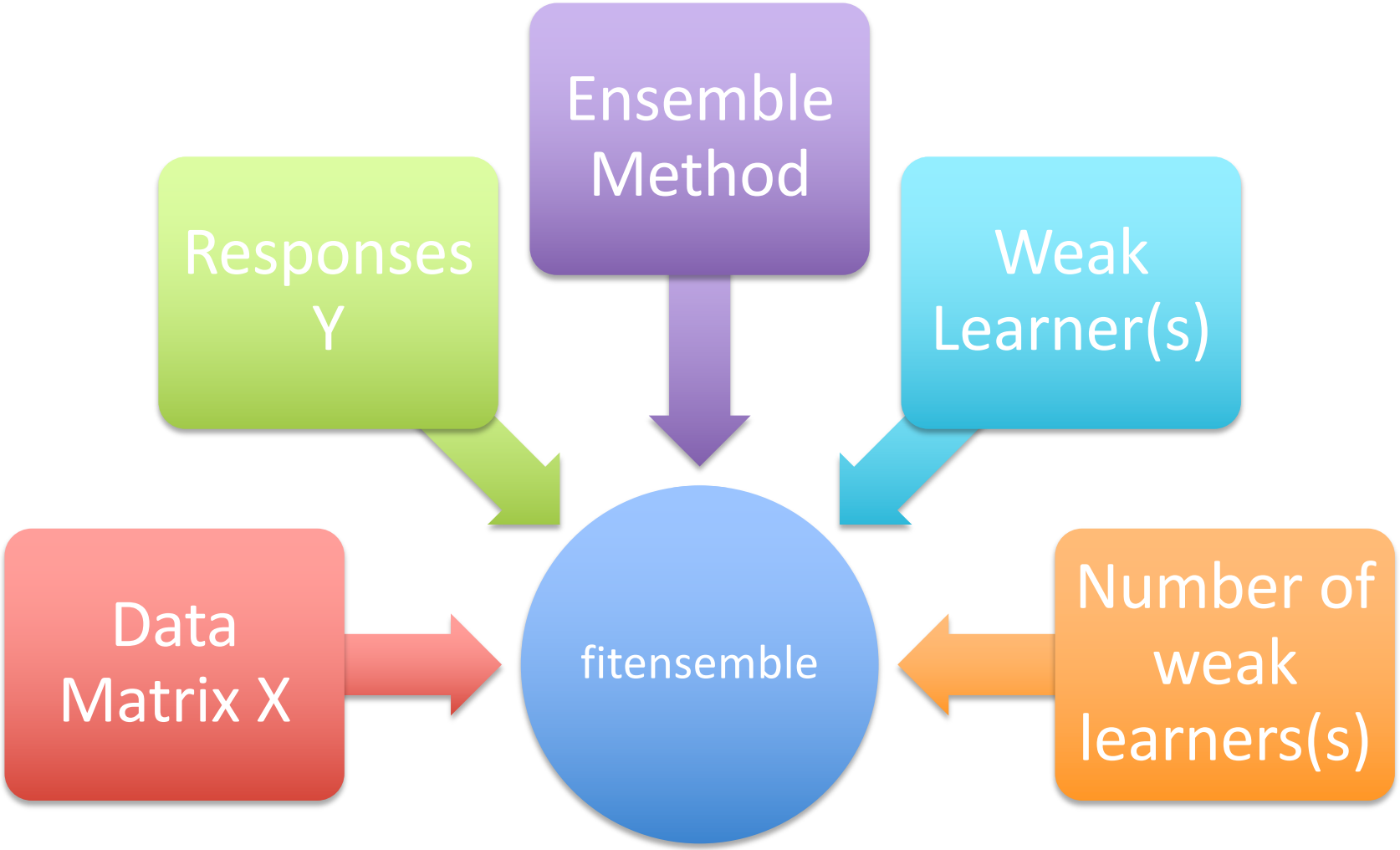
Reinforcement Learning

- Reinforcement learning was inspired by behavioral psychology and concerns how agents act in an environment so as to maximize reward.
- It differs from standard supervised learning in that correct input/output pairs are never presented, nor sub-optimal actions explicitly corrected.
- Applications include robot control, chess, backgammon, checkers, and other activities where a software agent can learn.

Reinforcement learning

- Reinforcement learning is applied on-line and therefore involves finding a balance between exploration of uncharted territory and exploitation of existing knowledge.
- The components of reinforcement learning are:
  - a set of environment states ;
  - a set of actions;
  - rules of transitioning between states;
  - rules that determine the *reward* of a transition; and
  - rules that describe what the agent observes.

Constructing an ensemble



Ensemble Method

Method	Binary Classification	Multi-class Classification	Regression
AdaBoostM1	X	X	
LogitBoost	X		
GentleBoost	X		
RobustBoost	X		
LPBoost	X	X	X
TotalBoost	X	X	
RUSBoost	X	X	
Subspace	X	X	
Bag	X	X	
LSBoost			X

Weak Learners

Name	Ensemble Method
Discriminant	for Subspace ensemble
KNN	for Subspace ensemble
Tree	for any ensemble except Subspace

Creating an ensemble

- The syntax of fitensemble is

  ens = fitensemble(X,Y,method,numberens,learners)
- **X** is the matrix of data. Each row contains one observation, and each column contains one predictor variable.
- **Y** is the responses, with the same number of observations as rows in X.
- **method** is a string naming the type of ensemble method.
- **numberens** is the number of weak learners in ens from each element of learners. So the number of elements in ens is numberens times the number of elements in learners.
- **learners** is a string naming a weak learner, a weak learner template, or a cell array of such strings and templates.

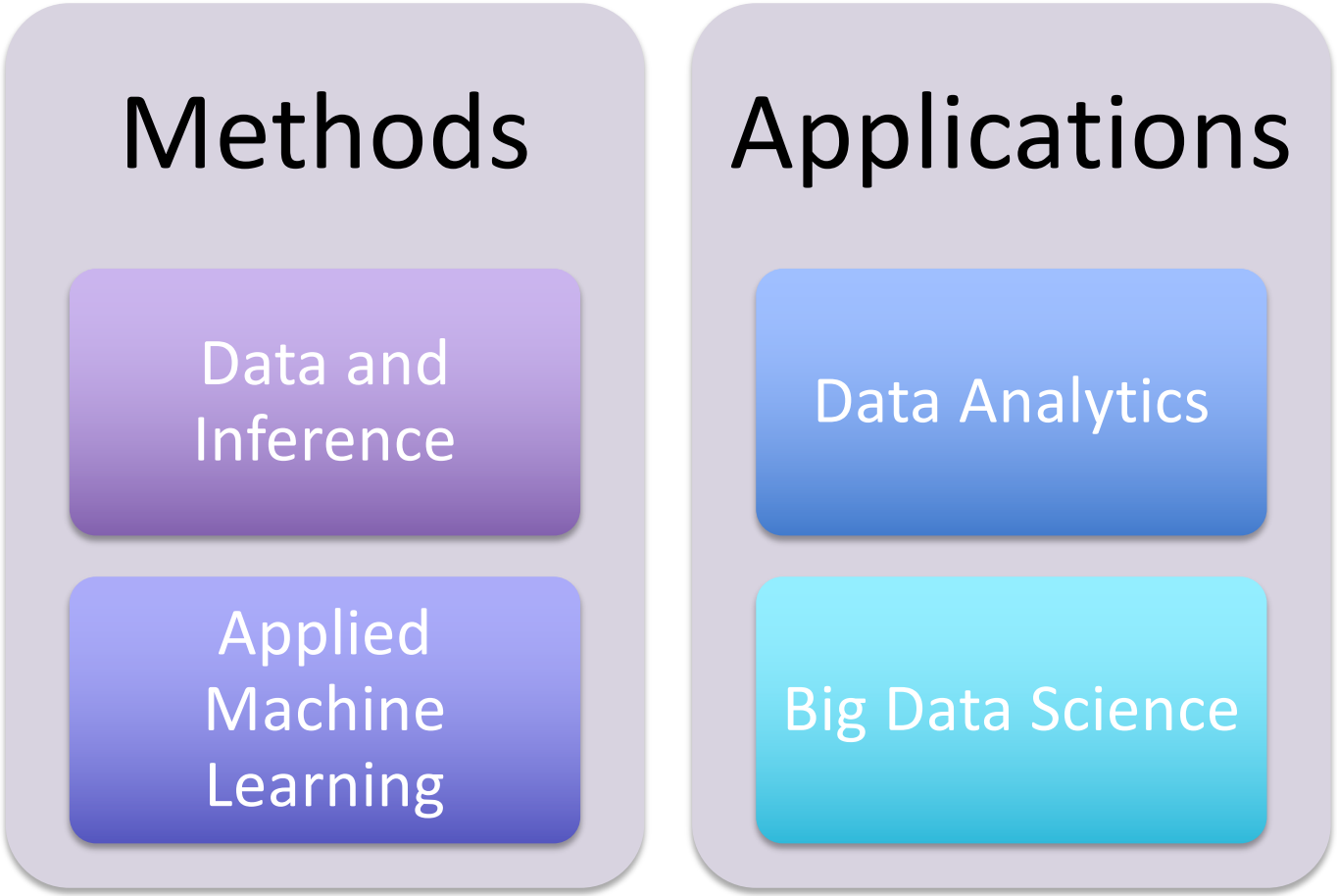
Evaluating an ensemble

- To obtain a better idea of the quality of an ensemble, use one of these three methods:

Description	When to use	Matlab function
Evaluate the ensemble on an independent test set	Useful when you have a lot of training data	loss
Evaluate the ensemble by cross validation	Useful when you don't have a lot of training data	kfoldloss
Evaluate the ensemble on out-of-bag data	Useful when you create a bagged ensemble with fitensemble	oobLoss

Matlab

- fitensemble
- resubloss
- Loss
- kfoldloss
- oobLoss


Methods
Data and Inference
Applied Machine Learning
Applications
Data Analytics
Big Data Science

Data and Inference	Data Analytics
Data types Matrix algebra Correlation Confidence Hypothesis testing Time series Linear models	Weather forecasting Finance Electricity demand Renewable energy Biomedical engineering Telemedicine Healthcare Education
Applied Machine Learning	Big Data Science
Nonlinearity Regime switching Nonparametric approaches Forecasting, Regression, Classification Unsupervised learning: - PCA, ICA - Clustering techniques, k-means Supervised learning: - Neural networks, KNN - Decision Trees, Support Vector Machines Ensemble Approaches	Twitter Sentiment Analysis Google Trends, Wikipedia Weather observations and futures Activity monitoring Satellite applications Catastrophe modelling Early warning systems Big data for development

Q&A