Data Analytics

Course: 18-787

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

Spring 2023

ICT Center of Excellence
Carnegie Mellon University

Data Analytics

WEEK 2A

Course outline

Week	Lecture A	Lecture B
1	Data Analytics	Weather forecasting
2	Renewable energy	Wind energy
3	Solar energy	Demand forecasting
4	Risk	Extreme events
5	Health	Biomedicine
6	Early warning systems	Economic forecasting

Today’s Lecture

No.	Activity	Description	Time
1	Challenge	Supply variability	10
2	Discussion	Forecasting supply	10
3	Case study	Ensemble forecasts	10
4	Analysis	Electricity prices	20
5	Demo	Renewable policy	20
6	Q&A	Questions and feedback	10

A1Q7: Stepwise variable selection

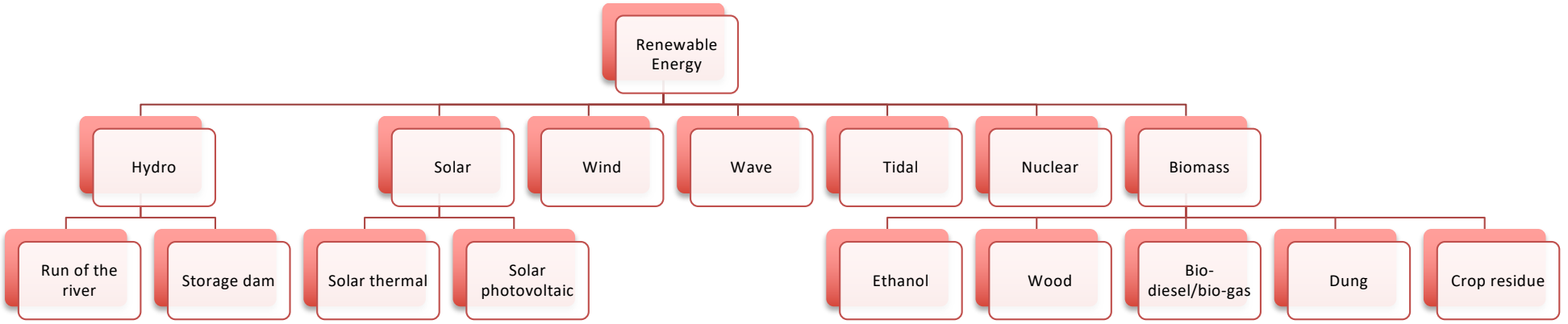
- When given a large number of candidate explanatory variables, it is necessary to select variables.
- Some variables have no predictive value, some are strongly correlated with others.
- There are different ways to select variables for constructing a model.
- Stepwise approaches this in an iterative fashion, starting with all variables and removing one at a time.

A1Q9: Day of the week

- Day of the week effects reflect that human behaviour
- Weekends different to weekdays
- This pattern can be represented using a dummy variable to pick up different levels for the different days of the week
- The statistical significance (p value) of each day can be tested

Renewable Energy

Renewable Sources



Renewable energy, often referred to as clean energy, comes from natural sources or processes that are constantly replenished.

Quiz

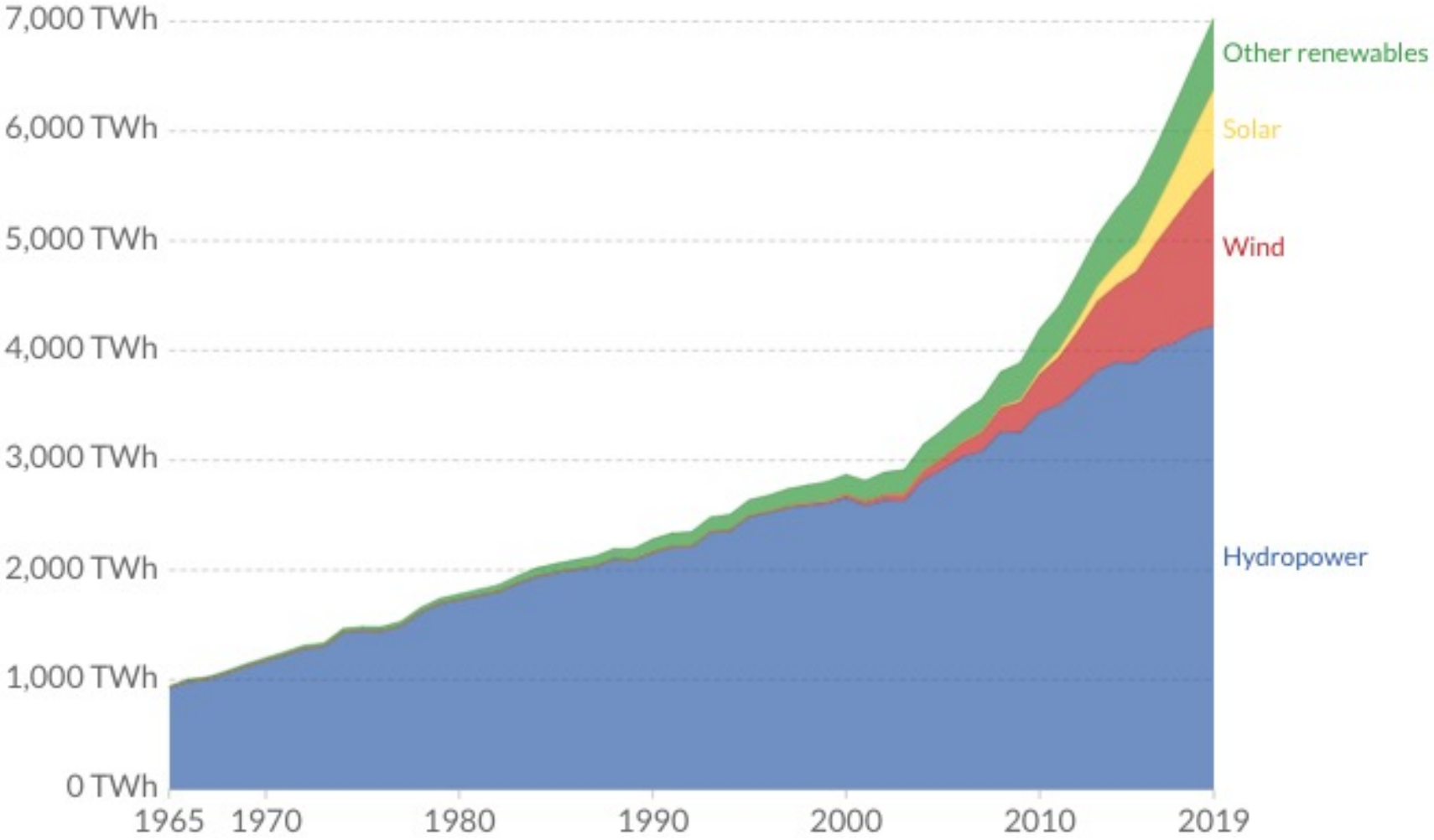
- In 2021, the share of renewables in global electricity generation reached:
  1) 10%
  2) 20%
  3) 30%
  4) 40%

- **Slido.com #78010**

Renewable energy generation, World

Our World in Data

⇄ Change country ☐ Relative



Source: BP Statistical Review of Global Energy

Note: 'Other renewables' refers to renewable sources including geothermal, biomass, waste, wave and tidal. Traditional biomass is not included.

Three Gorges, China (22.5 GW)

Source: www.wikipedia.com

Dinorwig, Wales (1.7 GW)

- Water is stored at a high altitude in Marchlyn Mawr reservoir and is discharged into Llyn Peris through the turbines during times of peak electricity demand.
- It is pumped back from Llyn Peris to Marchlyn Mawr during off-peak times. Although it uses more electricity to pump the water up than it generates on the way down, pumping is generally done at periods of low demand, when the energy is cheaper to consume.

Dinorwig, Wales (1.7 GW)

Source: www.wikipedia.com

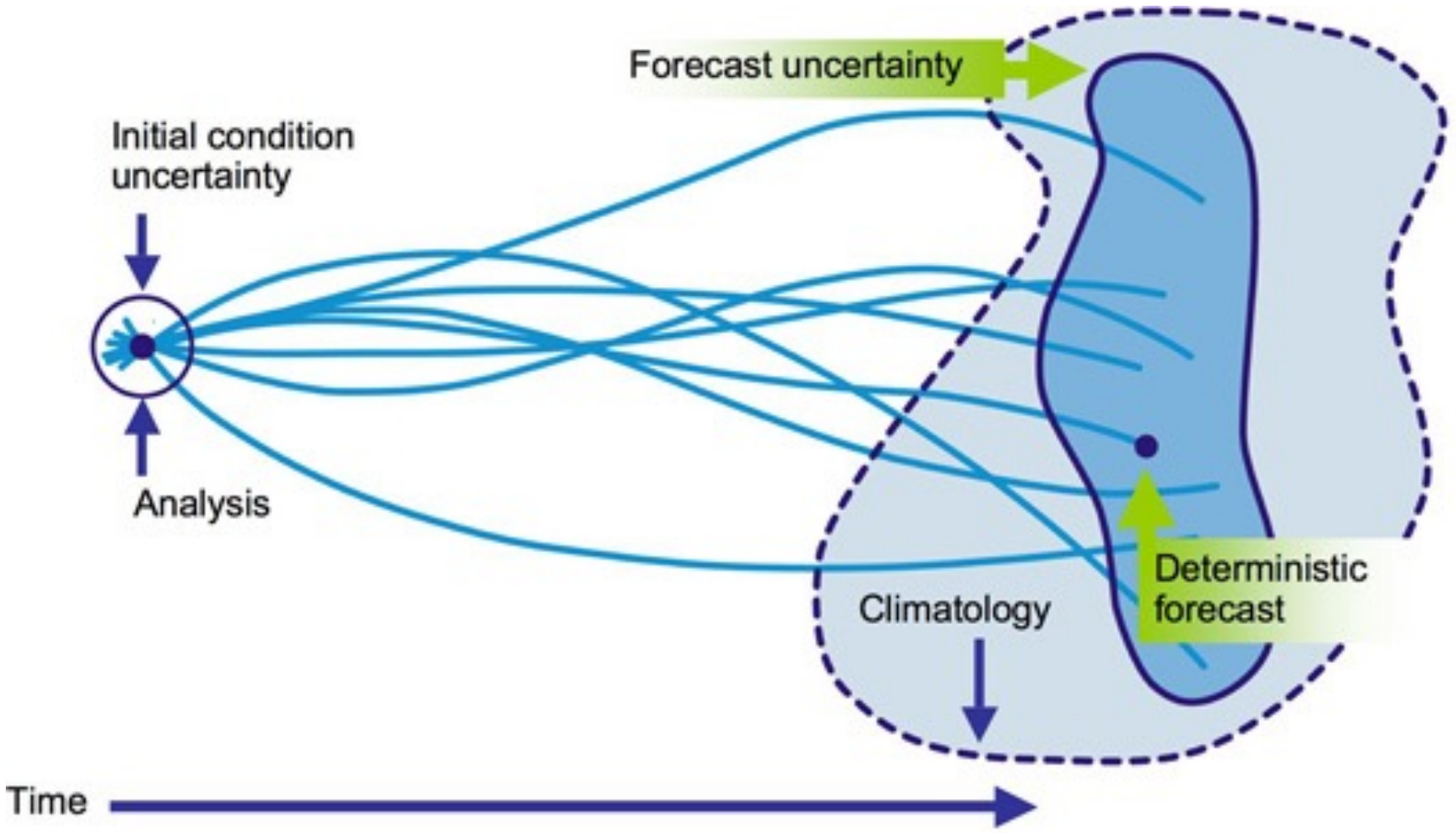
Uncertainty

- Uncertainty in the weather causes variability in the amount of power generated:
- Rainfall - hydropower
- Sunshine - solar energy
- Wind speed - wind power
- This uncertainty is best addressed using ensemble modelling approaches.

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 creation



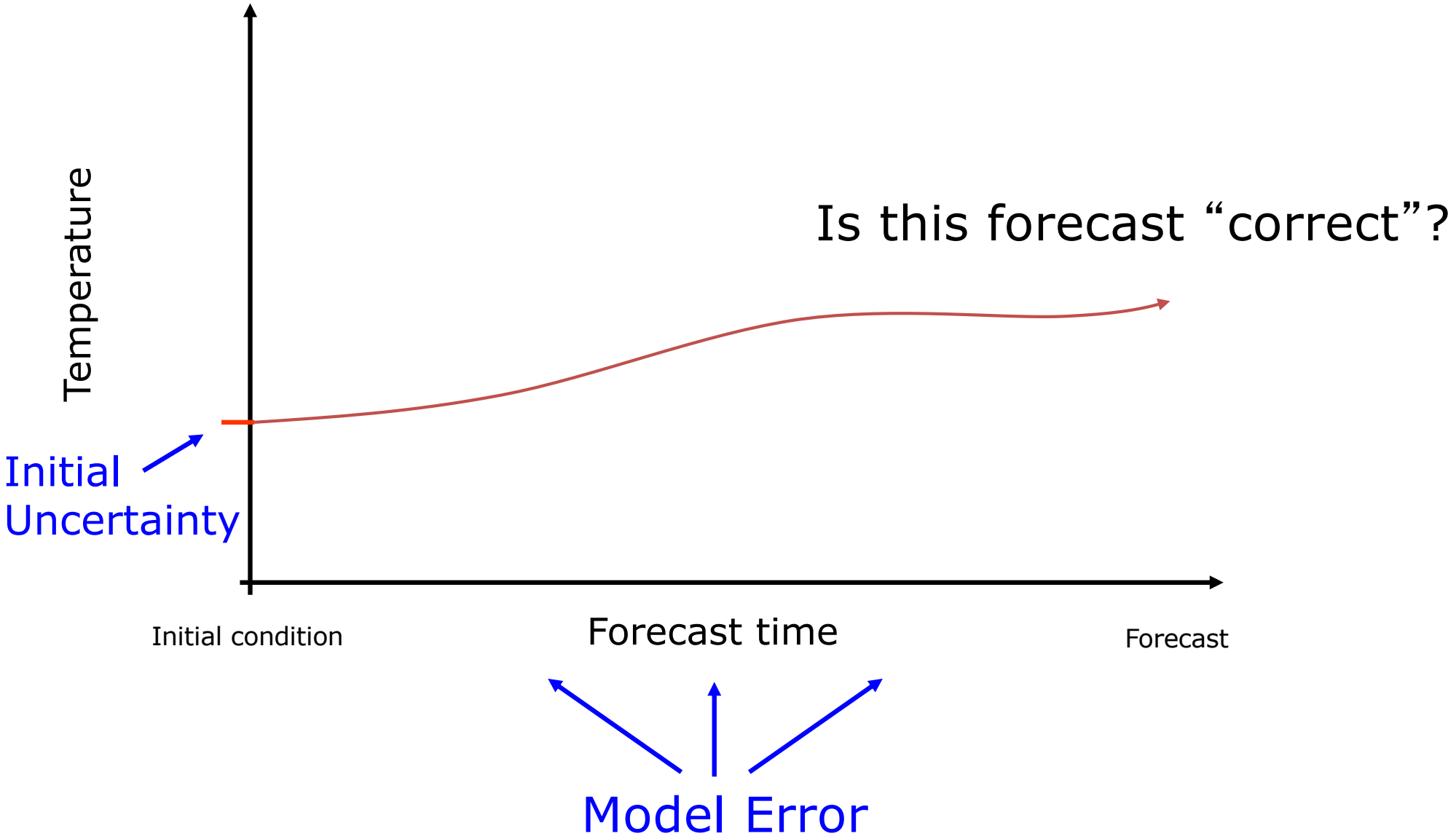
Source: www.easterbrook.ca

Ensemble benchmark

- When forecasting the temperature in New York in June, we could construct a benchmark ensemble known as “climatology” using:
  1) All daily temperatures over last 30 years
  2) All daily June temperatures over last 30 years

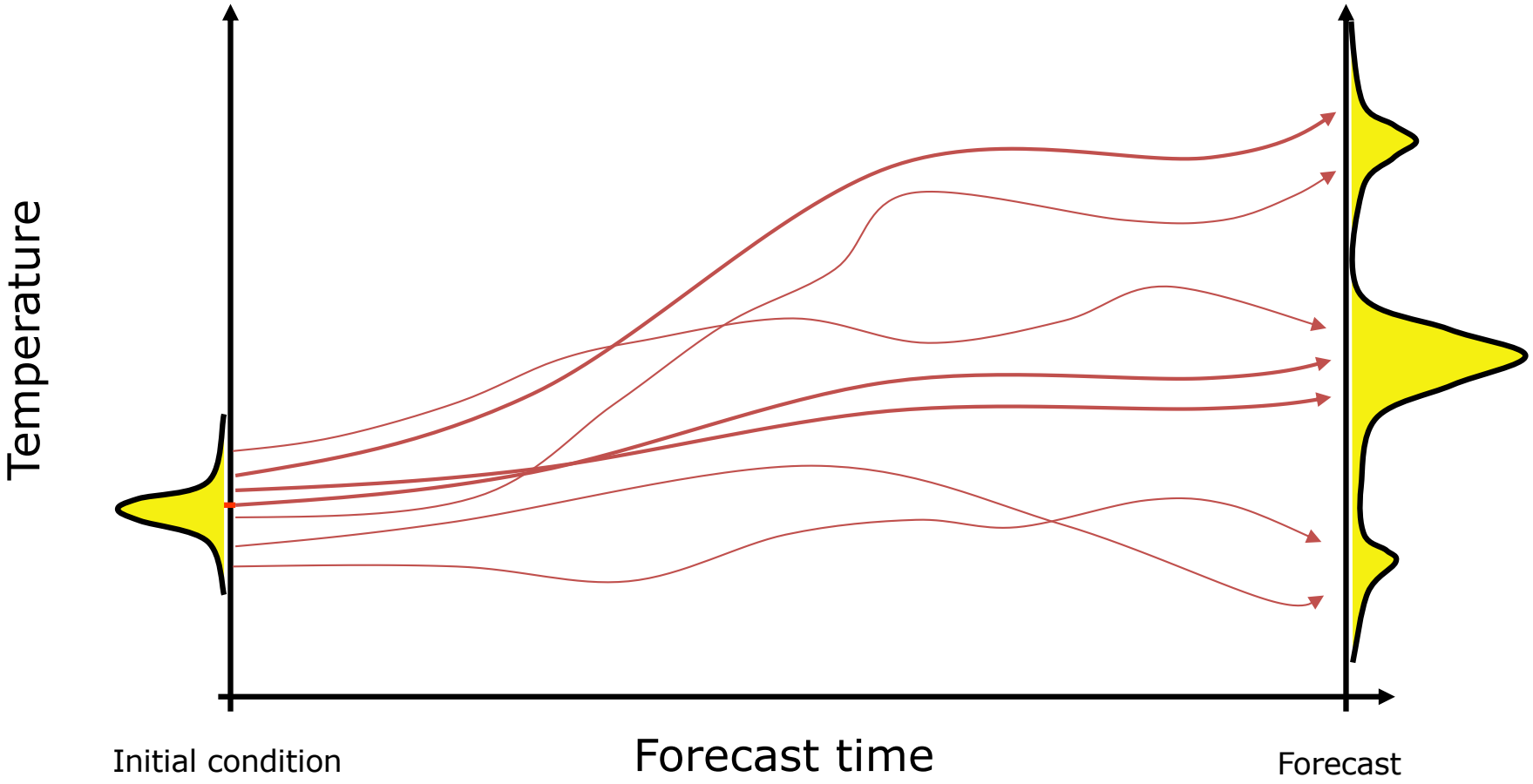
- **Slido.com #78010**

Deterministic Forecasting



Source: Renate Hagedorn, ECMWF

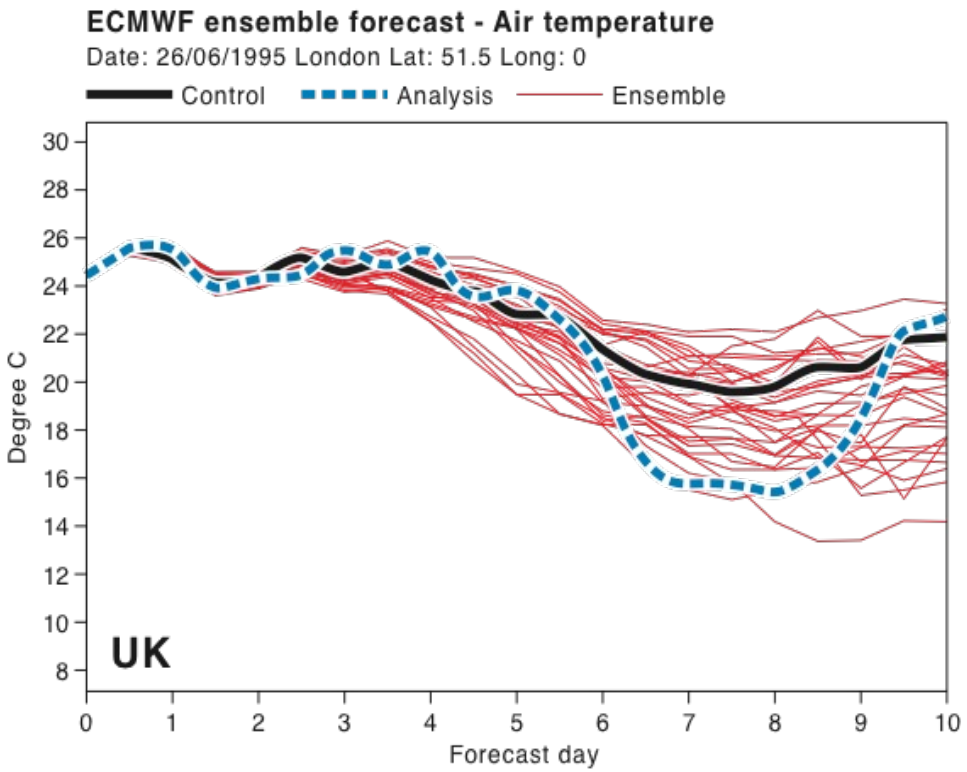
Ensemble Forecasting



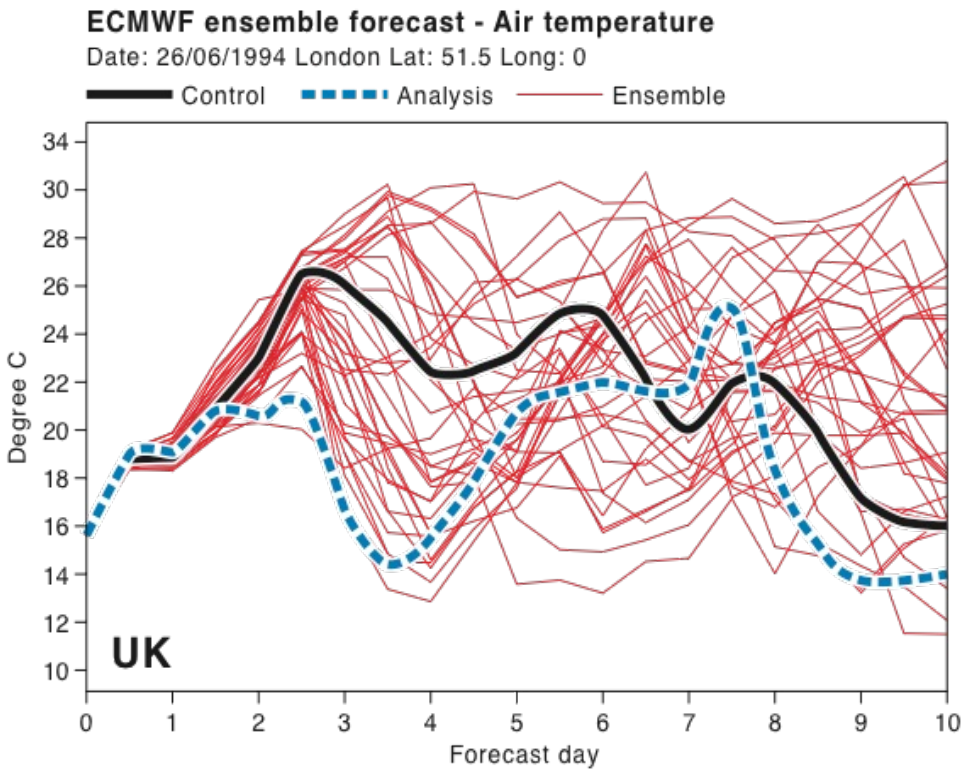
Description of weather prediction in terms of a Probability Density Function (PDF)

ECMWF Ensemble prediction

26th June 1995

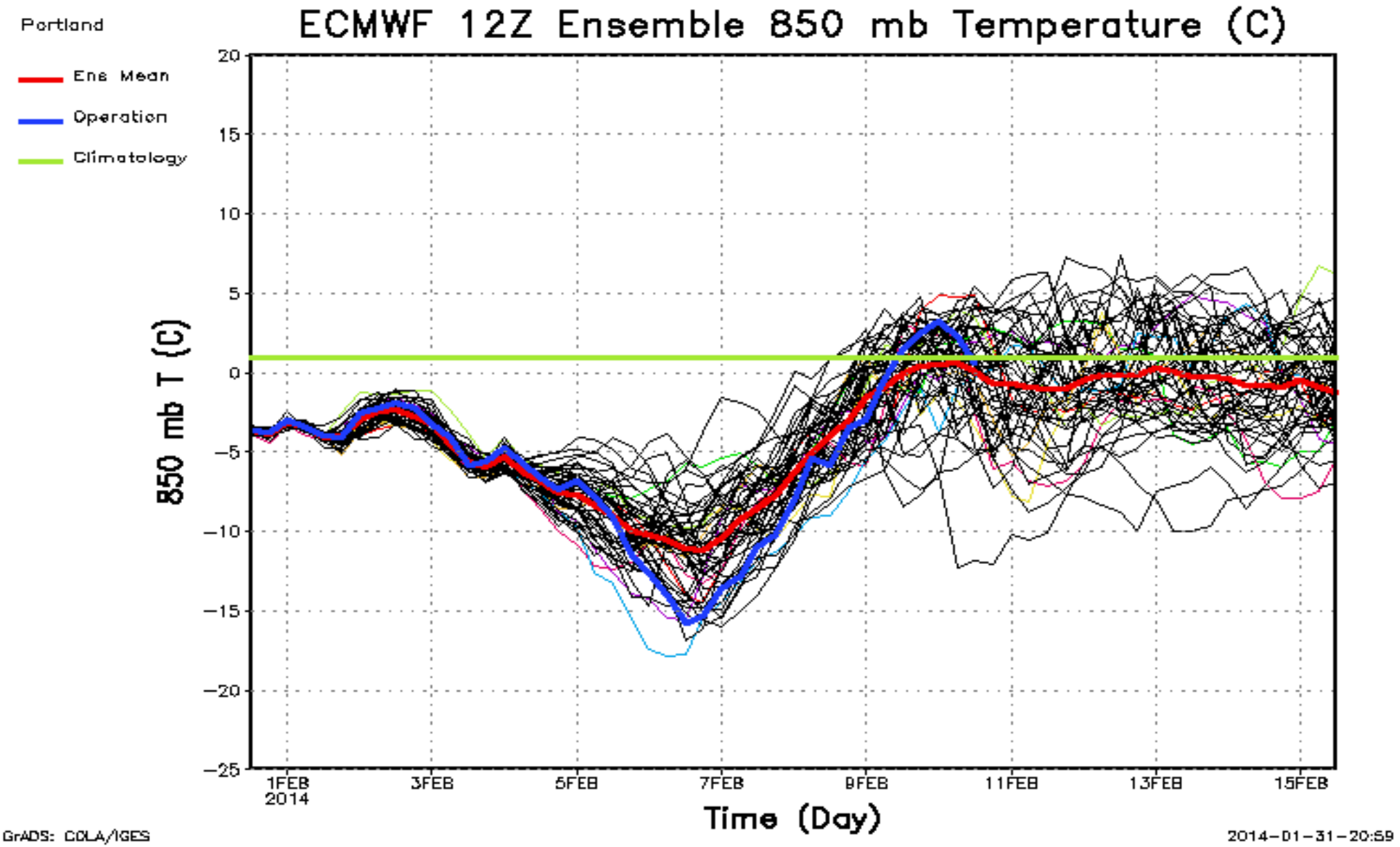


26th June 1994

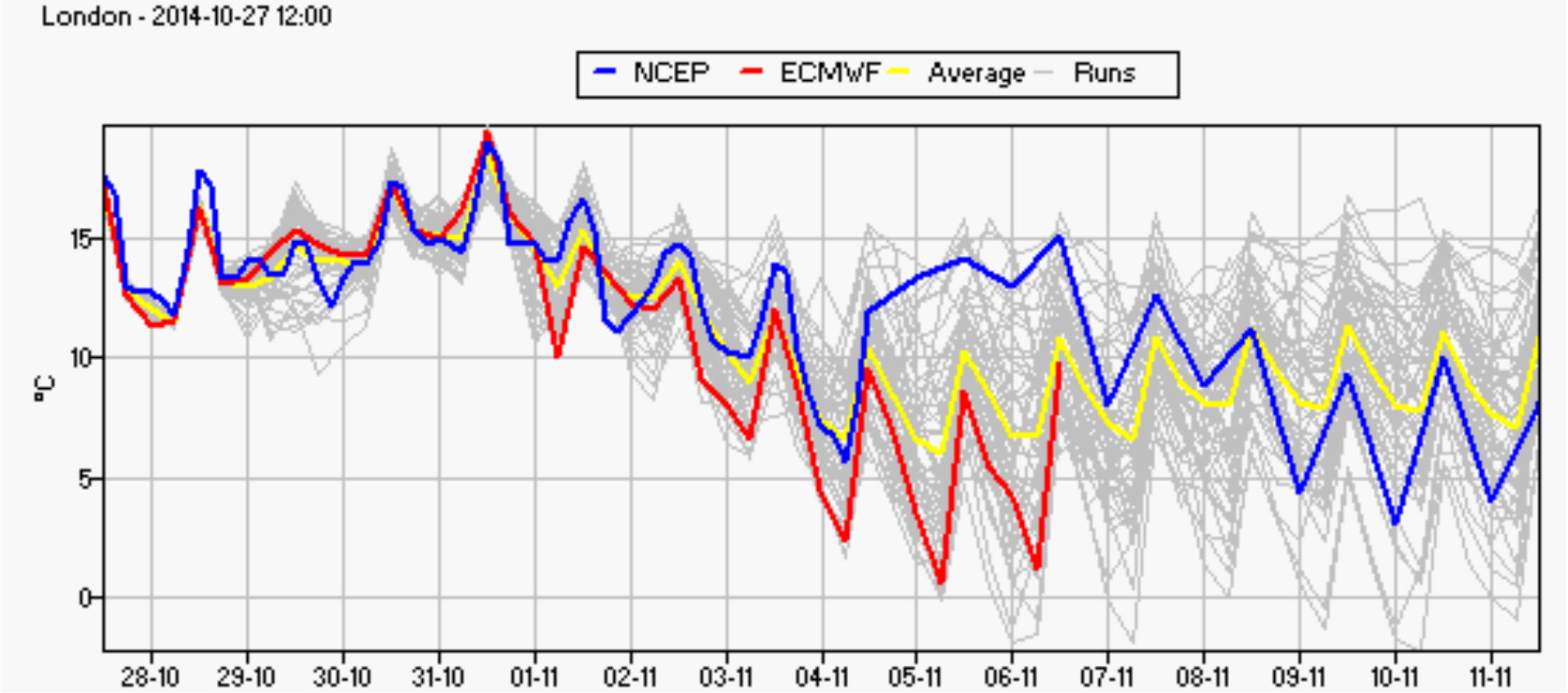


If the forecasts are coherent (small spread) the atmosphere is in a more predictable state than if the forecasts diverge (large spread)

ECMWF ensemble

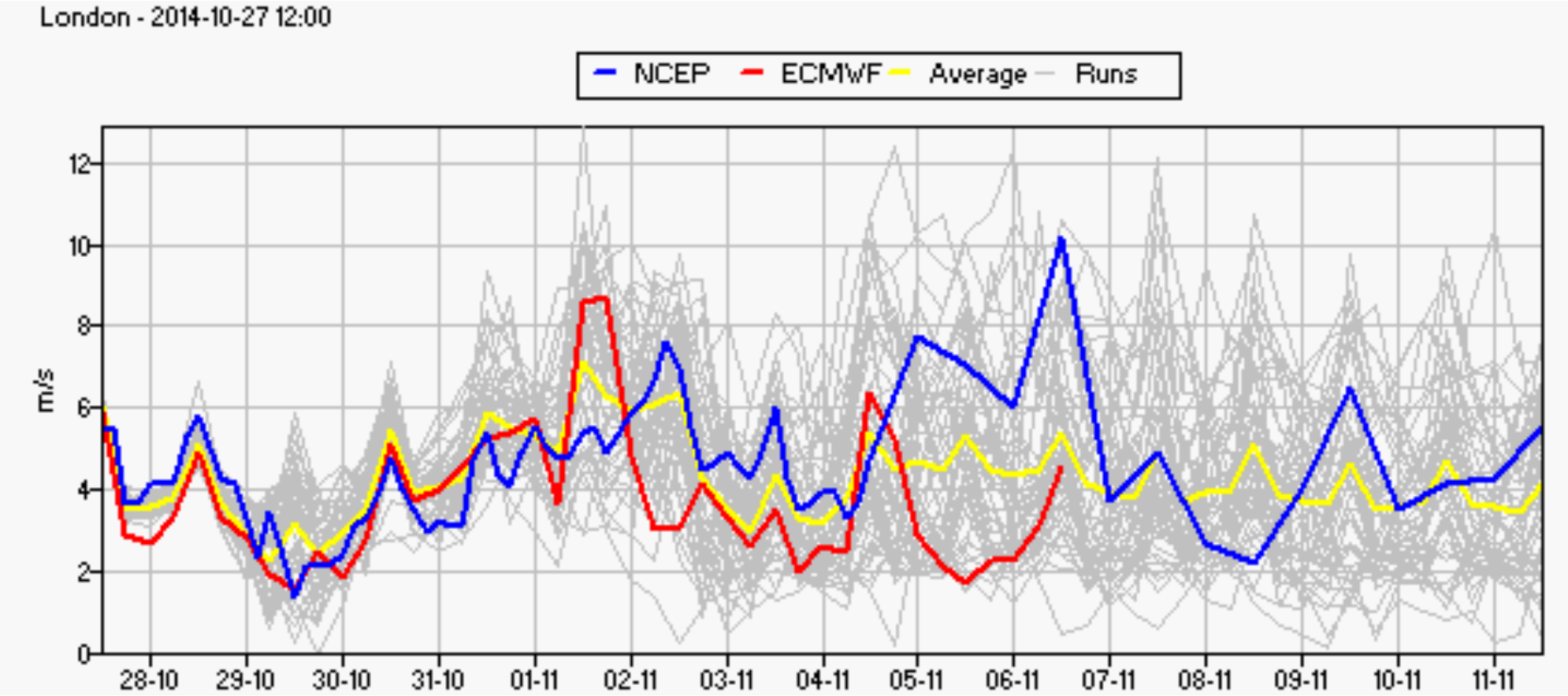


Multiple Temperature Ensembles



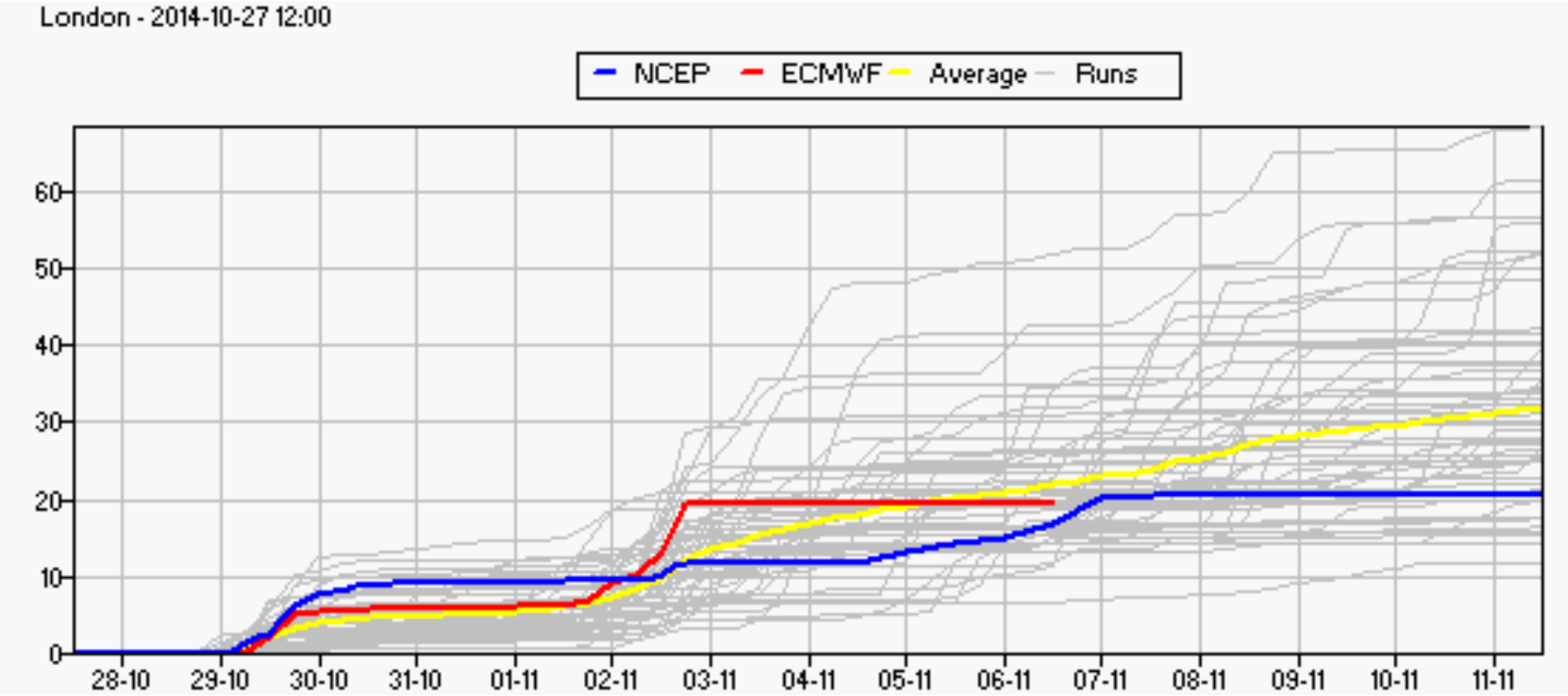
Source: www.weathercast.co.uk

Multiple Wind Ensembles



Source: www.weathercast.co.uk

Multiple Rainfall Ensembles



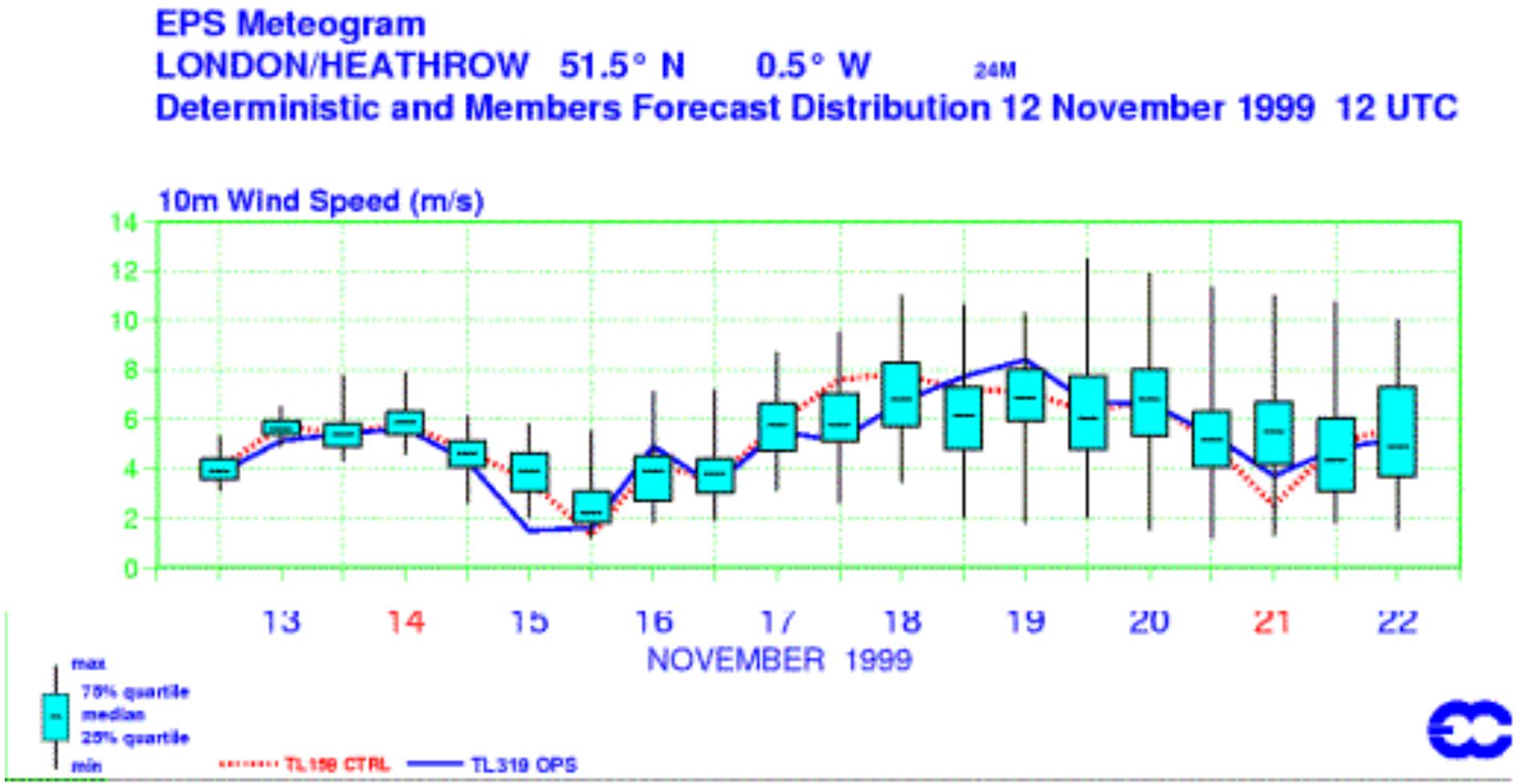
Source: www.weathercast.co.uk

Ensemble visualizations

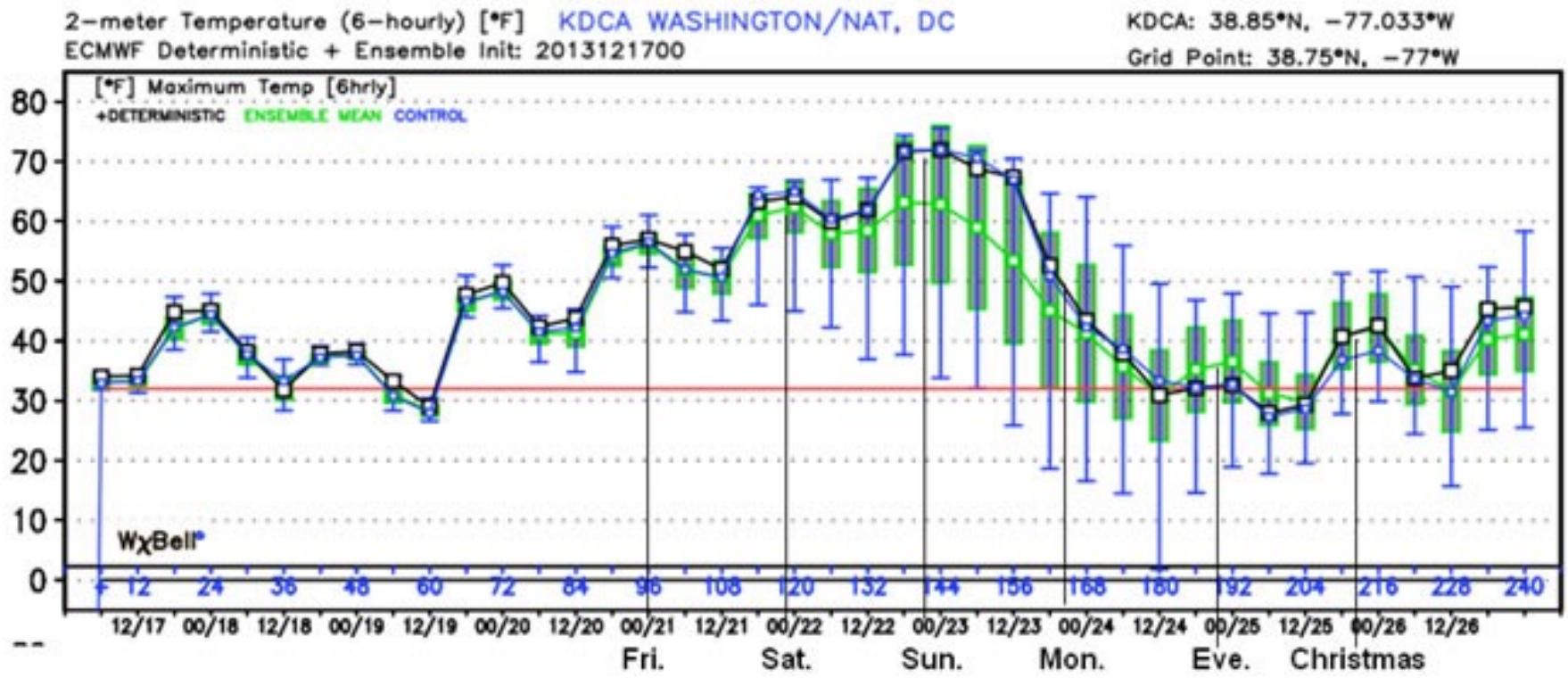
- Ensemble forecasts can be visualized using:
  1) Spaghetti diagrams
  2) Histograms
  3) Box plots
  4) All of the above

- **Slido.com #78010**

EPS Meteogram



ECMWF ensemble



Source: www.washingtonpost.com; www.weatherbell.com

Evaluation of ensembles

- Modelling
  - Numerical weather prediction models
  - Statistical time series models
- Nonlinearity, model error, uncertainty and risk assessment
- Ensemble forecasting, density forecasting
- Benchmarks & statistical testing
- Renewable energy forecasting

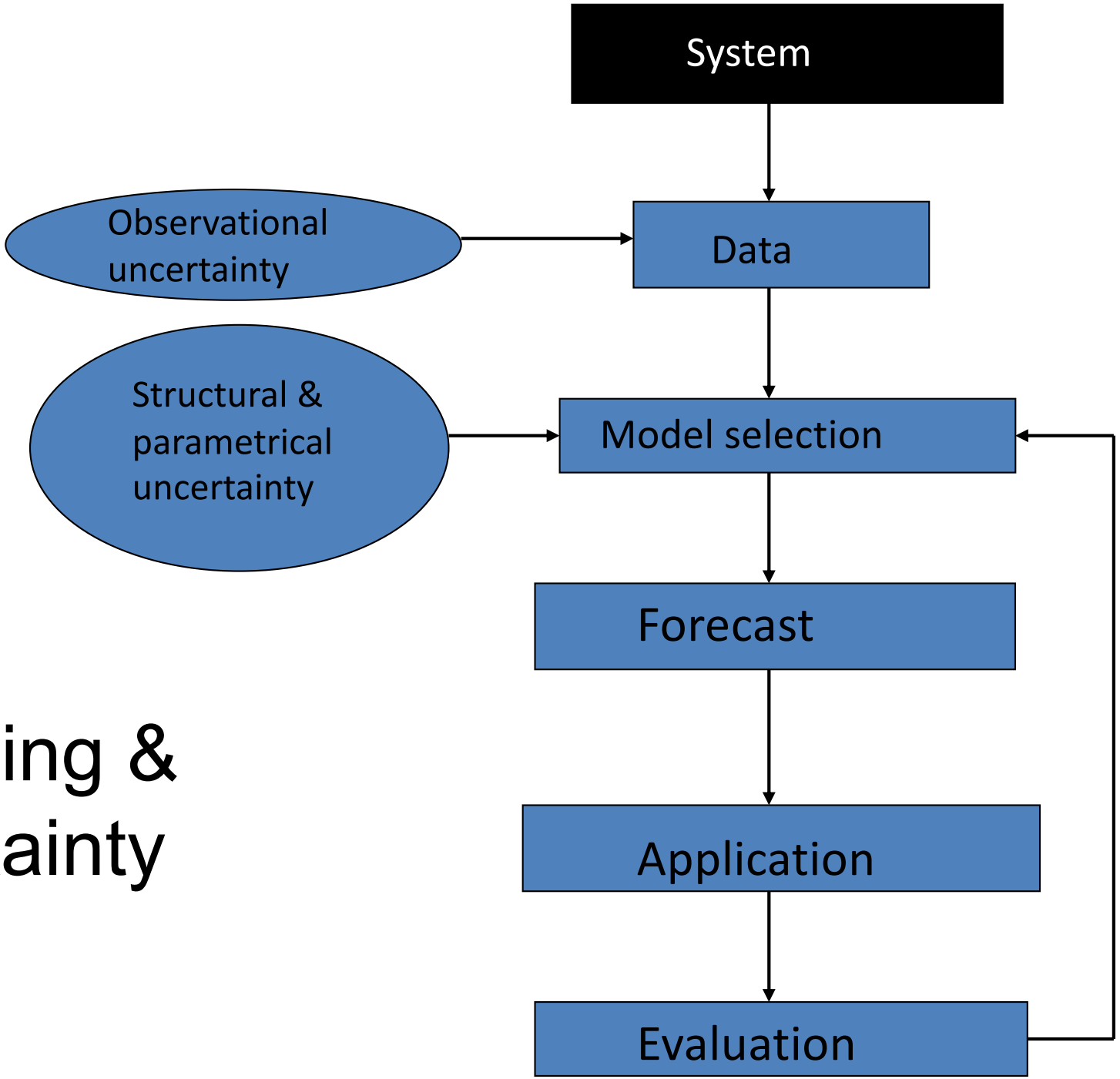
Data-based principled (DBP) modelling

- Statistical data-based modelling (DBM)
  - Time series modelling (ARIMA, TAR, Nonlinear models, state space models, Machine learning, Neural networks, Genetic algorithms)
- Mathematical modelling
  - Incorporation of prior knowledge
  - Newtons Law (mechanics), Navier-Stokes (fluids), Maxwell's Laws (electro-magnetism)
  - Conservation laws of nature (mass, momentum, force)
  - System constraints (non-negative distributions)
- DBP modelling aims to combine knowledge obtained from first principles, such as conservation laws of nature and system constraints, with information extracted from existing databases and real-time observations.

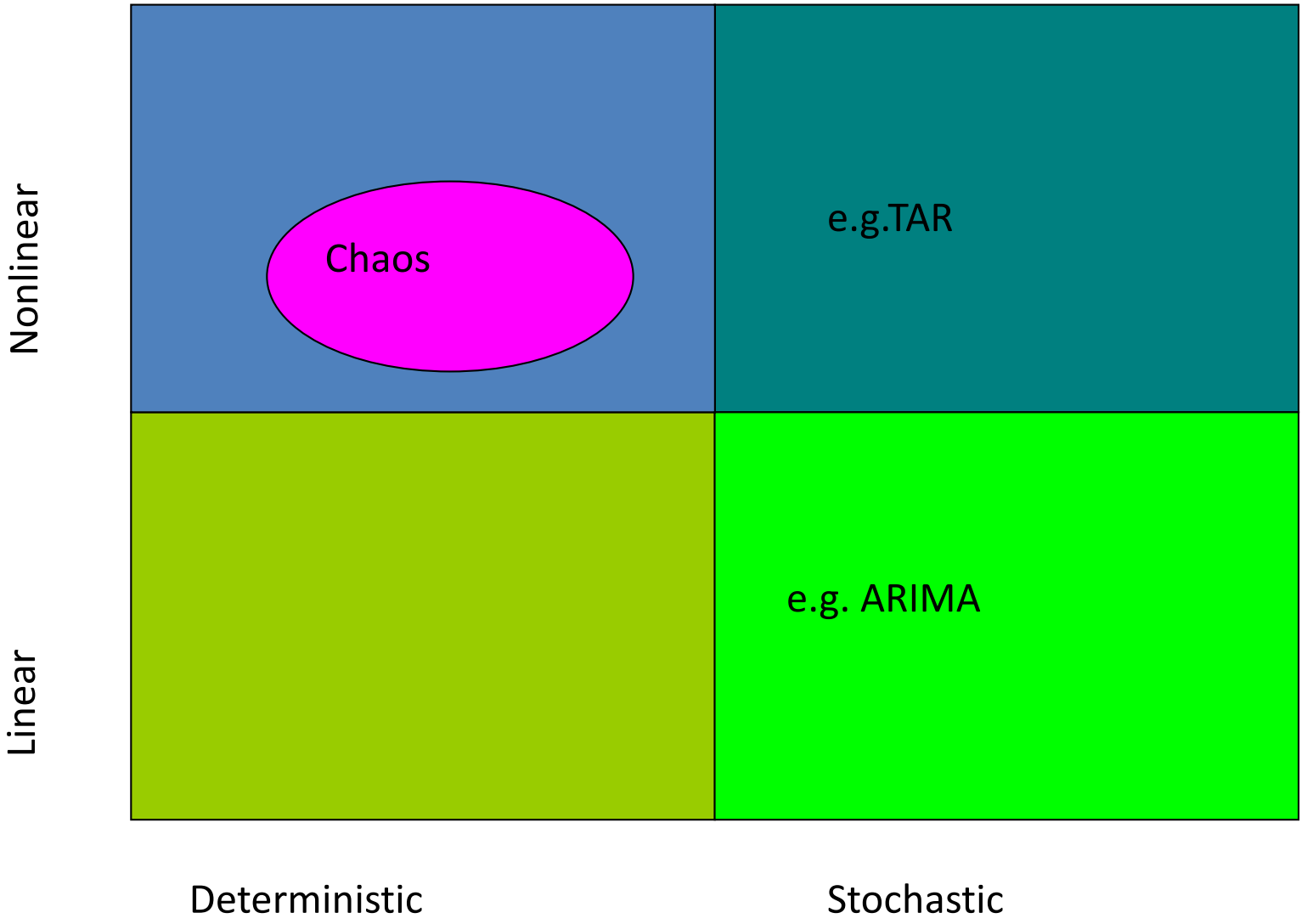
All models are wrong!

- “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 non-stationary: structural breaks, external influences
- Any useful model should account for the fact that the unknowable is likely to happen

Modelling & uncertainty

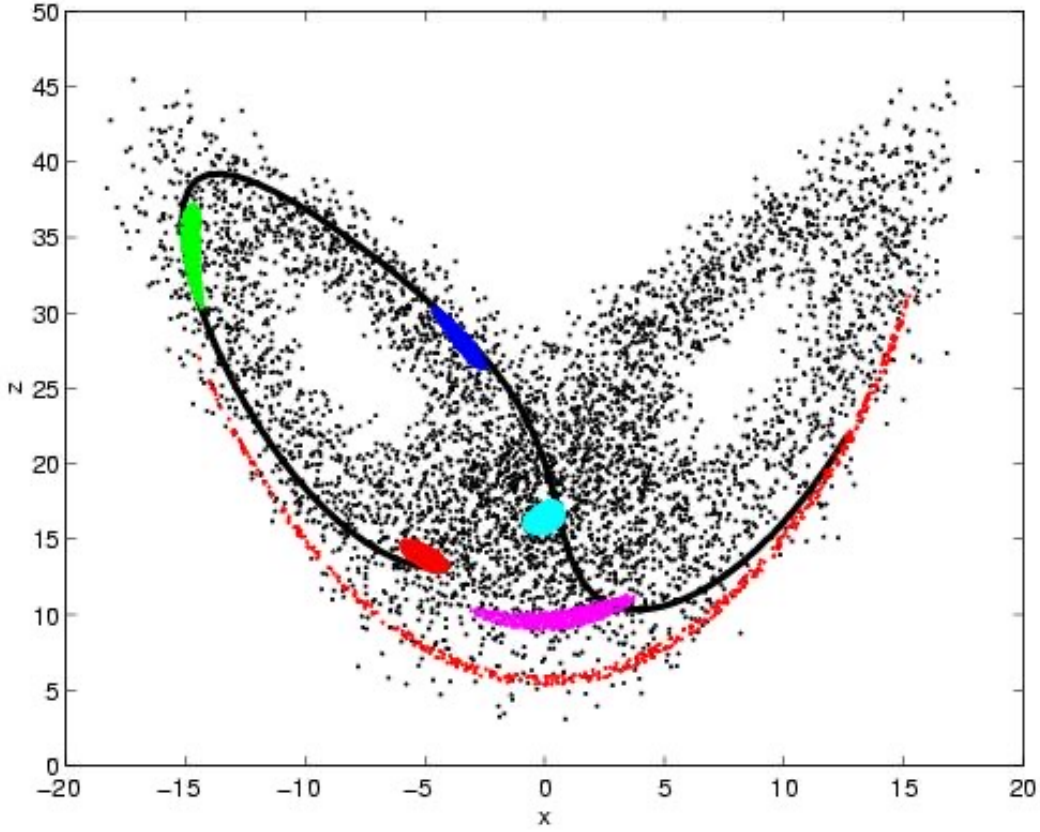


Model class and complexity

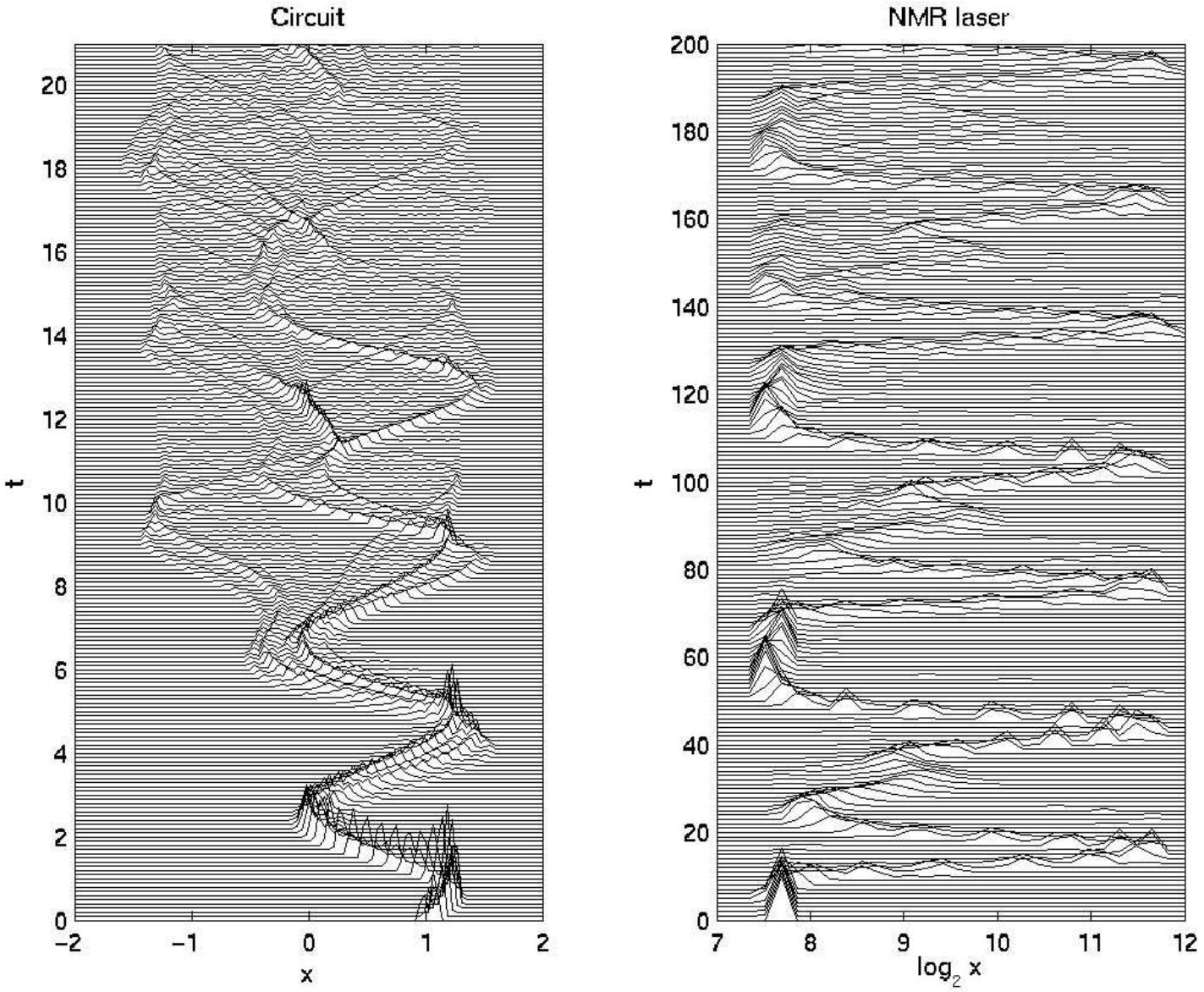


Chaos in the Lorenz system

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

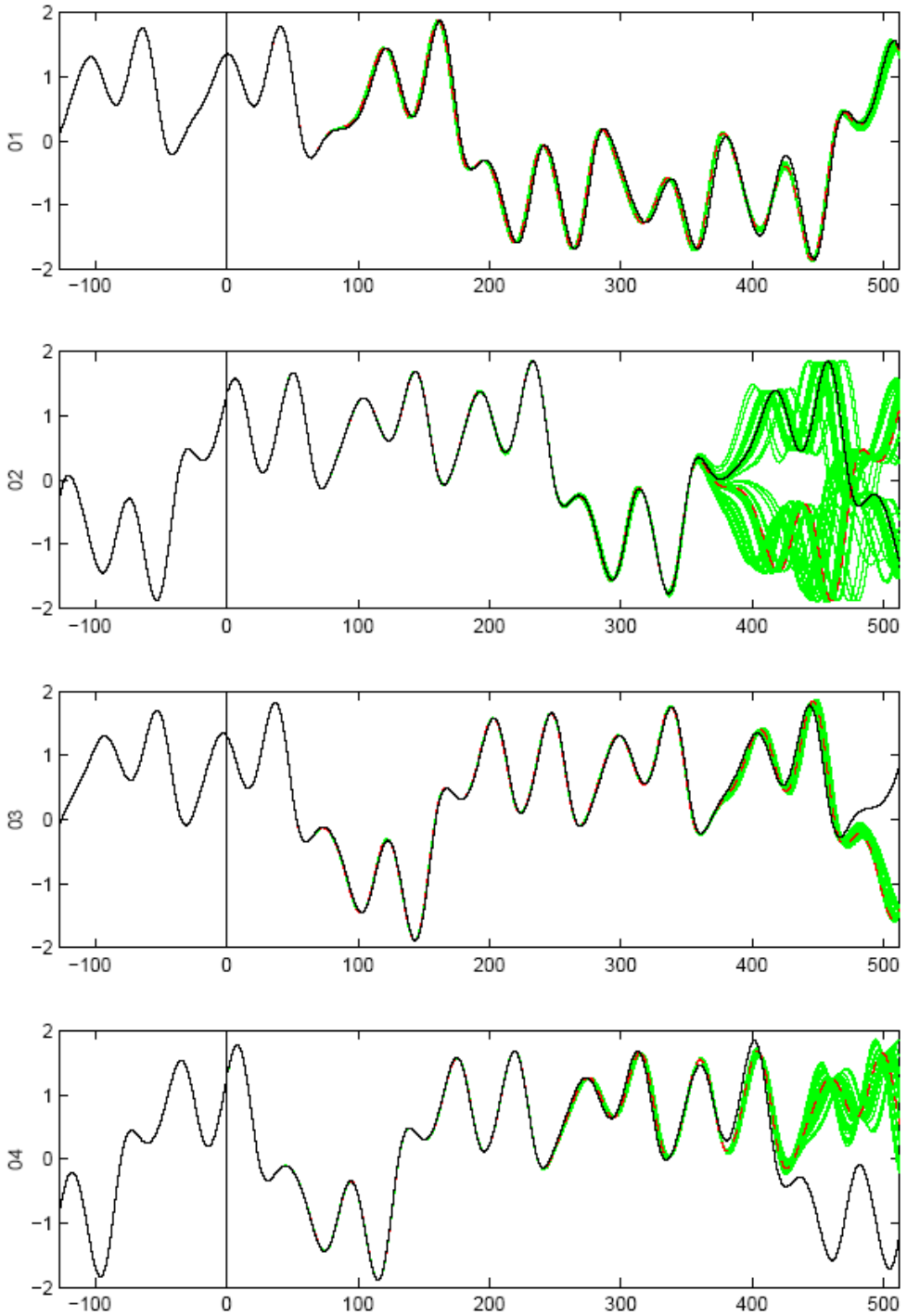


Uncertainty evolution

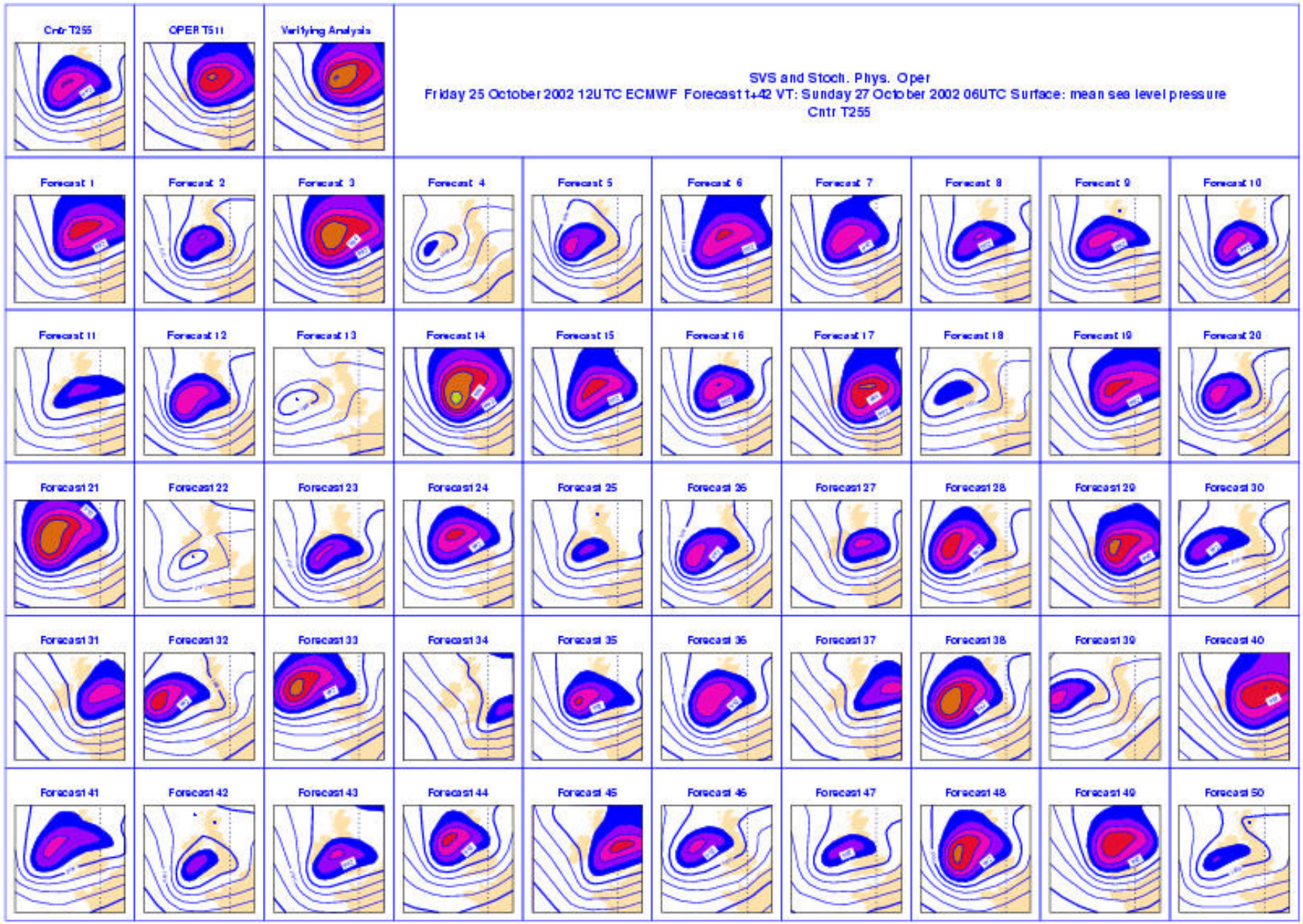


- Ensemble forecasts
- 128 points for each initial condition
- Forecast lead time of 512 steps
- Observations (black)
- Point forecast (red)
- Ensemble (green)
- Forecast quality depends on the initial condition
- Model inadequacy test

  McSharry & Smith. Physica D (2004)



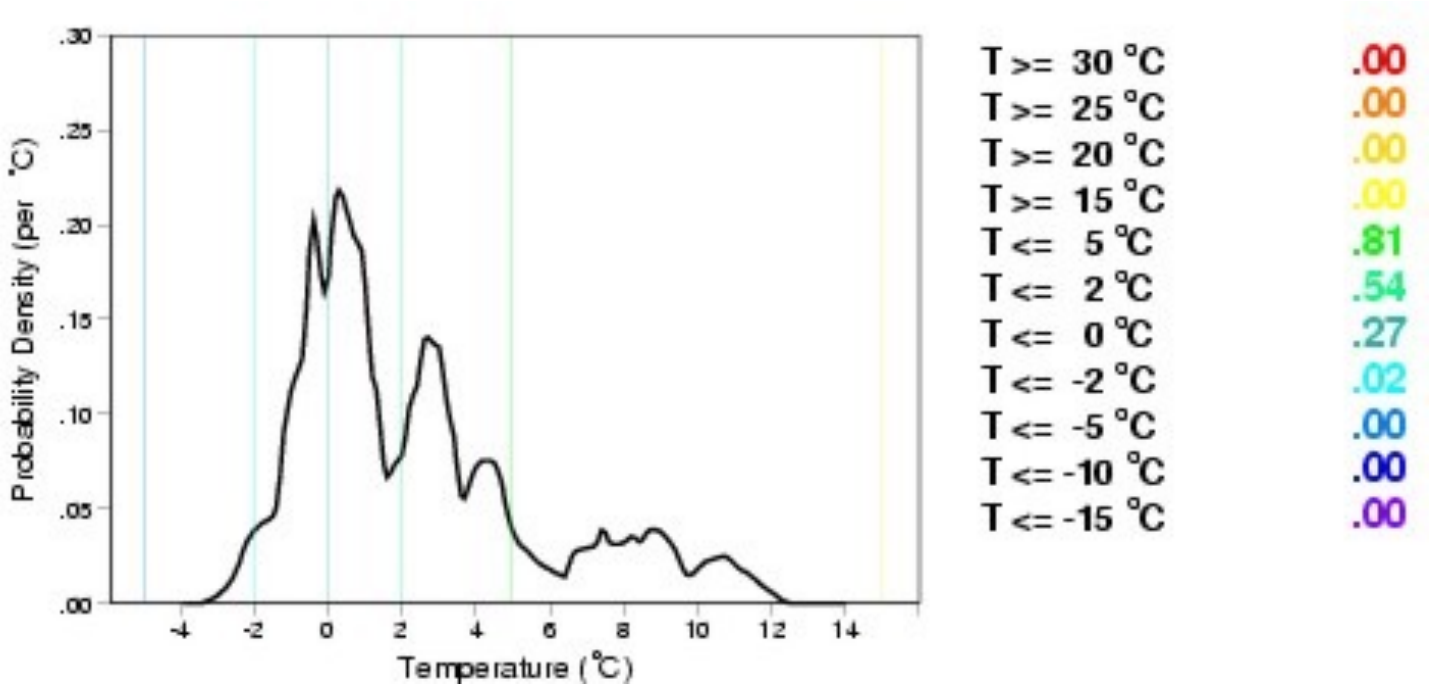
ECMWF 42 hour forecast for the October Storm of 2002



Each simulation looks physically reasonable, and most have storms…

Source: Lenny Smith

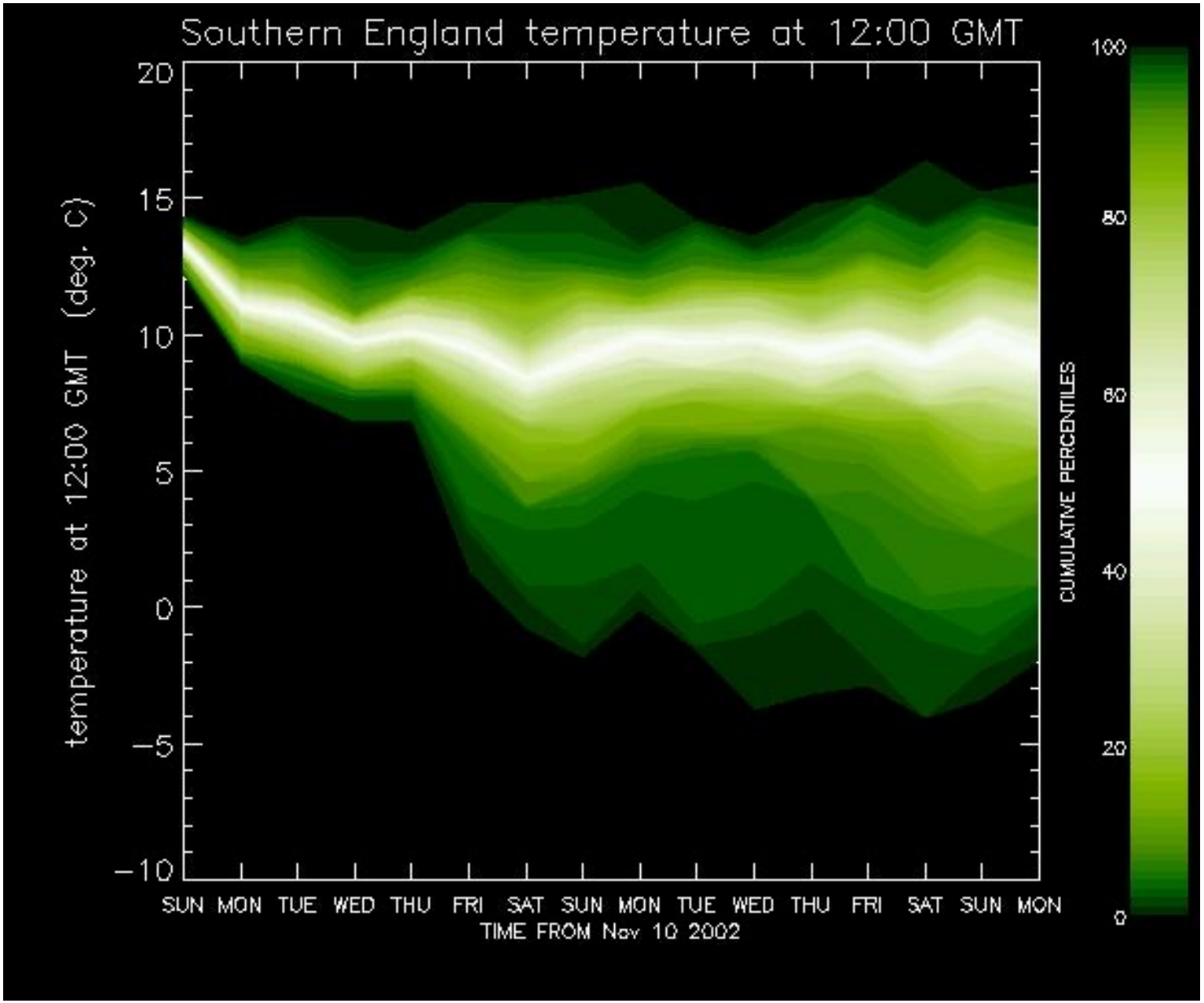
Temperature density forecast



- 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:



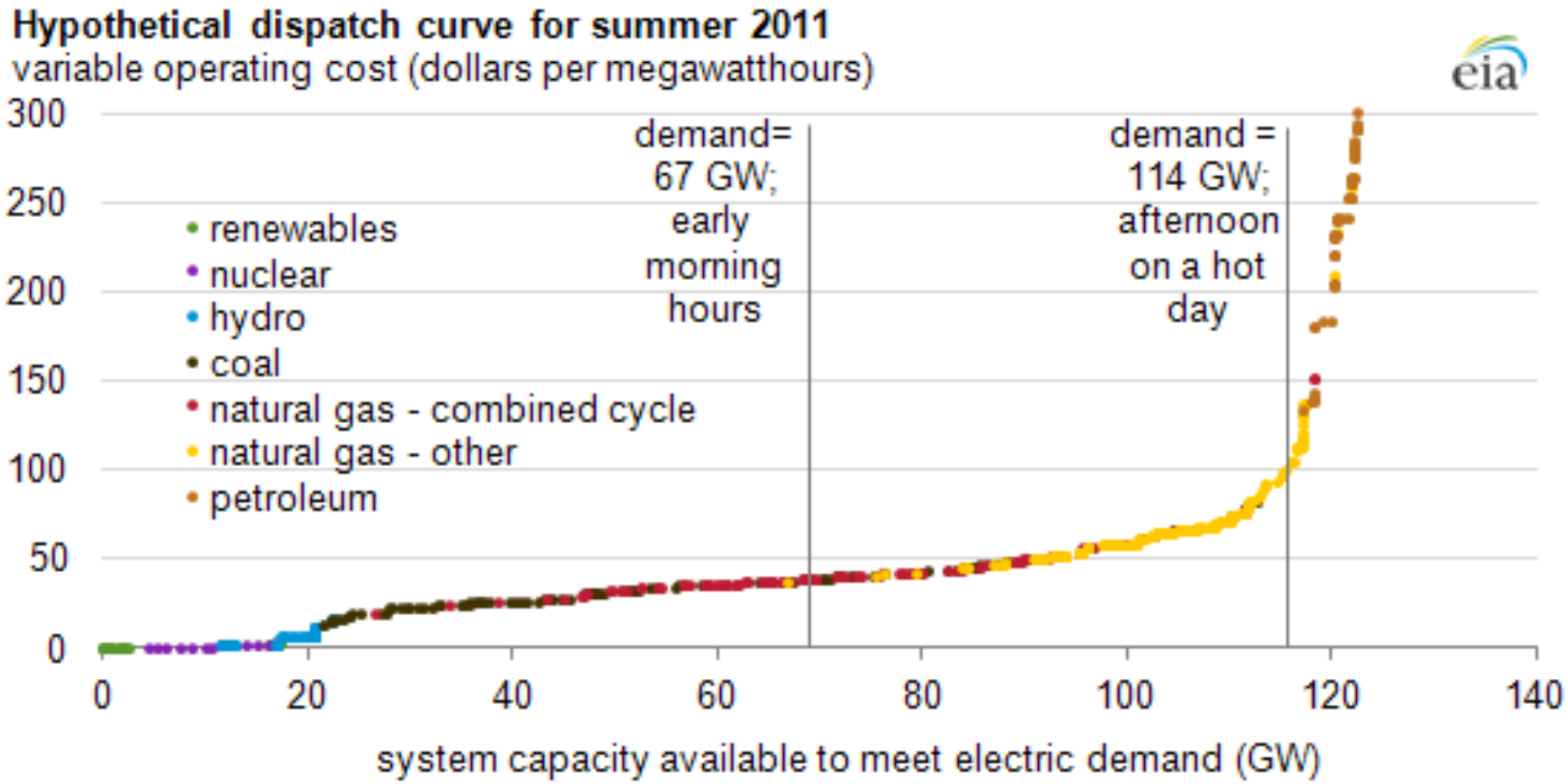
www.dime.lse.ac.uk

Ensemble visualizations Poll

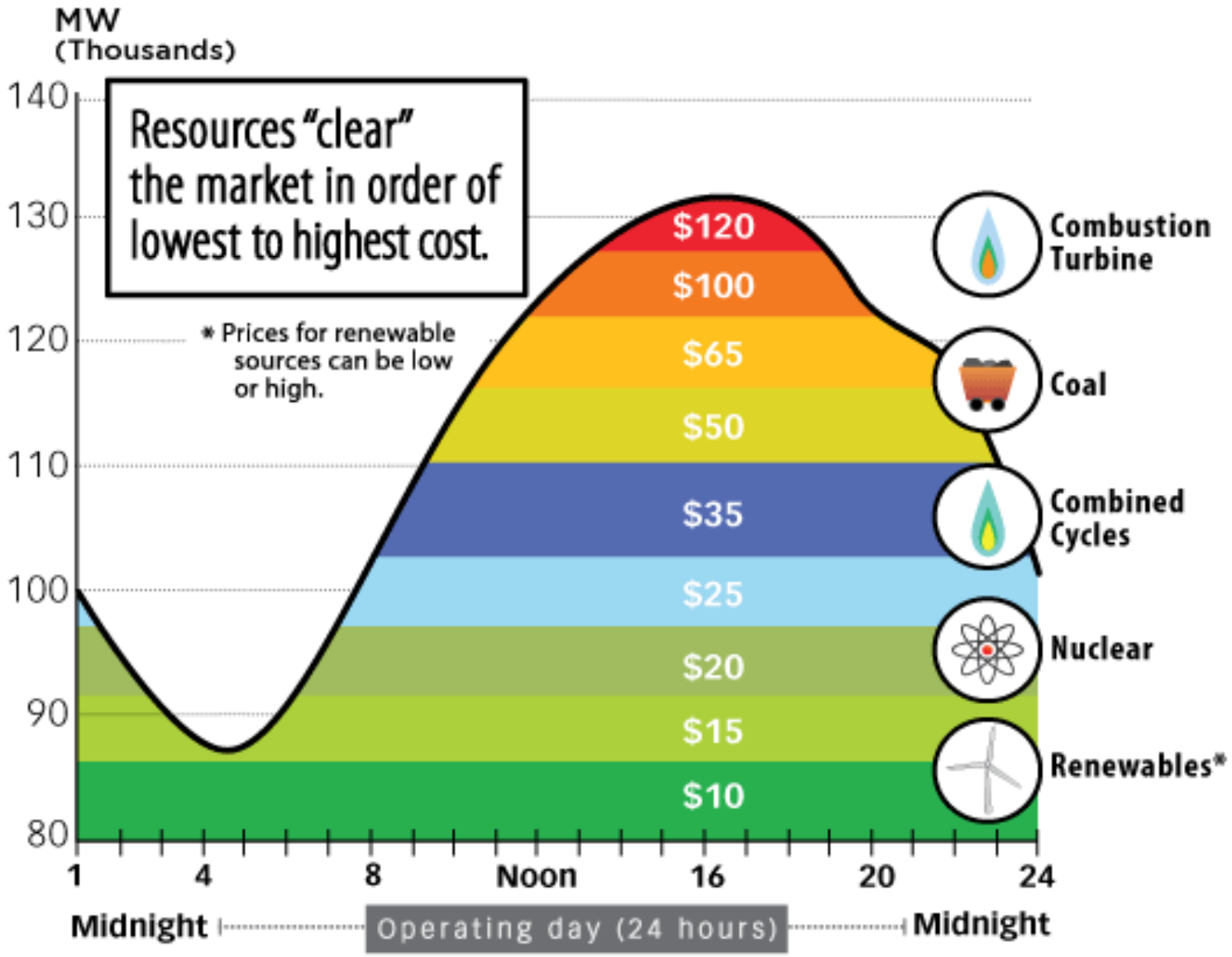
- Select your favorite way of visualizing ensemble forecasts:
  1) Spaghetti diagrams
  2) Histograms
  3) Box plots

- **Slido.com #78010**

Dispatch curve

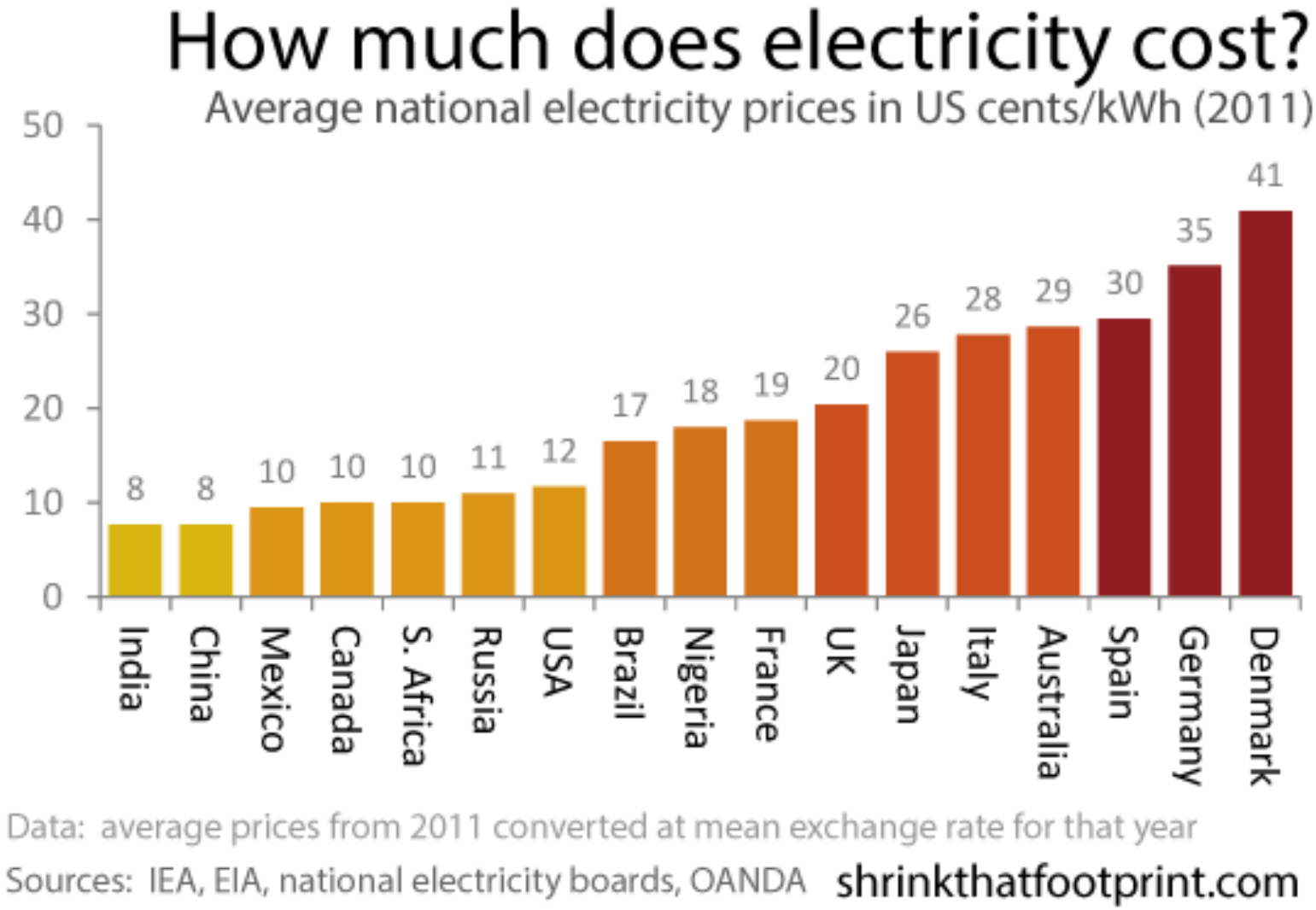


Daily demand profile



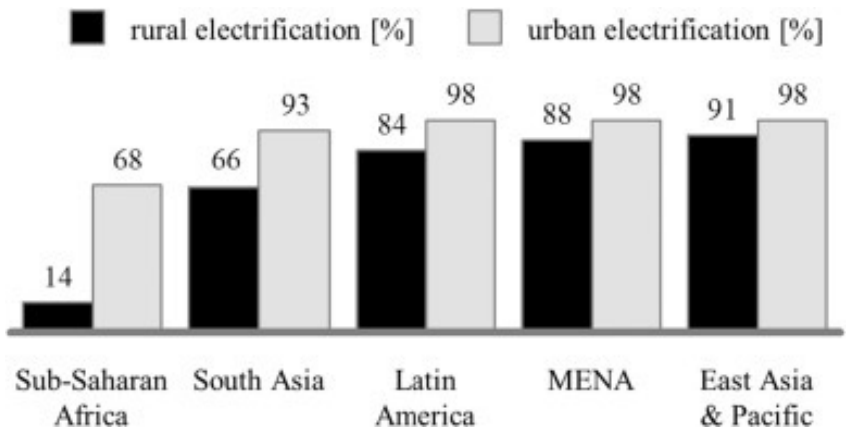
Source: [PJM]

Cost of electricity



Rural electrification: off-grid solutions

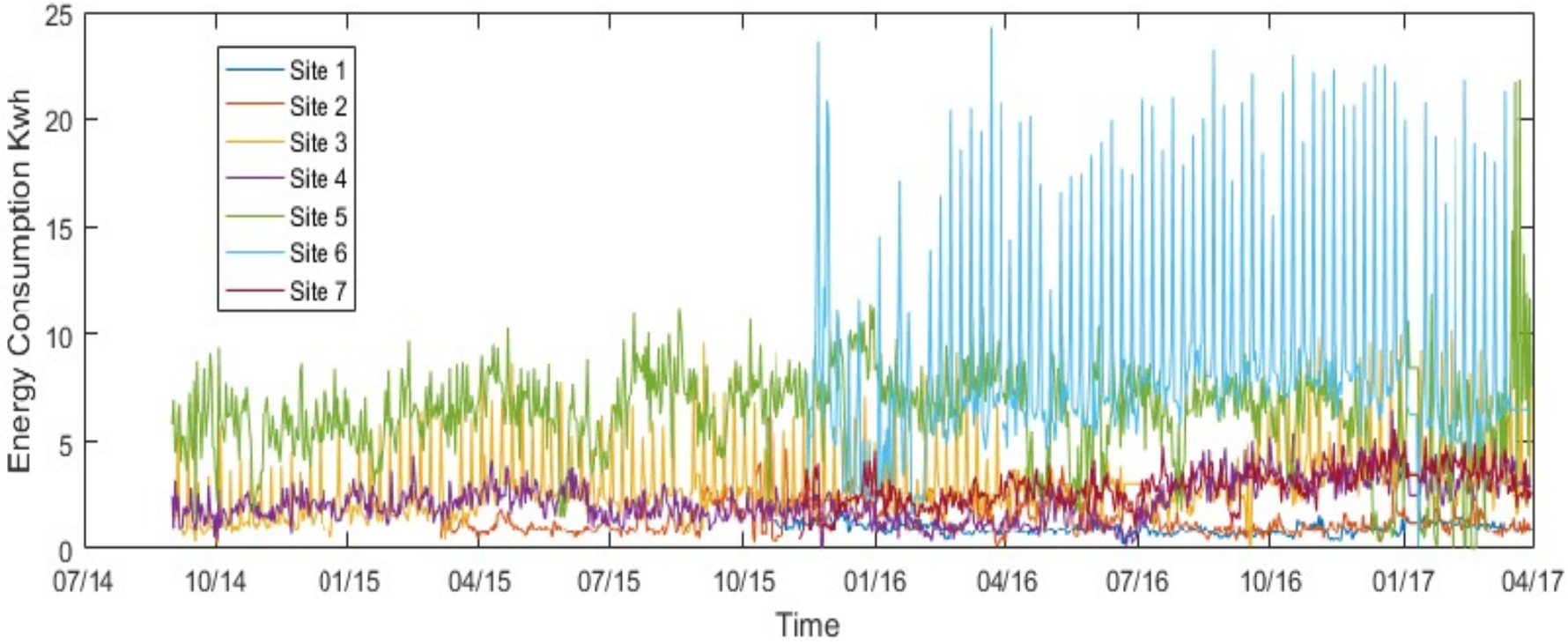
- In Sub-Saharan Africa nearly 600 million people or about 70% of the population lives without electricity (IEA).
- Sub-Saharan Africa's rural electrification rate is 14% (Trotter, 2016).
- The International Energy Agency foresees that microgrids and off-grid solutions will have a huge role in providing energy for 70% of all rural populations in developing countries.



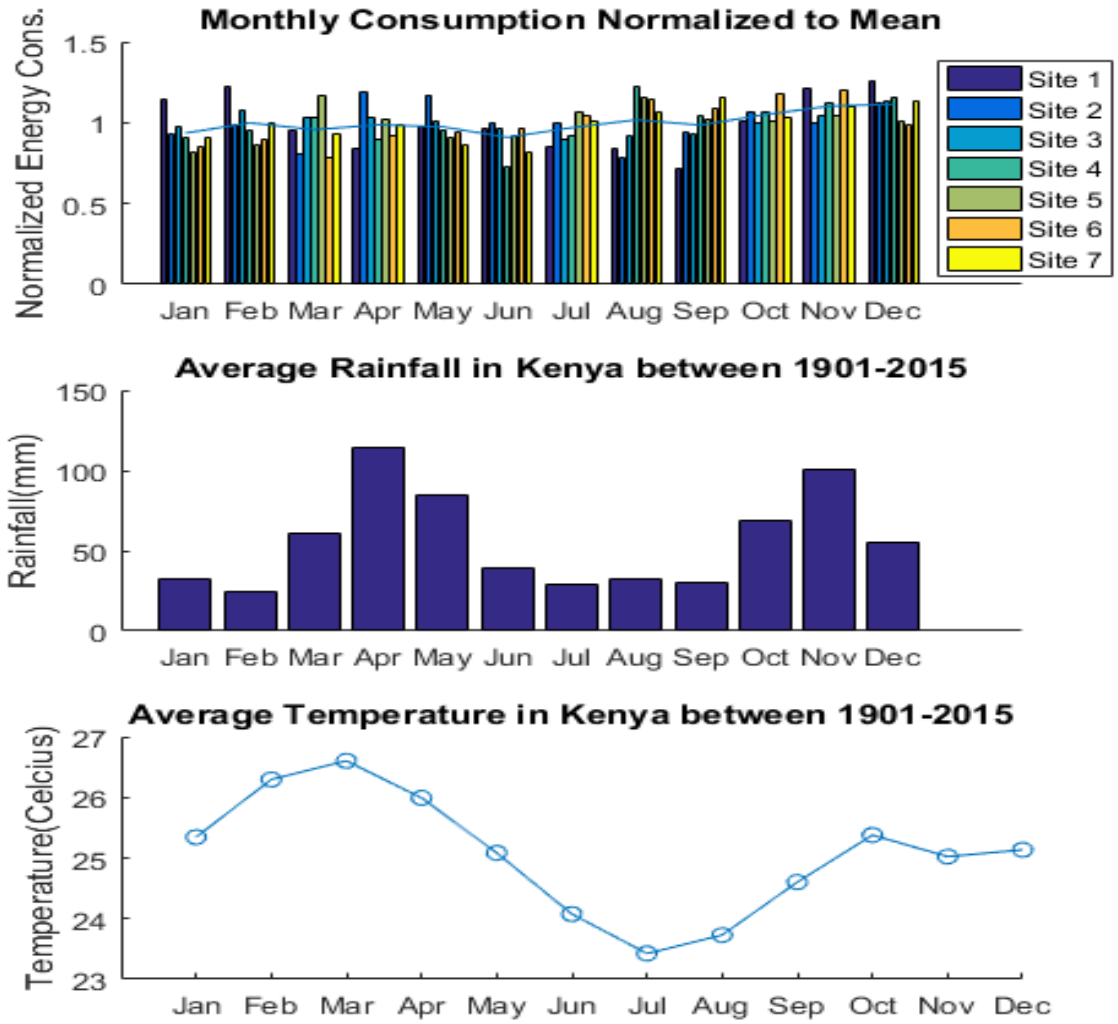
B. Otieno & P.E. McSharry (2018). Customer segregation of an East African Microgrid. IEEE PES&IAS PowerAfrica Conference, Cape Town, South Africa.
F. Otieno & P.E. McSharry (2018). Forecasting energy demand for microgrids over multiple horizons. IEEE PES&IAS PowerAfrica Conference, Cape Town, South Africa.

Independent studies with Fred Otieno and Brian Otieno.

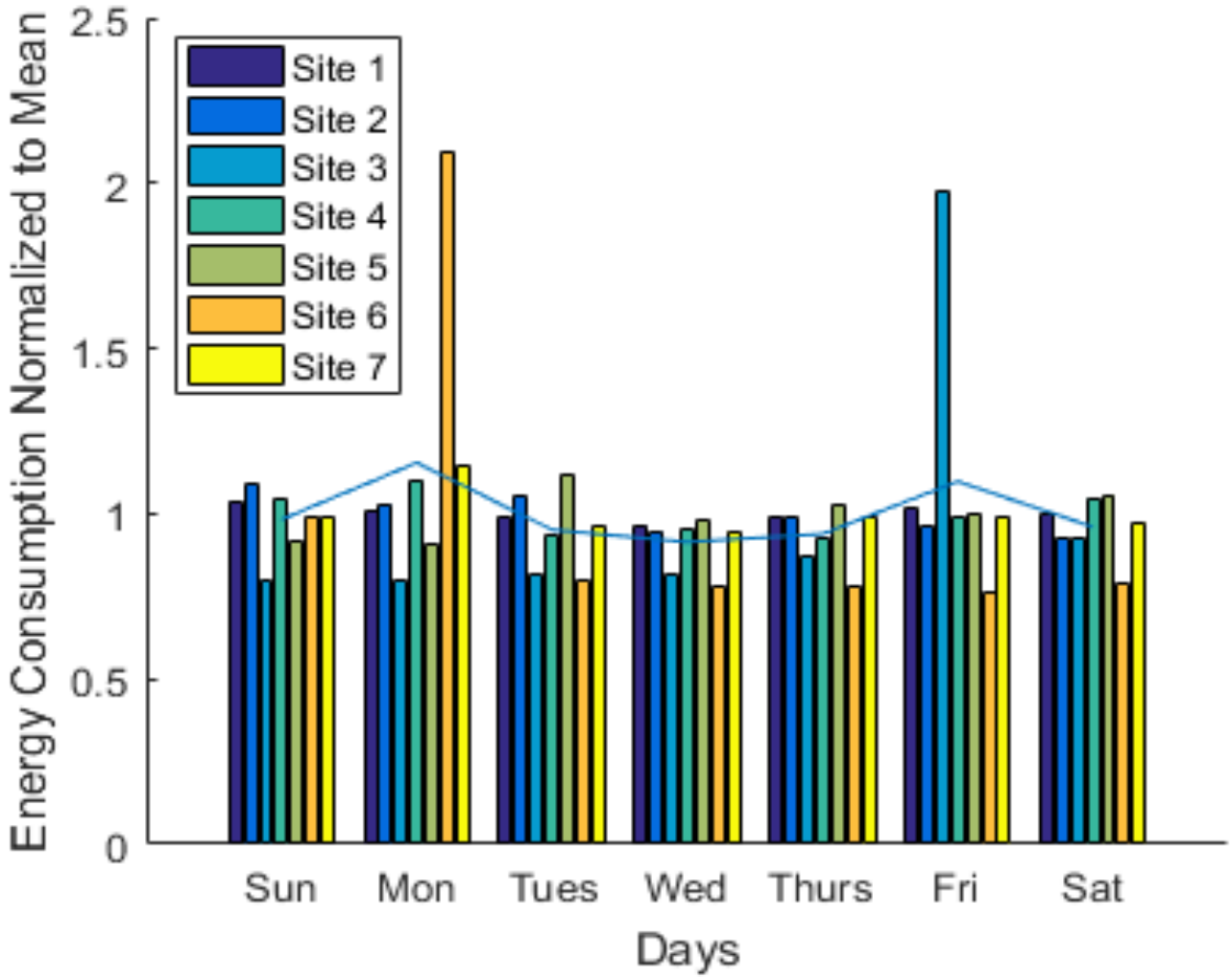
Energy Consumption



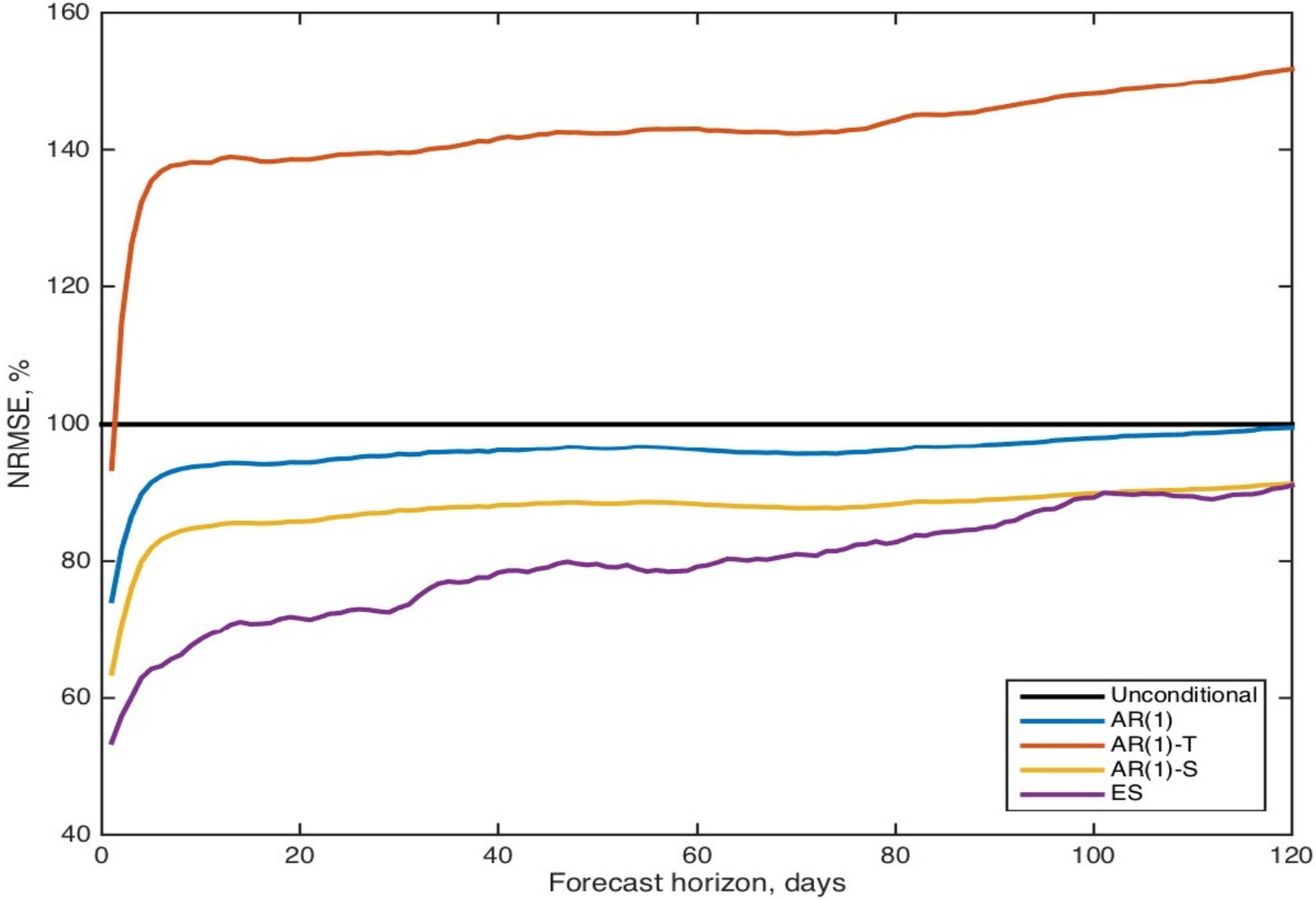
Intra-annual seasonality



Day of the week effect



Forecasting Demand



Data Analytics

WEEK 2B

Course outline

Week	Lecture A	Lecture B
1	Data Analytics	Weather forecasting
2	Renewable energy	Wind energy
3	Solar energy	Demand forecasting
4	Risk	Extreme events
5	Health	Biomedicine
6	Early warning systems	Economic forecasting

Today's Lecture

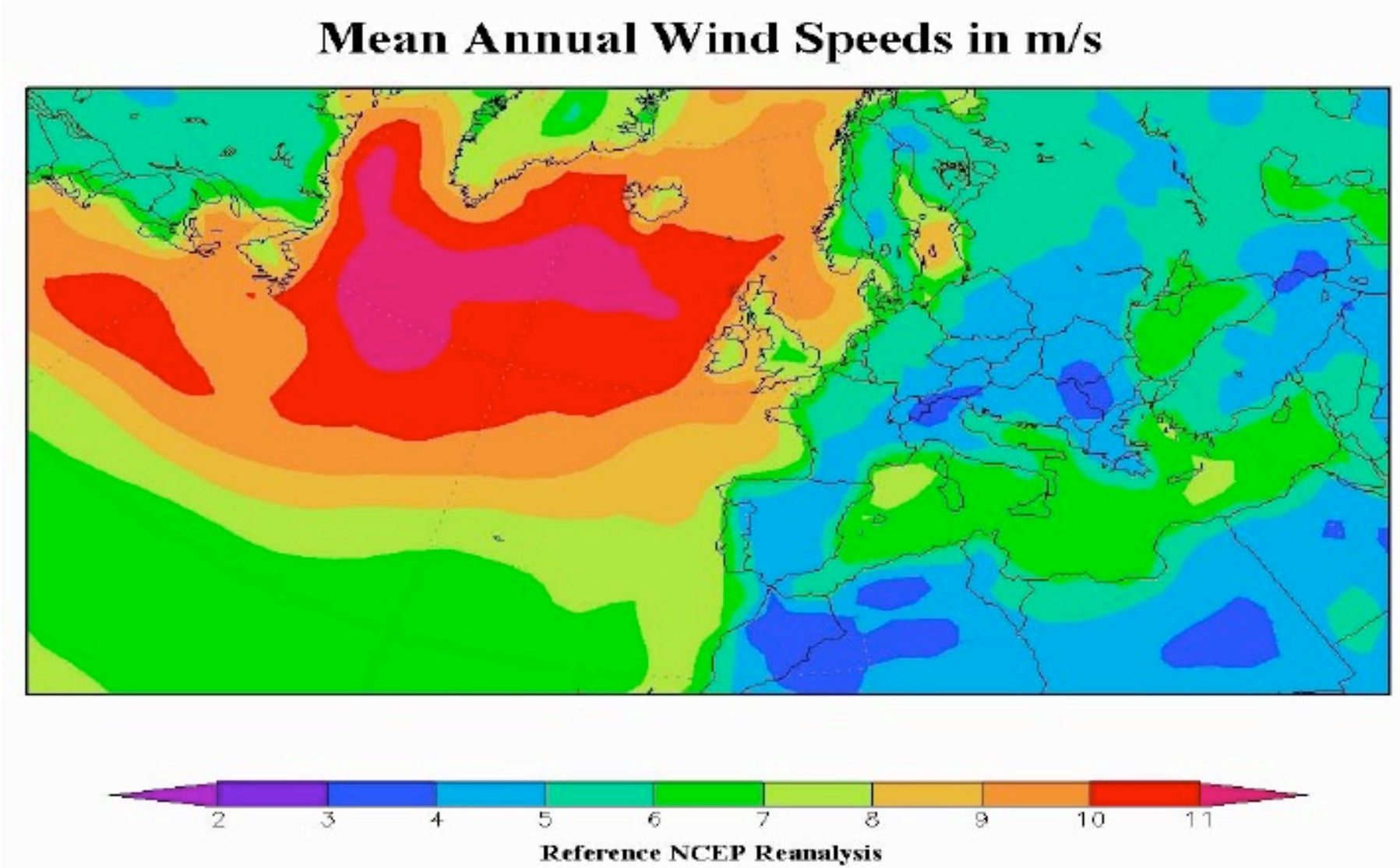
No.	Activity	Description	Time
1	Challenge	Wind energy integration	10
2	Discussion	Wind energy growth	10
3	Case study	Wind farm forecasting	10
4	Analysis	Probabilistic forecasting	20
5	Demo	Variability indices	20
6	Q&A	Questions and feedback	10

Origin of Wind Energy

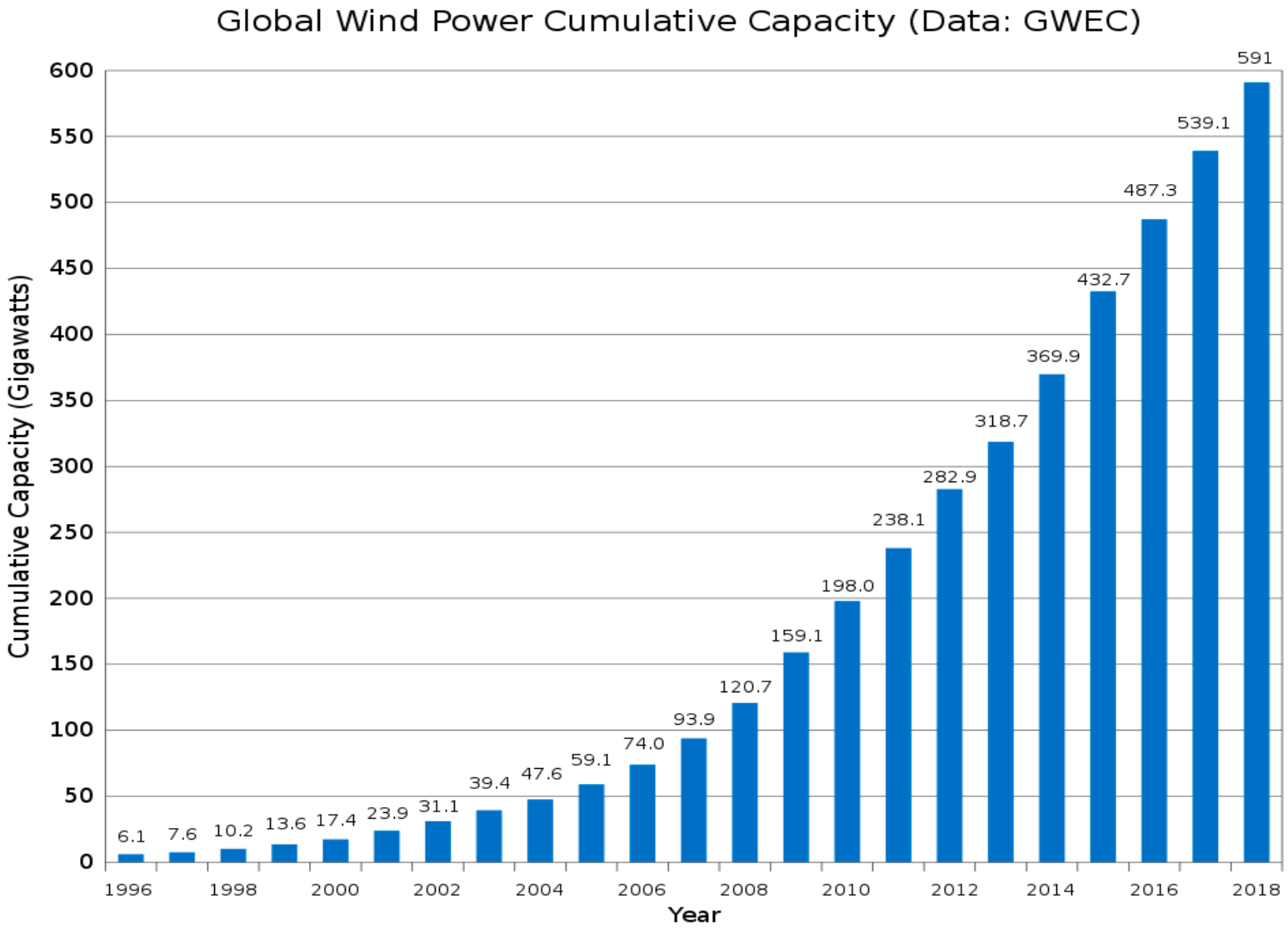
- 1% to 3% of energy from the Sun that hits the earth is converted into wind energy.
- The Earth is unevenly heated by the sun; poles receives less energy from the sun than the equator.
- Dry land heats up (and cools down) more quickly than the seas.
- Differential heating powers a global atmospheric convection system
- Power is directly proportional to the density of the air, the area swept out by the rotor, and the cube of the wind speed:

$$P=0.5\rho\pi R^2v^3$$

Wind Energy Potential



Worldwide Installed Wind Capacity



Source: Global Wind Energy Council, 2020

Poll

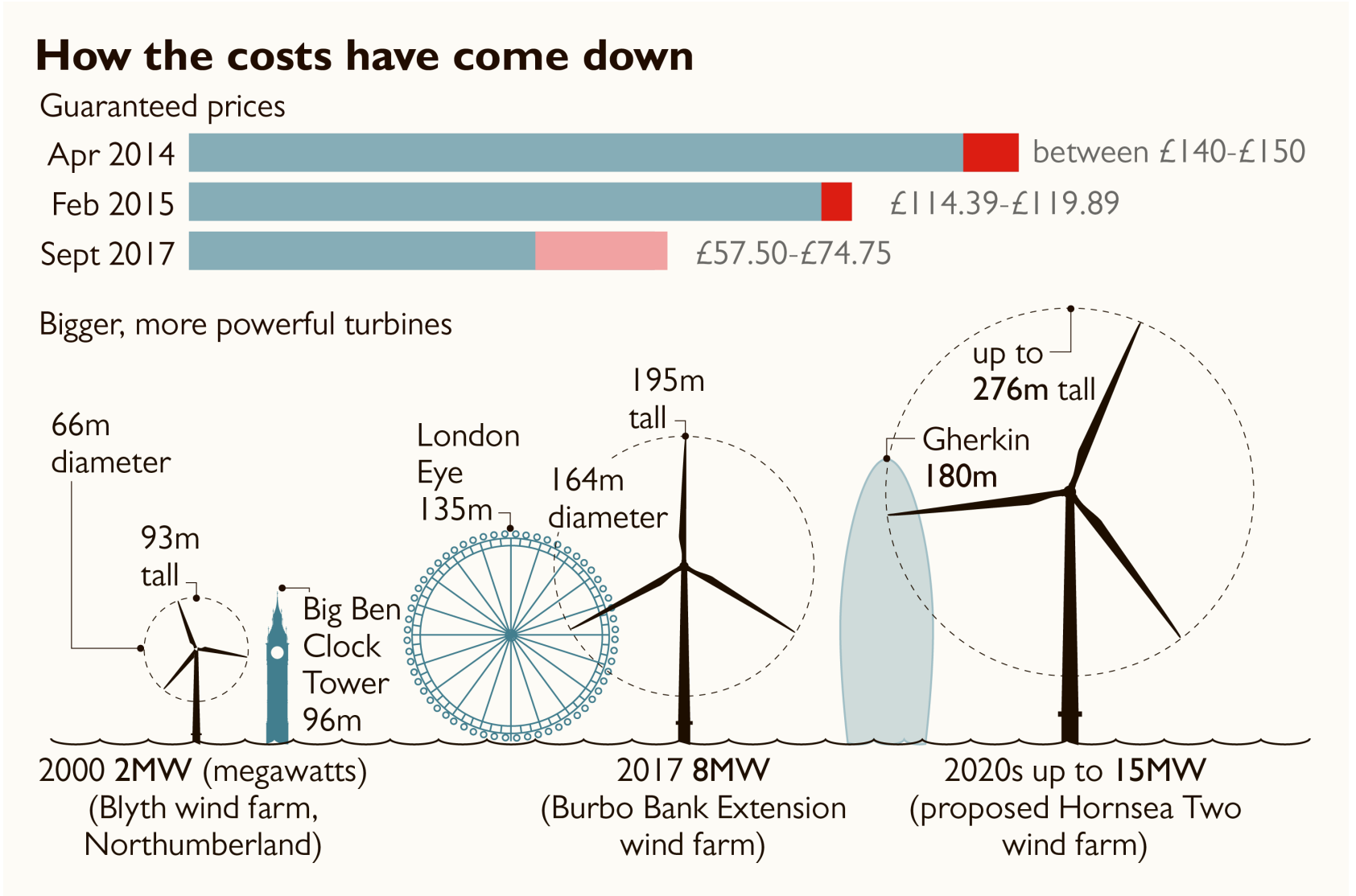
- Wind turbines are now bigger than:
  1) Boeing 747-8 (76m long)
  2) London’s Big Ben (93m tall)
  3) The London Eye (135m diameter)
  4) London’s Girkin Building (180m tall)

Slido.com #73896

Turbines


Bigger. Cheaper. Greener.
Boeing 747-8
Length: 76m
Vindeby
Year: 1991
Diameter: 35m
Tower Height: 35m
Capacity: 0,45MW
Middelgrund
Year: 2000
Diameter: 76m
Tower Height: 64m
Capacity: 2,00MW
Nysted
Year: 2003
Diameter: 82,4m
Tower Height: 69m
Capacity: 2,30MW
Horns Rev 2
Year: 2009
Diameter: 93m
Tower Height: 68m
Capacity: 2,30MW
Anholt
Year: 2012
Diameter: 120m
Tower Height : 82m
Capacity: 3,60MW
Westermost Rough
Year: 2014
Diameter: 154m
Tower Height: 102m
Capacity: 6,00MW
Burbo Bank Extension
Year: 2016
Diameter: 164m
Tower Height: 113m
Capacity: 8,00MW
DONG energy

Cost


How the costs have come down
Guaranteed prices
Apr 2014
between £140-£150
Feb 2015
£114.39-£119.89
Sept 2017
£57.50-£74.75
Bigger, more powerful turbines
66m diameter
93m tall
Big Ben Clock Tower 96m
London Eye 135m
164m diameter
195m tall
up to 276m tall
Gherkin 180m
2000 2MW (megawatts) (Blyth wind farm, Northumberland)
2017 8MW (Burbo Bank Extension wind farm)
2020s up to 15MW (proposed Hornsea Two wind farm)

Quiz

- Which country has the highest per capita electricity generatation from wind?
  1) China
  2) Denmark
  3) Germany
  4) USA

Slido.com #73896

Electricity generation from wind per capita

Rank	Country	kWh per person
1	Denmark	2825
2	Sweden	2702
3	Ireland	2251
4	Germany	1605
5	Finland	1384

Source: https://ourworldindata.org/grapher/wind-electricity-per-capita

Wind Energy Forecasting

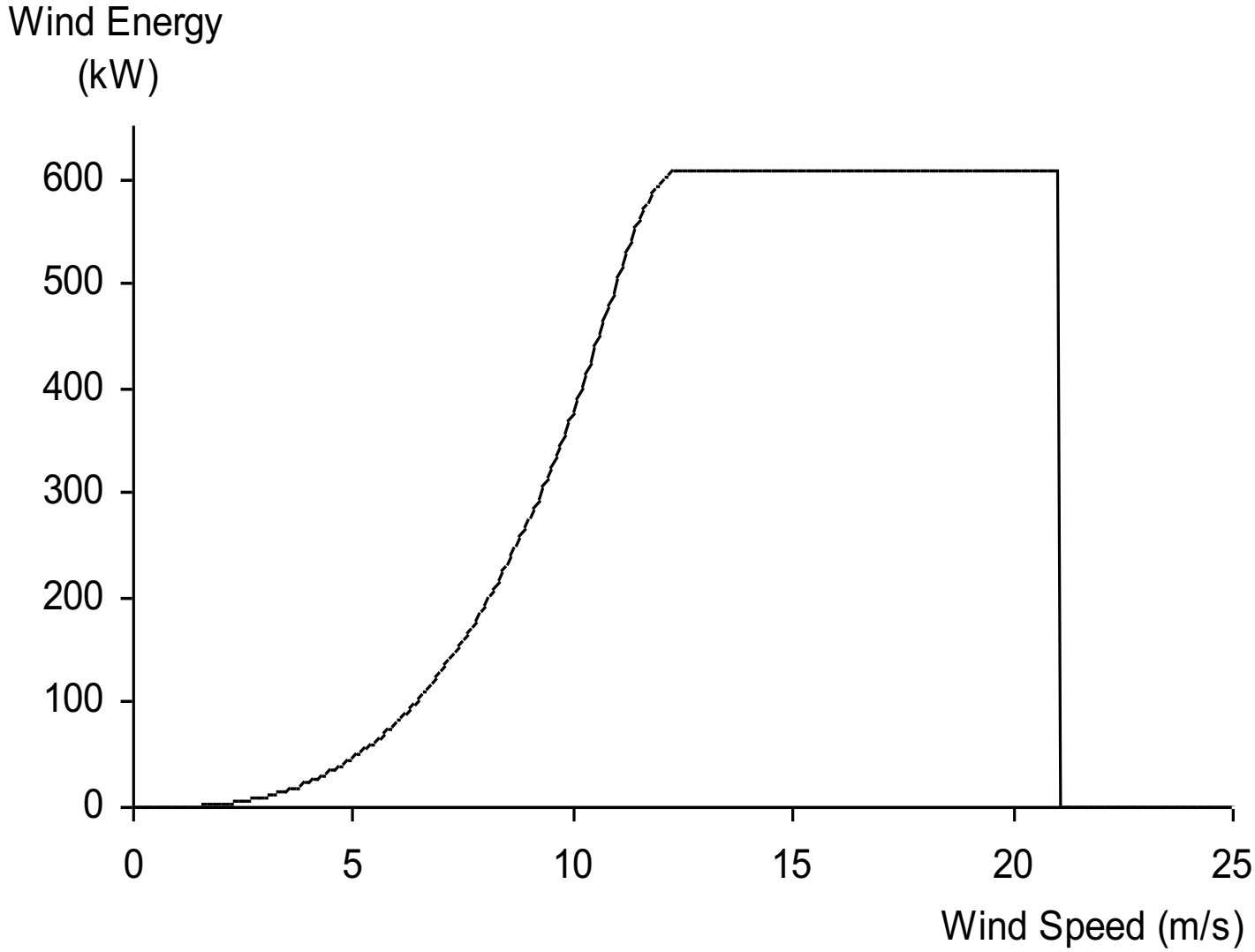
- Compared with other renewable energy sources, wind power is efficient, relatively cheap and has large resource.
- Prediction of generated wind power is tough because of erratic nature of earth's atmosphere. Uncertainty is problematic for electricity system operators and wind farm operators.
- We develop methods to predict midday wind power density for lead times from 1 to 10 days ahead for 5 UK wind farm locations.
- We consider AR-GARCH, ARFI-GARCH time series models and ensemble predictions produced by an atmospheric model.

Our Wind Farm Locations

Online	Wind farm	Turbine model	Turbines	MW Capacity	Annual homes equivalent	Operator
Dec-1992	Blood Hill, Norfolk	Vestas V27-225	10	2.25	1258	E.on Renewables
Oct-1997	Llyn Alaw, Anglesey	Bonus	34	20.4	11407	npower renewables
Sep-2001	Bears Down, Cornwall	Bonus	16	9.6	5368	npower renewables
Mar-2002	Bu Farm, Orkney	GE Wind 900S	3	2.7	1510	I & H Brown
Mar-2002	Cemmaes, Powys	Vestas	18	15.3	8555	Cumbria Wind Farms

Wind Power Curve

- For simplicity, we use the same power curve for all locations:

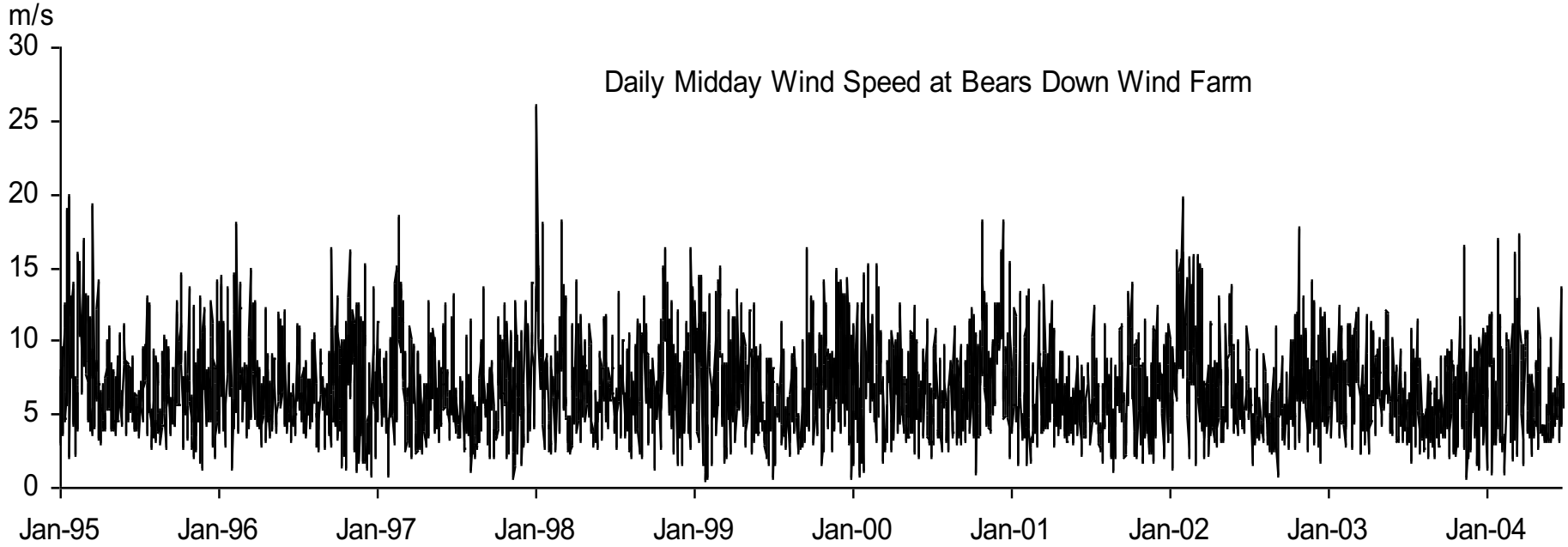


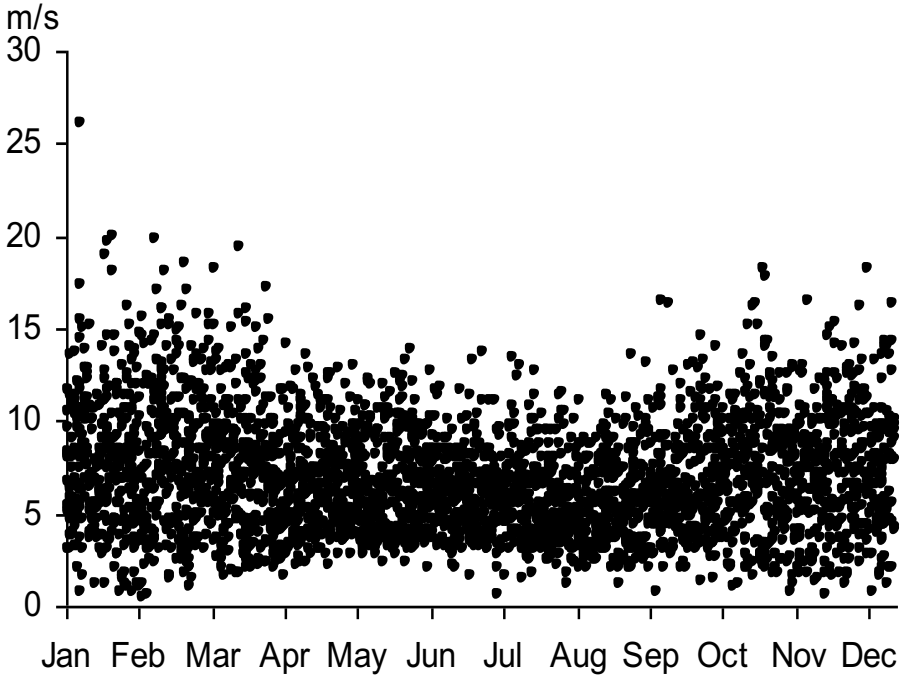
Poll

- When modelling daily wind speed data, we need to describe the following:
  1) Intraday variability
  2) Intraweek variability
  3) Intrayear variability

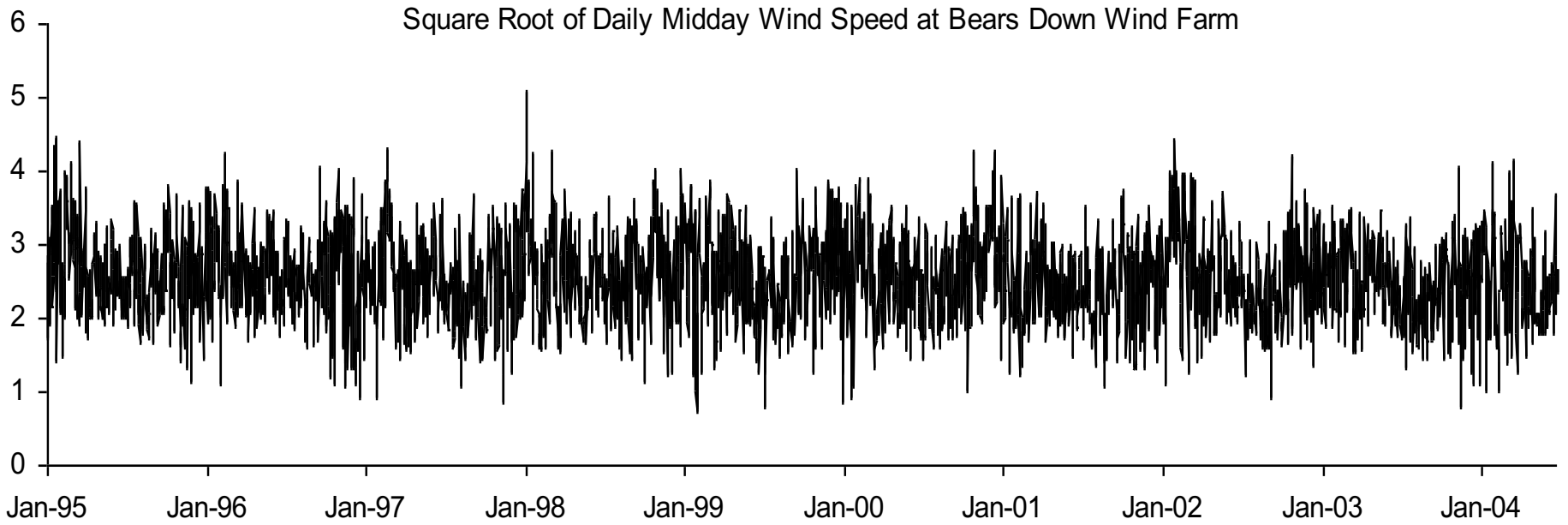
Slido.com #73896

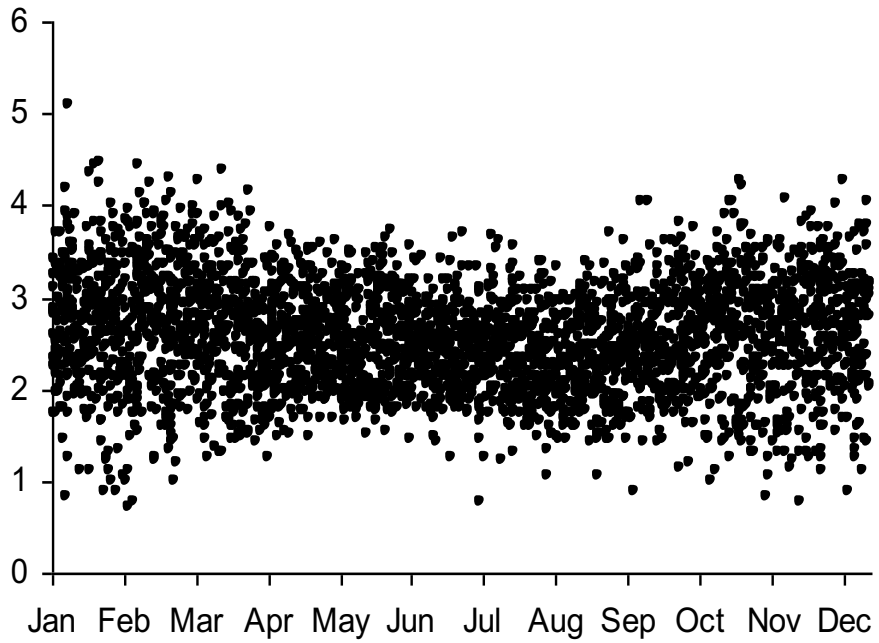
Wind Speed Time Series





Time Series of Square Root of Wind Speed





AR-GARCH Models

- We fitted AR-GARCH models to square root of wind speed:

$$y_t = S(\boldsymbol{\mu}, t) + \sum_{i=1}^{7} \phi_i \, y_{t-i} + \varepsilon_t$$

$$\varepsilon_t = \sigma_t \, \eta_t$$

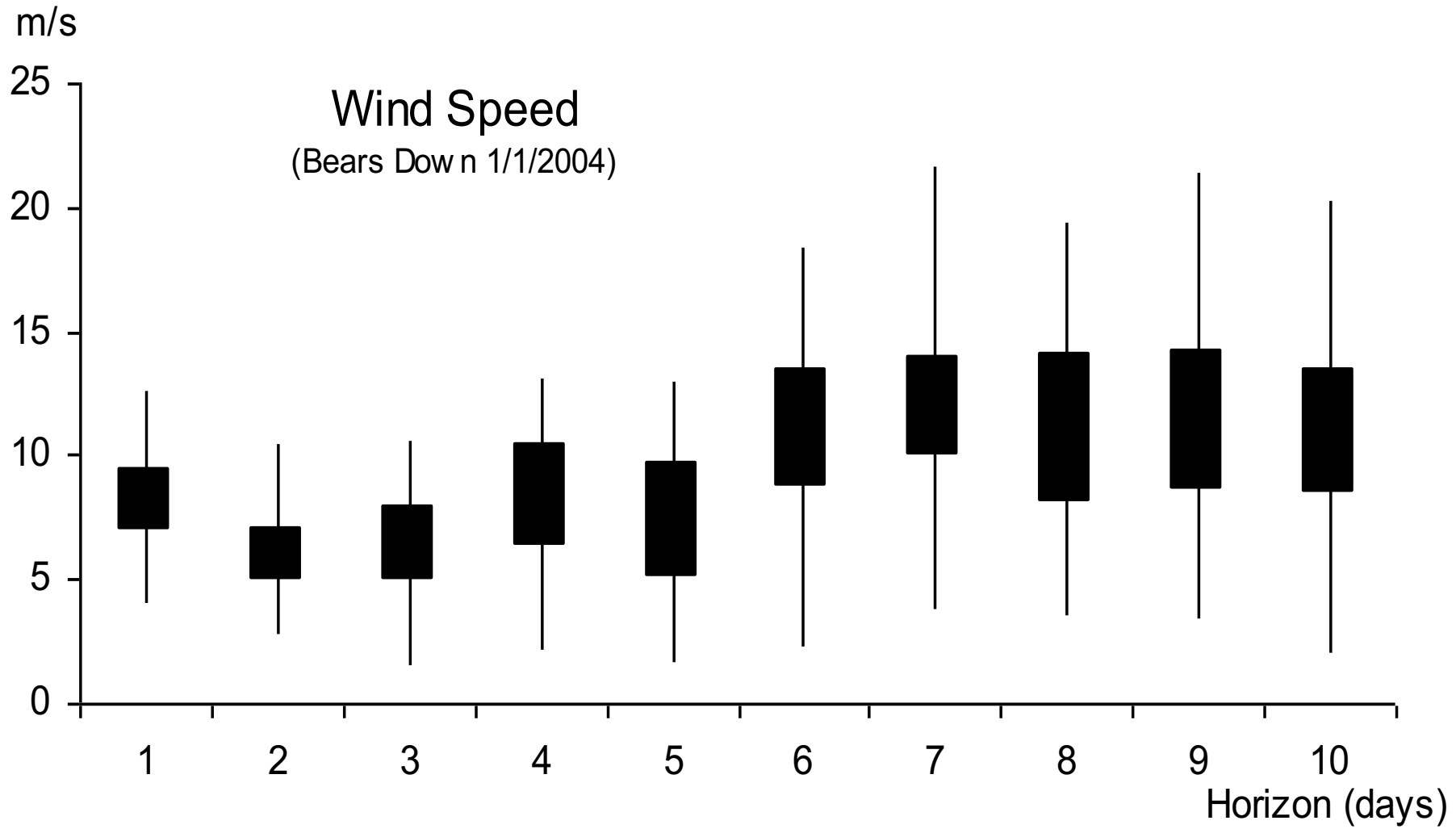
$$\sigma_t^2 = \boldsymbol{\omega} + \alpha \, \varepsilon_{t-1}^2 + \beta \, \sigma_{t-1}^2$$

$$S(\mu, t) = \mu_0 + \mu_1 \sin\left(2\pi \tfrac{d(t)}{365}\right) + \mu_2 \cos\left(2\pi \tfrac{d(t)}{365}\right) + \mu_3 \sin\left(4\pi \tfrac{d(t)}{365}\right) + \mu_4 \cos\left(4\pi \tfrac{d(t)}{365}\right)$$

- Wind speed density constructed using Gaussian or empirical distribution. Convert to wind power density using Monte Carlo simulation and power curve.

Wind Speed Ensemble Predictions

- An ensemble consists of 51 members (weather scenarios).



Maximum likelihood approach

- The ith member of the calibrated square root of wind speed ensemble forecast:

$$\hat{y}^{i}_{t+k|t} = \mu^{ENS}_{t+k|t} - b_k + \lambda_k \left(\tilde{y}^{i}_{t+k|t} - \mu^{ENS}_{t+k|t} \right)$$

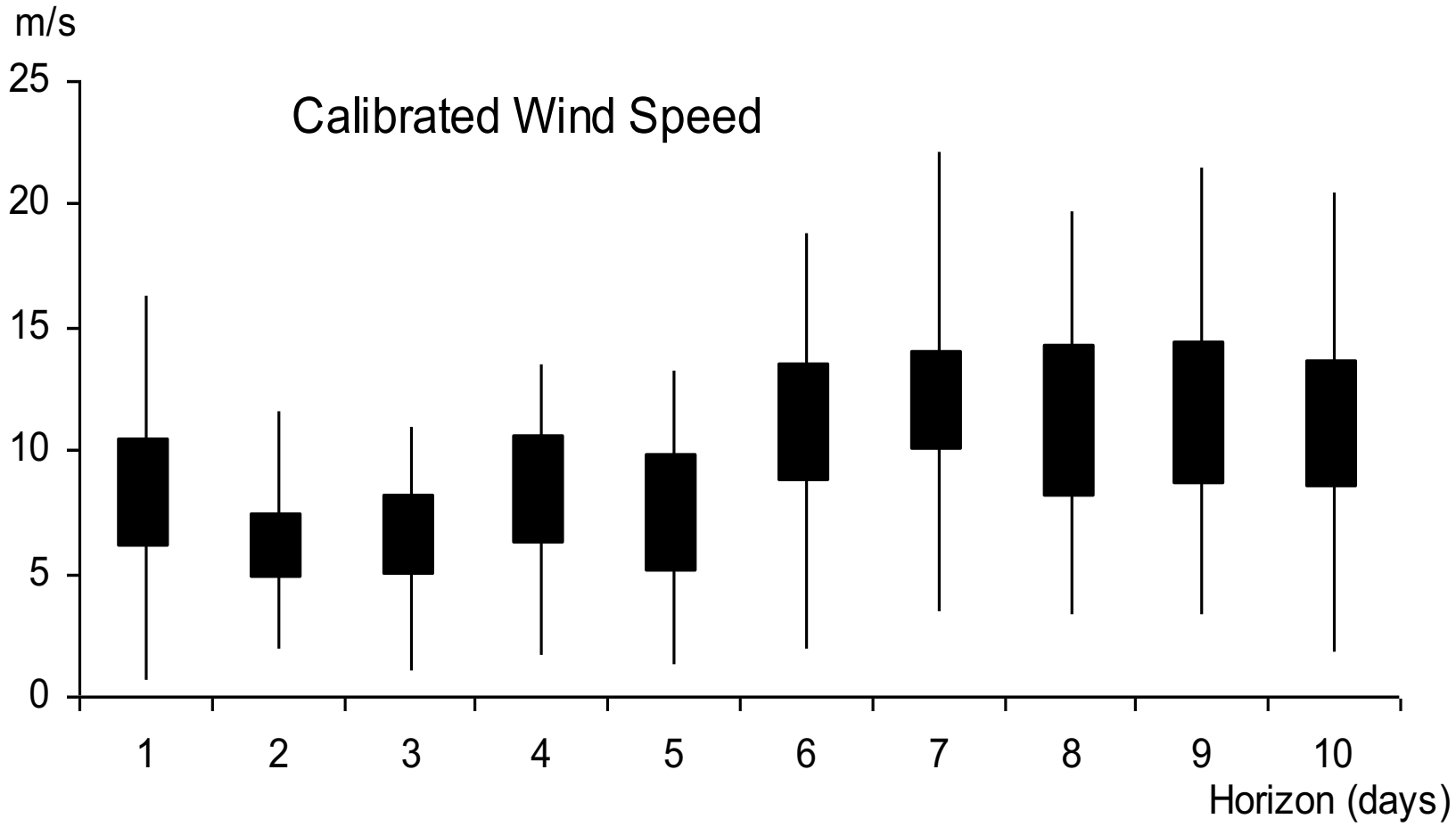
- $\tilde{y}^{i}_{t+k|t}$ is the square root of the original wind speed ensemble member; $\mu^{ENS}_{t+k|t}$ is the mean of the 51 ensemble members for the square root of wind speed; t is the forecast origin; k is the forecast horizon; and b_k and λ_k are calibration parameters
- Kernel density estimation with bandwidth h_k:

$$\hat{p}_{t+k|t}(y) = \frac{1}{51} \sum_{i=1}^{51} K(y, \hat{y}^{i}_{t+k|t}, h_k), \qquad \int_{0}^{\infty} K(y, \hat{y}^{i}_{t+k|t}, h_k) dy = 1,$$

- Historical forecasts/verifications; kernel smoothing and calibration jointly estimated:

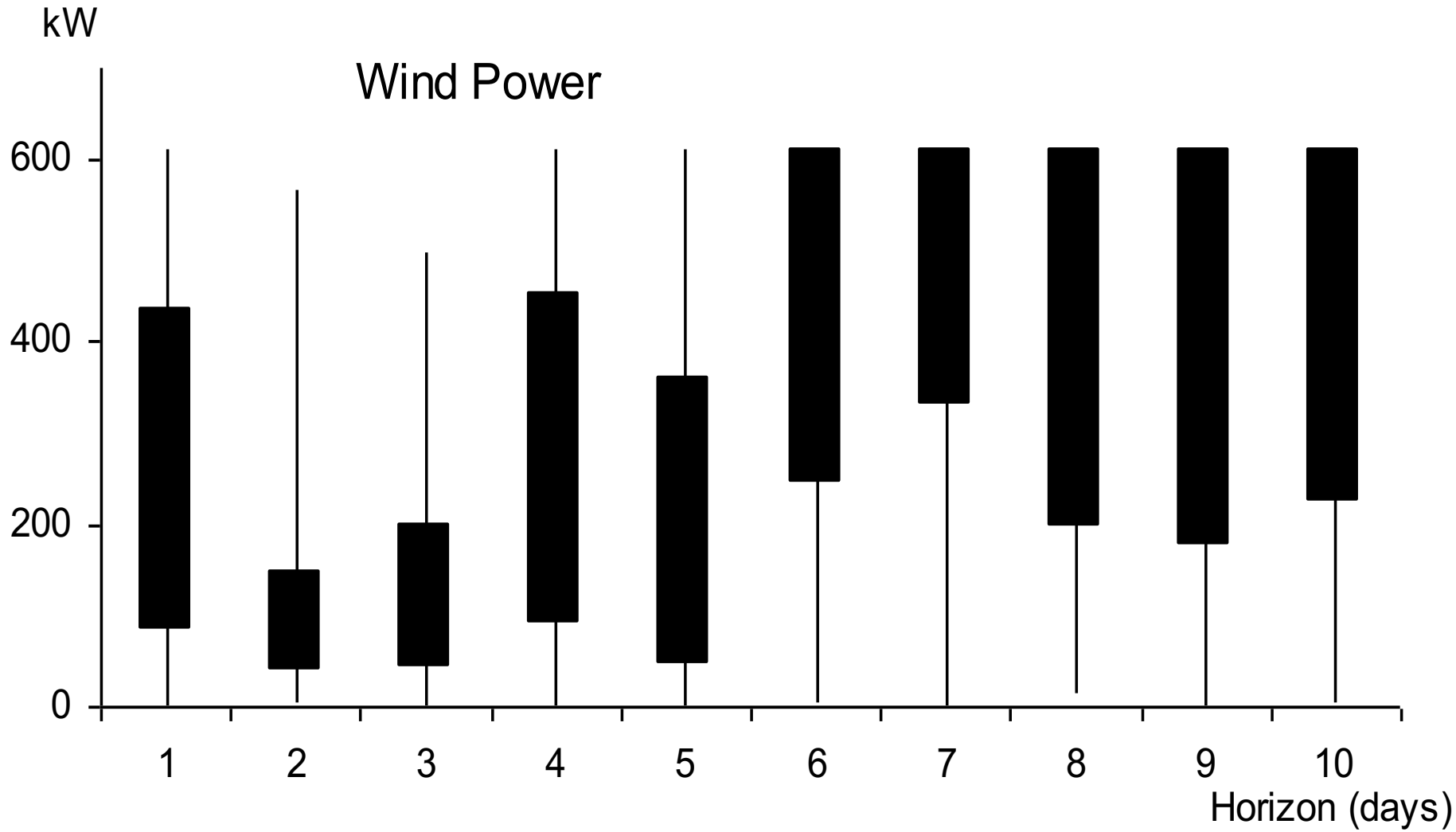
$$L(b_k, \lambda_k, h_k) = \sum_{t} \ln \hat{p}_{t+k|t}(y_{t+k}),$$

Calibrated Wind Speed Ensemble Predictions

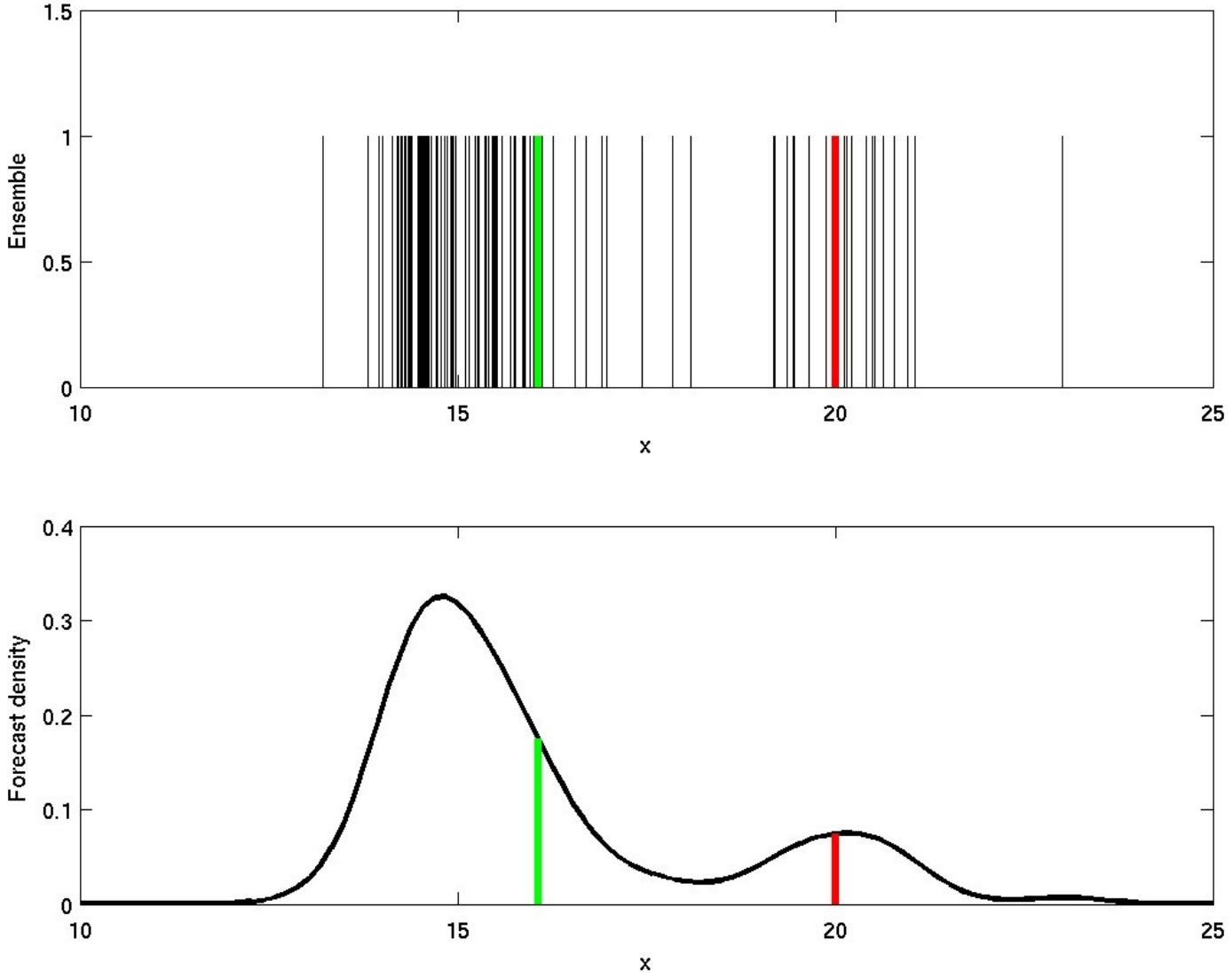


Wind Power Scenarios

- Converted 51 calibrated wind speed ensemble members into 51 wind power scenarios.

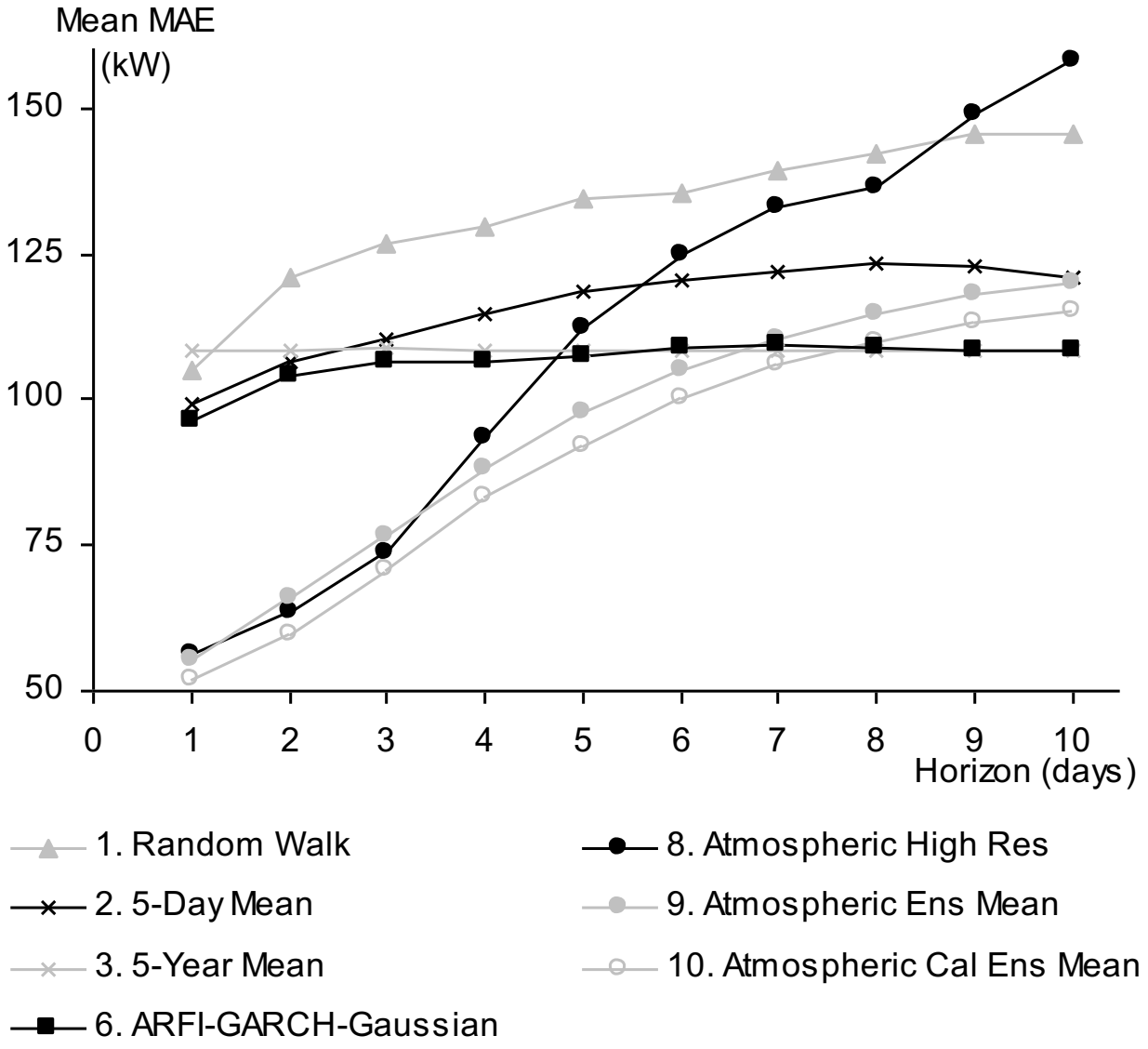


Log-likelihood evaluation

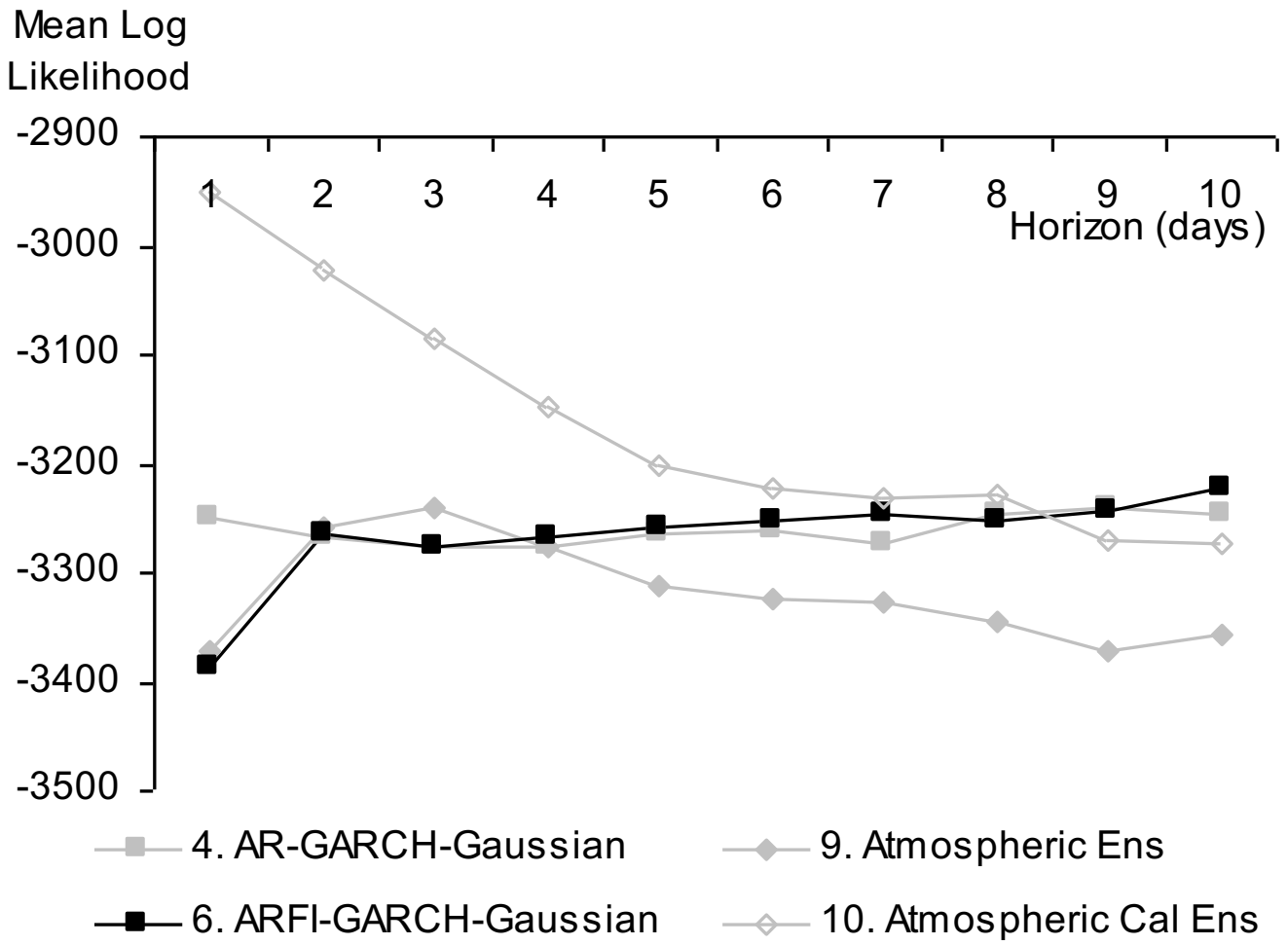


Evaluating Wind Power Point Forecasting

- Estimated ARFI-GARCH using 1995 to 2003 and calibrated ensembles using 1997 to 2003. Evaluated methods using 2003 to mid 2004.



Evaluating Density Forecasts - Log Likelihood



Summary of results

- Wind generation is fastest growing source of renewable energy. Risk management requires wind energy density forecasts.

- Ensemble predictions provide an alternative to AR-GARCH for wind energy density forecasting.

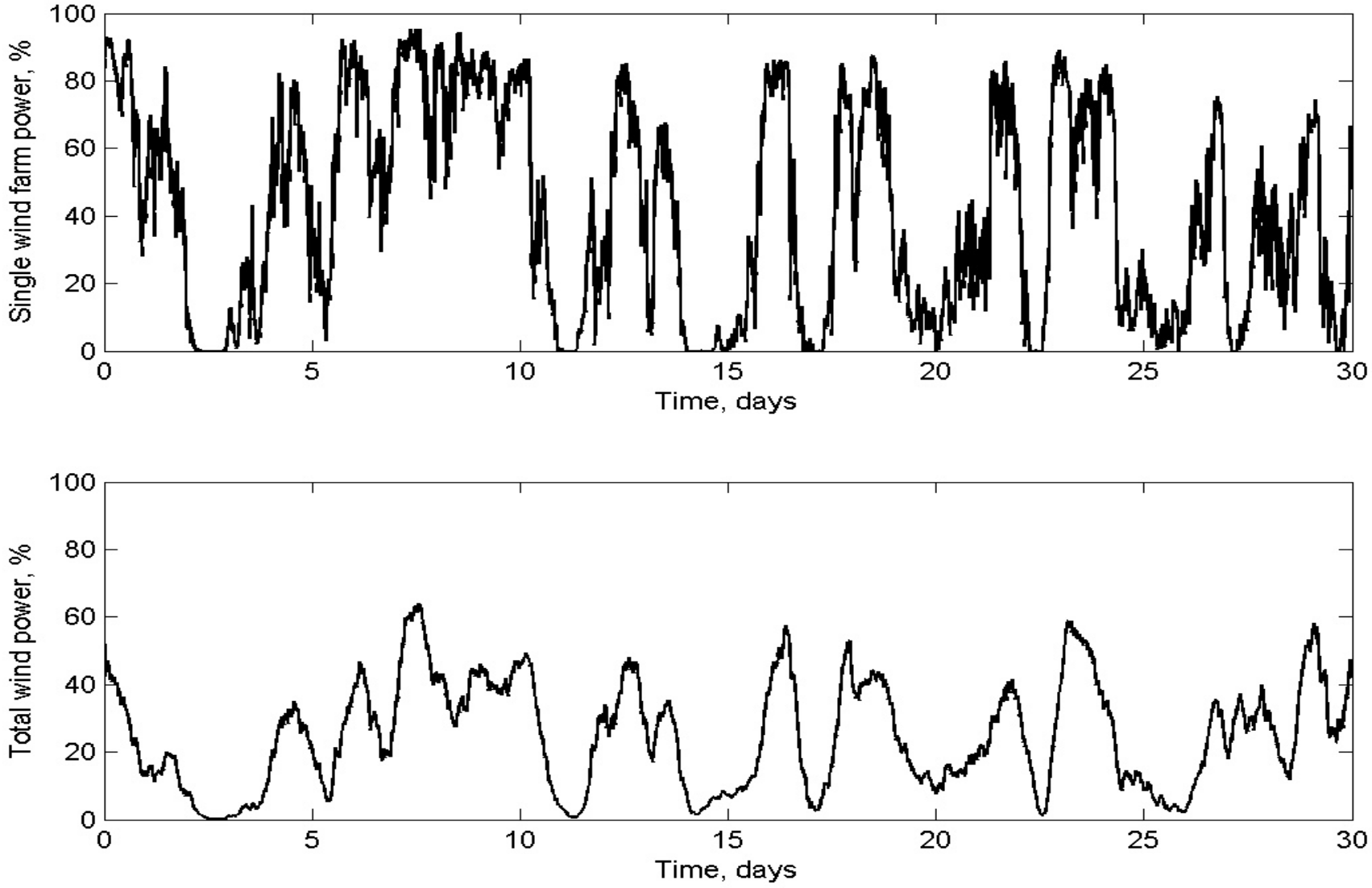
- Point forecasting:
  - Ensemble mean was best.

- Density forecasting:
  - Log likelihood

Wind Power Integration

- Integration of wind power is limited by variability
- Uncertainty is problematic for electricity system operators and wind farm operators.
- Wind power is often considered intermittent but in reality it varies due to meteorological conditions
- Variability depends on spatial locations of wind farms and diversification helps to smooth supply
- Accurate and reliable forecasts are required for horizons of 1-3 days ahead

Wind Power Generation



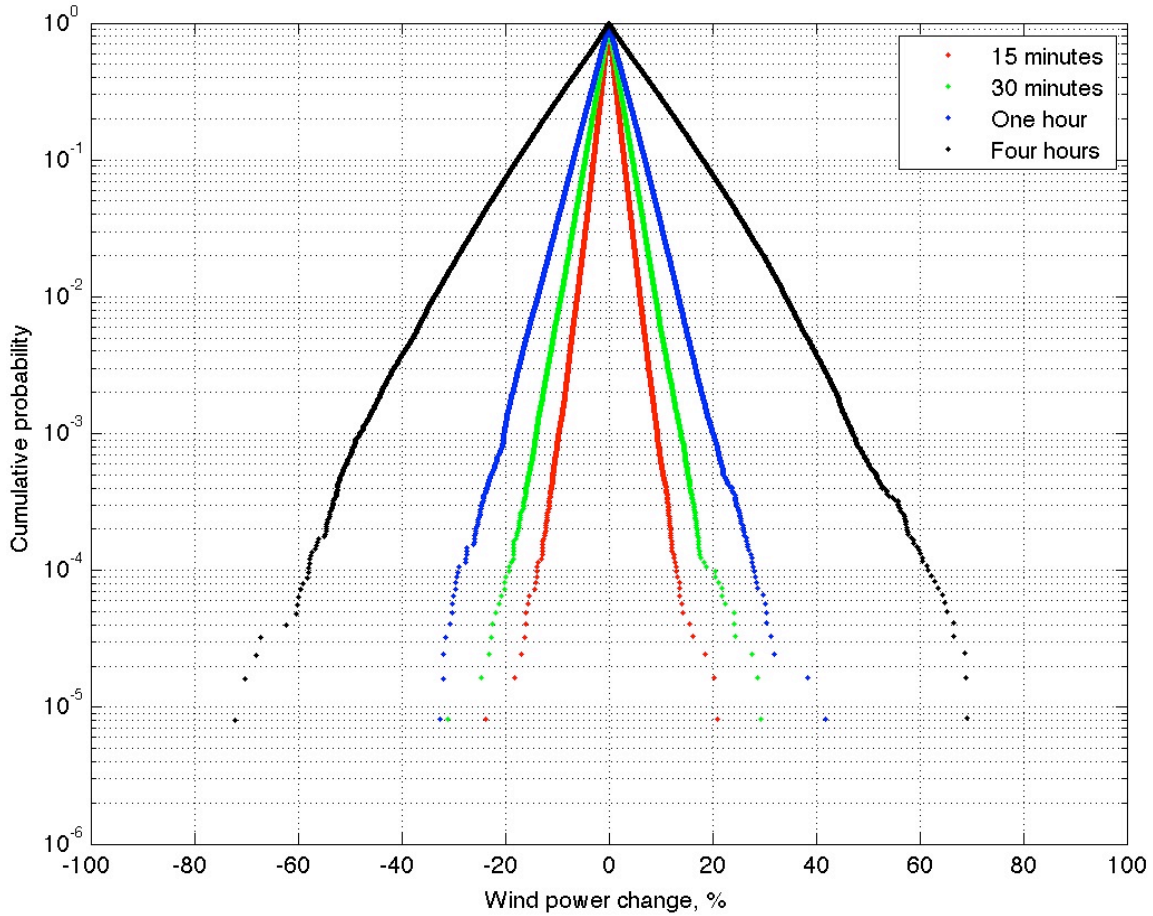
Source: EirGrid

Poll

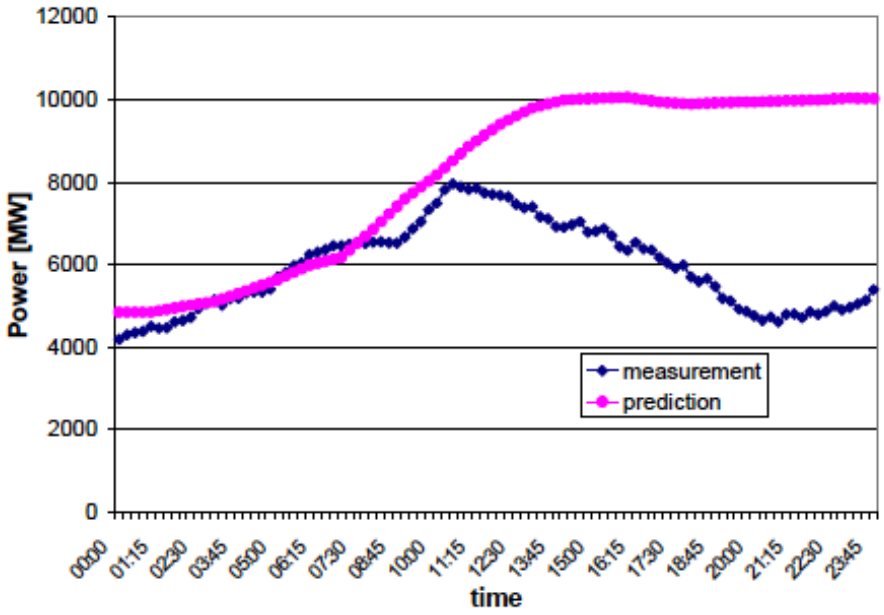
- Wind power changes (known as ramps) affect power system operation and are typically analysed by considering:
  1) Positive only
  2) Negative only
  3) Absolute changes
  4) Positive and negative separately

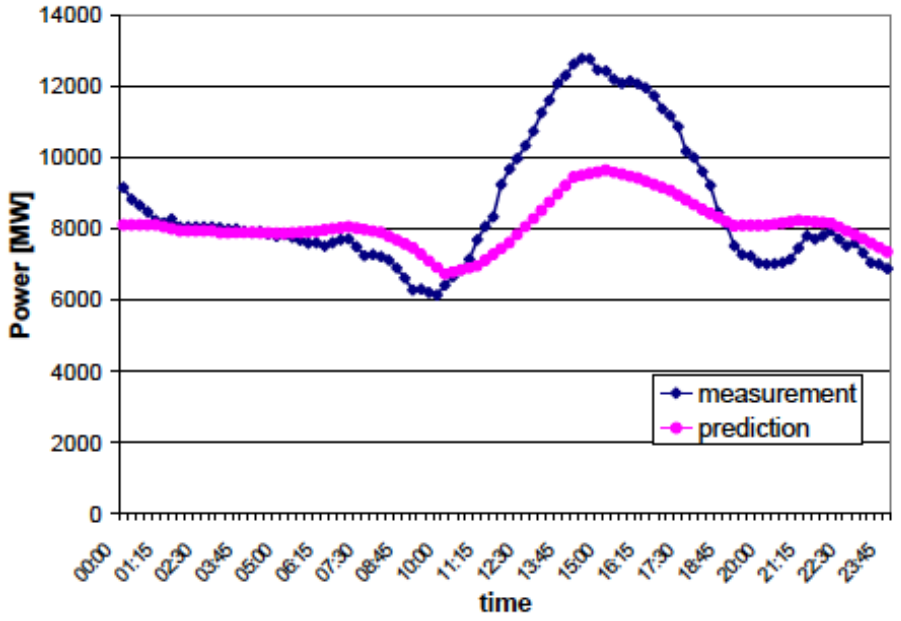
Slido.com #73896

Wind power fluctuations



Forecast Errors





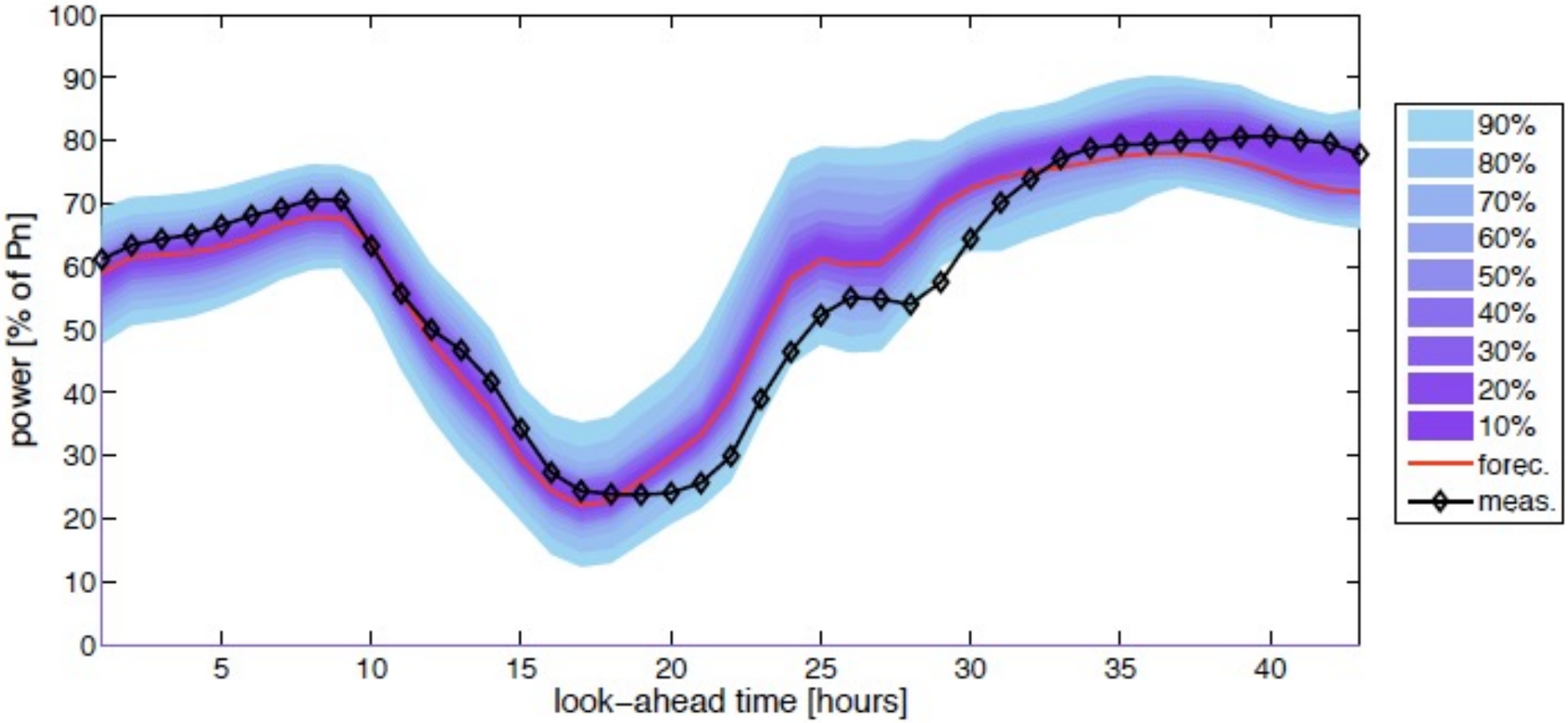
Extreme events in intraday wind power prediction for Germany.

(Left) Path of low-pressure system was different than predicted, maximum error: 5500 MW could have been avoided by extreme event correction.

(Right) Unexpected rise in pressure gradient in high-pressure situation. Maximum error: 2300 MW could have been avoided.

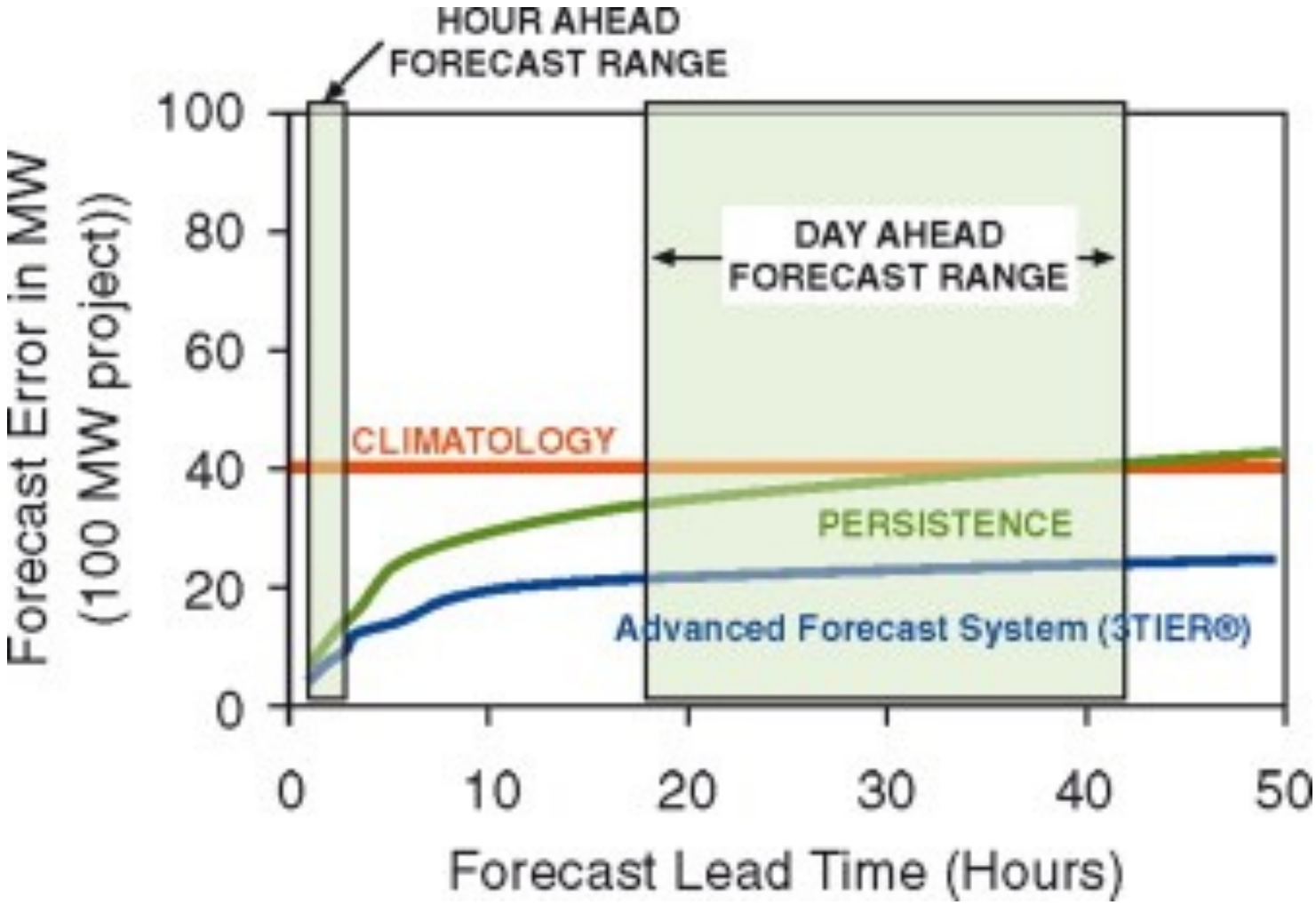
Source: Safewind

Probabilistic forecast of wind power in Denmark



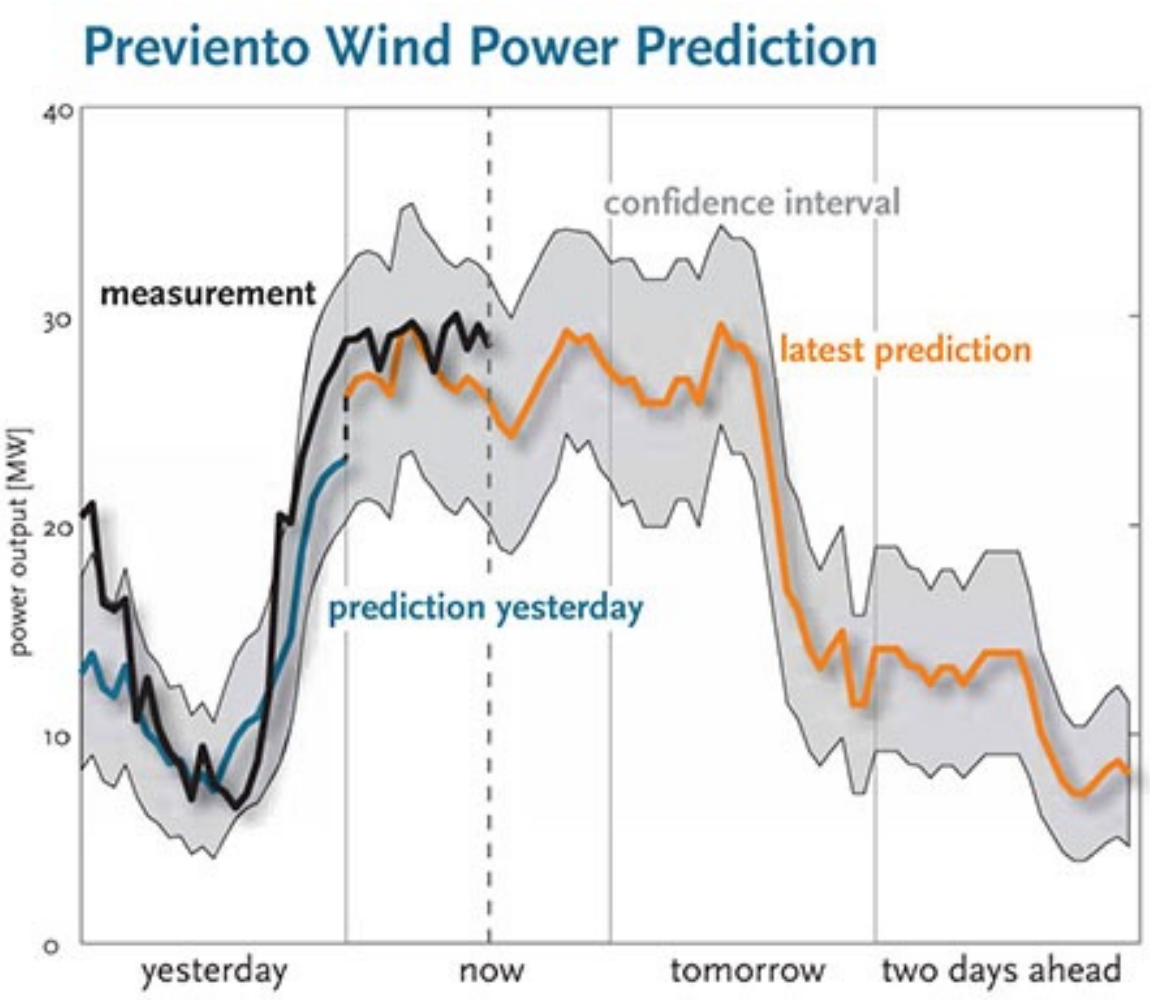
Source: SafeWind, colaboration with Pierre Pinson

Forecast horizon



Source: www.3tier.com

Previento forecasts



Source: www.energymeteo.com

Evaluation

- Comparison to forecasting benchmarks (persistence and unconditional) and linear ARIMA models
- Assessment of point and probabilistic forecasts (mean-absolute-error, quantile check function and continous ranked probability score)

Evaluating probabilistic forecasts

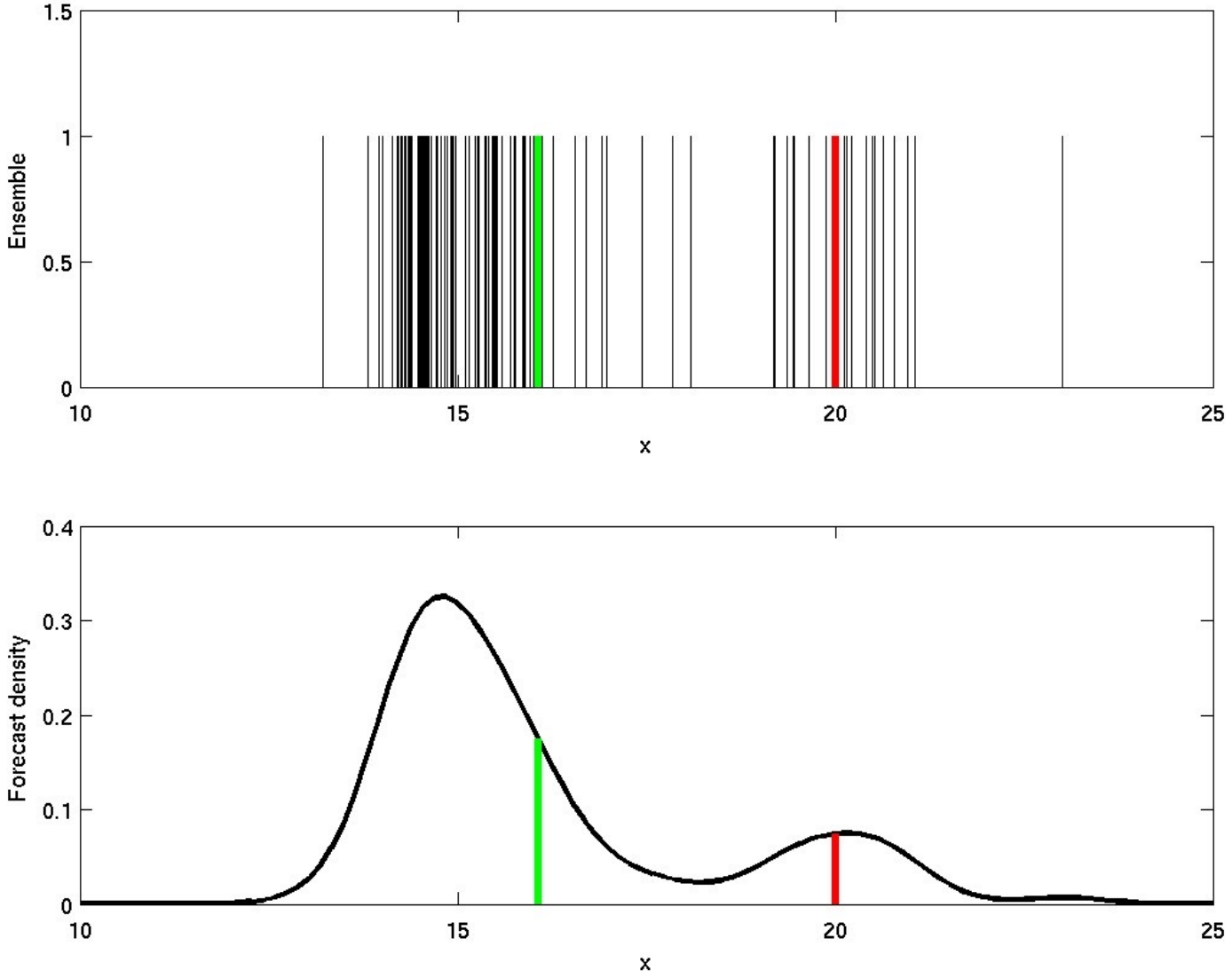
- Value: we can assess a forecast based on its value (economic) for improving decision-making
- Quality (statistical assessment)
  - Reliability: probabilistic correctness of the forecasts
  - Sharpness: ability to concentrate distribution
  - Resolution: accuracy conditional on explanatory variables
  - Skill: overall assessment of quality

Source: P. Pinson, P. McSharry, H. Madsen (2010). Quarterly Journal of the Royal Meteorological Society 136(646), pp. 77-90

Point and density forecasts

- We need density forecasts in order to account for uncertainty (observational, parametrical, model error)
- Decision-makers like certainty but point forecasts generally convey over-confidence
- Reliable and skillful density forecasts can provide improved point forecasts
- Point forecasts could be supplemented using prediction intervals or risk indices

Log-likelihood evaluation



Brier score

- Brier (1950) introduced the Brier score to measure the accuracy of a set of probability assessments
- It quantifies the average deviation between predicted probabilities for a set of events and their outcomes, so a lower score represents higher accuracy
- Suppose a forecaster gives a probability p of a particular event occurring
- Let $x = 1$ if the event occurs and $x = 0$ otherwise
- The Brier score for this single forecast/verification is:

  $BS = (p - x)^2$

- A Brier score of zero indicates a perfect forecast and one indicates poor performance

Brier score

- Suppose that an event is defined by the observable, y_{t+k}, falling below a specific threshold value y
- The Brier score for the CDF forecast, $\hat{F}_{t+k|t}(y)$, with a forecast horizon of k, is given by averaging across N forecast instances:

$$BS(k,y) = \frac{1}{N}\sum_{t=1}^{N}\left(\hat{F}_{t+k|t}(y) - I(y_{t+k} \leq y)\right)^2$$

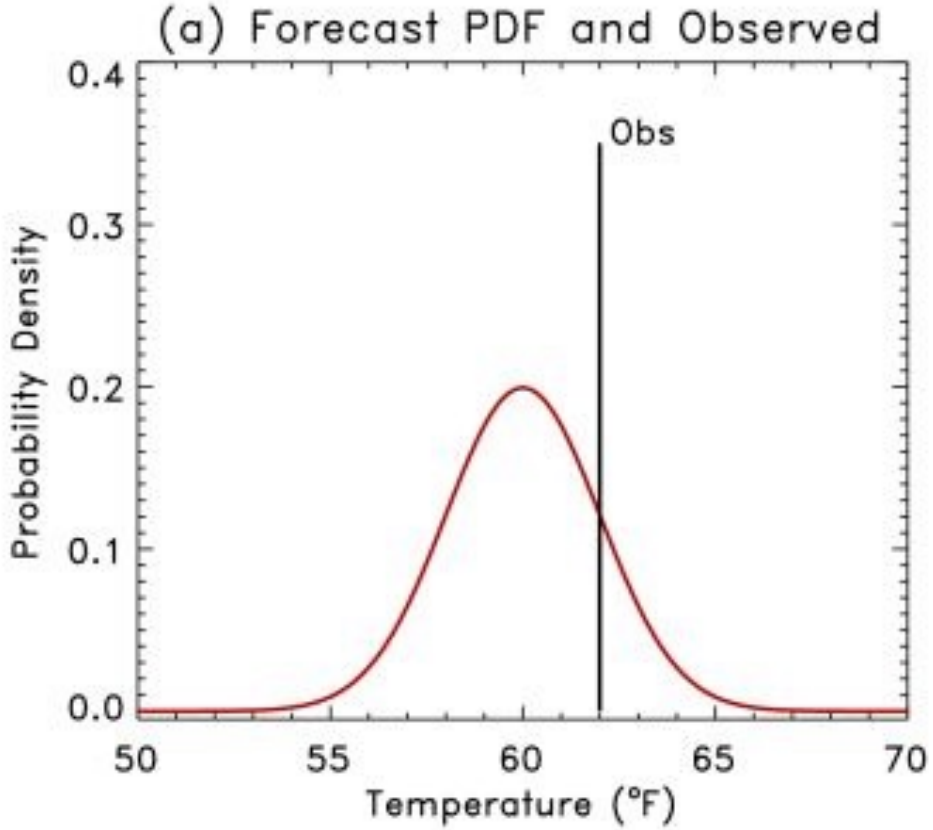
where I(.) is an indicator function

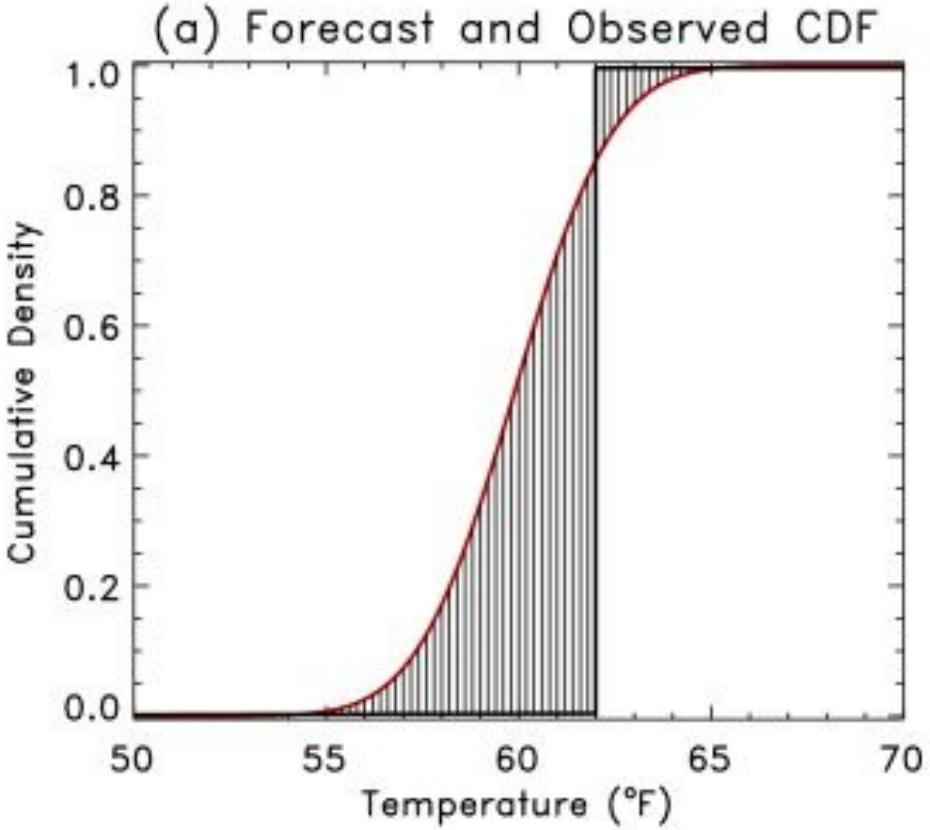
Continuous ranked probability score

- The crps for a CDF forecast, $\hat{F}_{t+k|t}(y)$, and corresponding verification, y_{t+k}, is defined by taking the integral of the Brier scores for the associated binary probability forecasts at all real-valued thresholds:

$$crps(\hat{F}_{t+k|t}, y_{t+k}) = \int_{-\infty}^{\infty} \left(\hat{F}_{t+k|t}(y) - I(y \geq y_{t+k}) \right)^2 dy$$

Graphic of CRPS





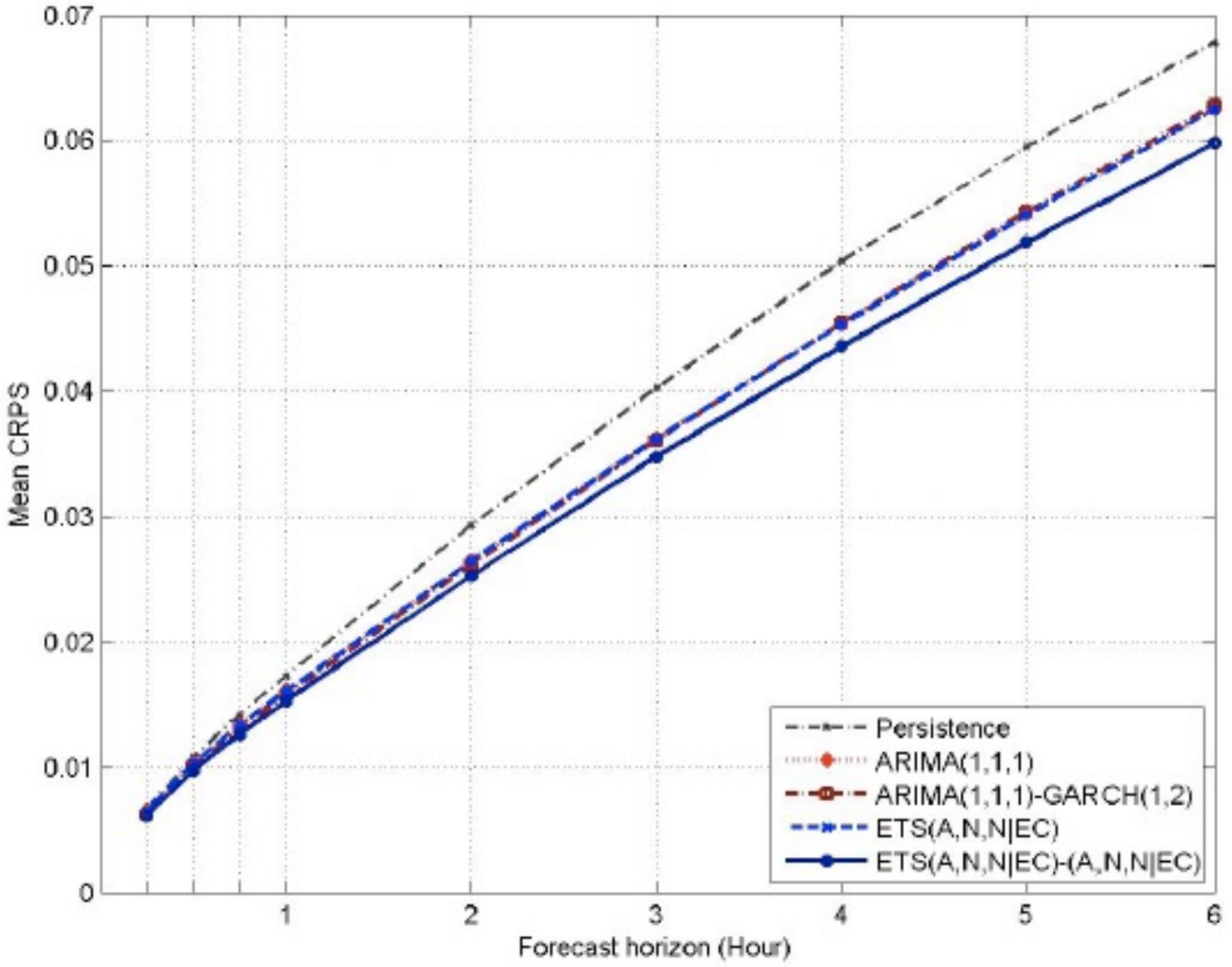
CRPS

- The average of these crps values over each forecast instance provides a score for each forecast horizon k

$$CRPS(k) = \frac{1}{N}\sum_{t=1}^{N} crps(\hat{F}_{t+k|t}, y_{t+k})$$

- CRPS provides what is known as a proper score, in that the forecaster minimizes the expected score for an observation drawn from the probabilistic forecast, F, by issuing F rather than any other competing probabilistic forecast
- Another useful property of the CRPS score arises from the fact that for point forecasts CRPS reduces to the mean absolute error (MAE)

Short-term wind forecasts



Lau & McSharry (2010), Annals of Applied Statistics (online).