Energy Meteorological In-Situ Big Data Analytics

Morris Riedel[1,2], Jonas Berndt[1,3], Charlotte Hoppe[1,3], Hendrik Elbern[1,3]

[1]Institue of Energy and Climate Research (IEK-8), Forschungszentrum Jülich GmbH, Jülich, Germany
[2]University of Iceland, Reykjavik, Iceland
[3]Rhenish Insitute for Environmental Research at the University of Cologne, Cologne, Germany

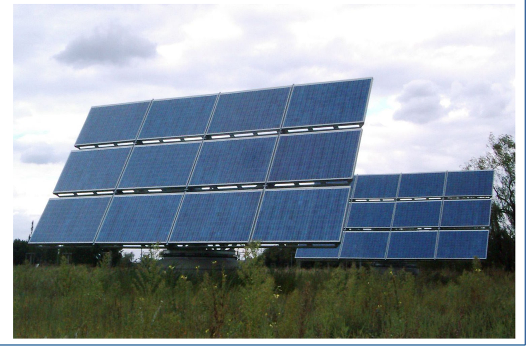
contact: m.riedel@fz-juelich.de
j.berndt@fz-juelich.de

Background & Objective

The stochastic nature of weather imposes wind and solar power as an uncertain source of electrical energy. Stable power grid management and energy trade on stock markets call for **improvement of probabilistic wind and solar power forecasts**. The major potential lies in the improvement of the underlying weather forecast.

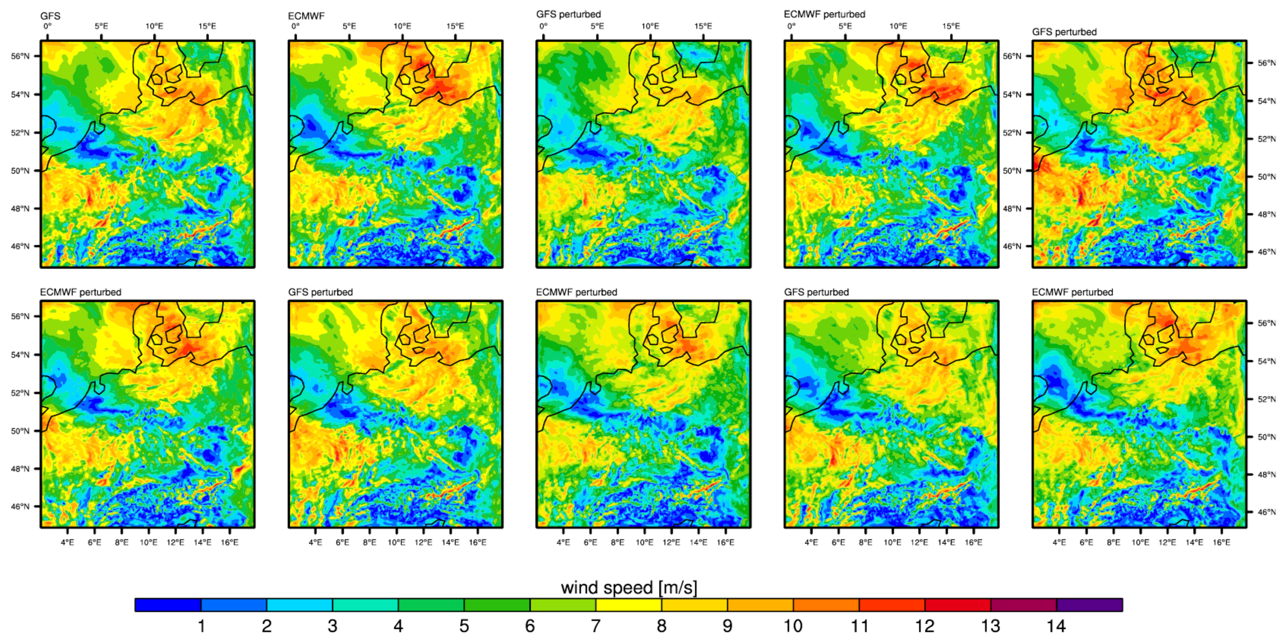
We make use of various perturbation techniques in the frame of a regional **meteorological ensemble with *O*(1000) members** to capture extreme error events and to improve skill scores of short and shortest range forecasts of wind speed at hub height (~ 100m) and irradiance.

A **data mining application** shall identify the relationship between **observation compliance and perturbation techniques**. This information serves as a basis to improve the further generation of ensemble members within a **particle filter algorithm**.



Meteorological Ensemble

- An ultra large ensemble version of the Weather Research and Forecast model (WRF) as part of ESIAS (Ensemble for Stochastic Integration of Atmospheric Simulation), which provides a comprehensive probability density evolution of the model state
- Computational efficient implementation on the JUQUEEN, which realizes communication between the ensemble members by introducing a second stage of MPI parallelism
- Initial values and lateral boundary values from the global ECMWF and GFS ensembles
- A broad variaty of state-of-the-art techniques of uncertainty representation within the model (SKEBS – Stochastic Kinetic-Energy Backckscatter Scheme, SPPT – Stochastic Perturbed Parameterization Tendency, perturbation of surface values, etc.)



Stamp plots of a 12-member selection with either GFS or ECMWF boundary and initial conditions and SKEBS model perturbation.

Particle Filter/Smoother Approach

An ensemble size of (*O*)1000 member offers the possibility to apply a Sequential Importance Resampling Filter (SIRF, or Sequential Monte Carlo method) as a **novel, non-linear data assimilation technique in atmospheric science**.

Parcticle filtering consists of representing the initial density of the model state by an ensemble of size *N*

$$p(\Psi) = \frac{1}{N}\sum_{i=1}^{N}\delta(\Psi - \Psi_i).$$

By applying Bayes' theorem, we estimate the posteriori density of the model state, given the observations *d*, is

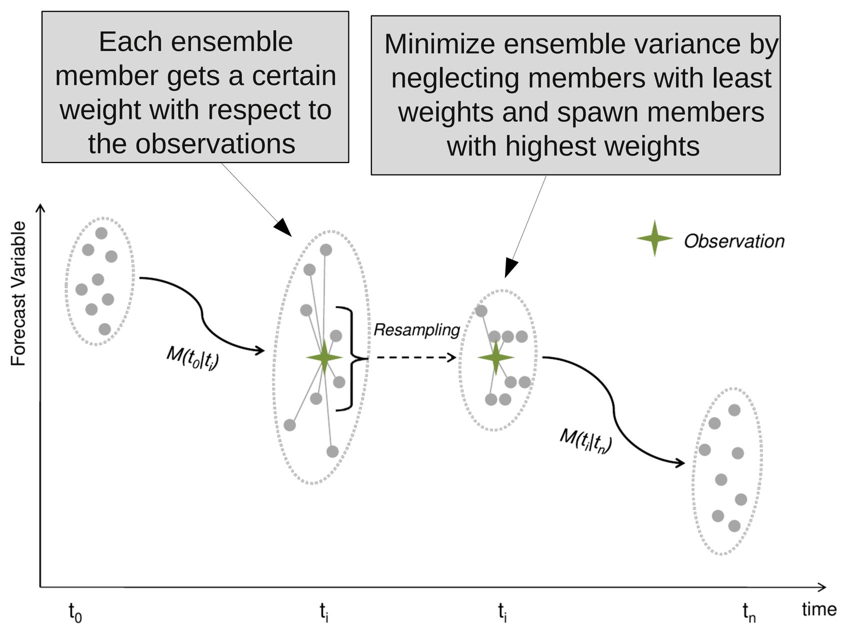
$$p(\Psi|d) \;=\; \sum_{i=1}^{N} w_i\delta(\Psi - \Psi_i)$$

Whereas the weights are given by

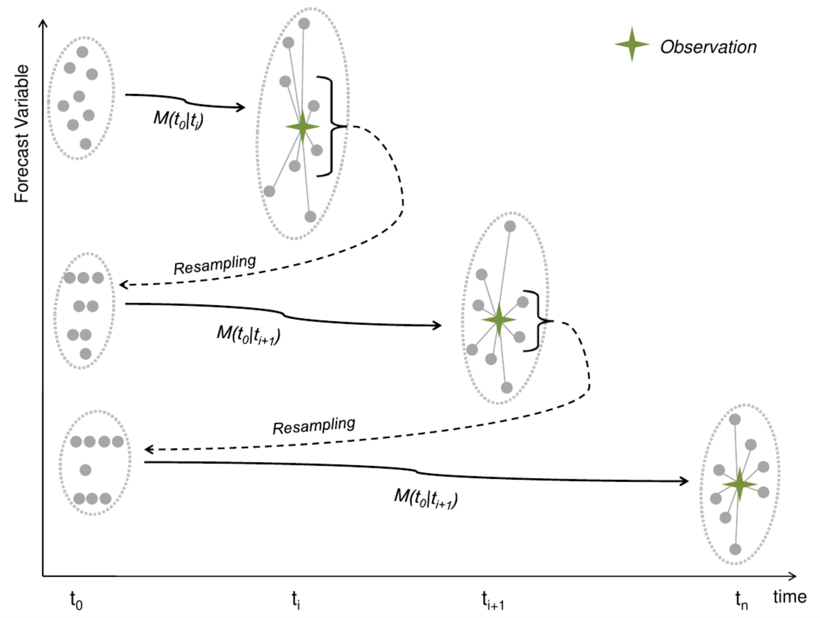
$$w_i = \frac{p(d|\Psi_i)}{\sum_{j=1}^{N} p(d|\Psi_j)}.$$

The obstacle of **filter degeneracy in high-dimensional systems** has be overcome. Our approaches include a smoothing algorithm, localization and/or sampling from a proposal density.

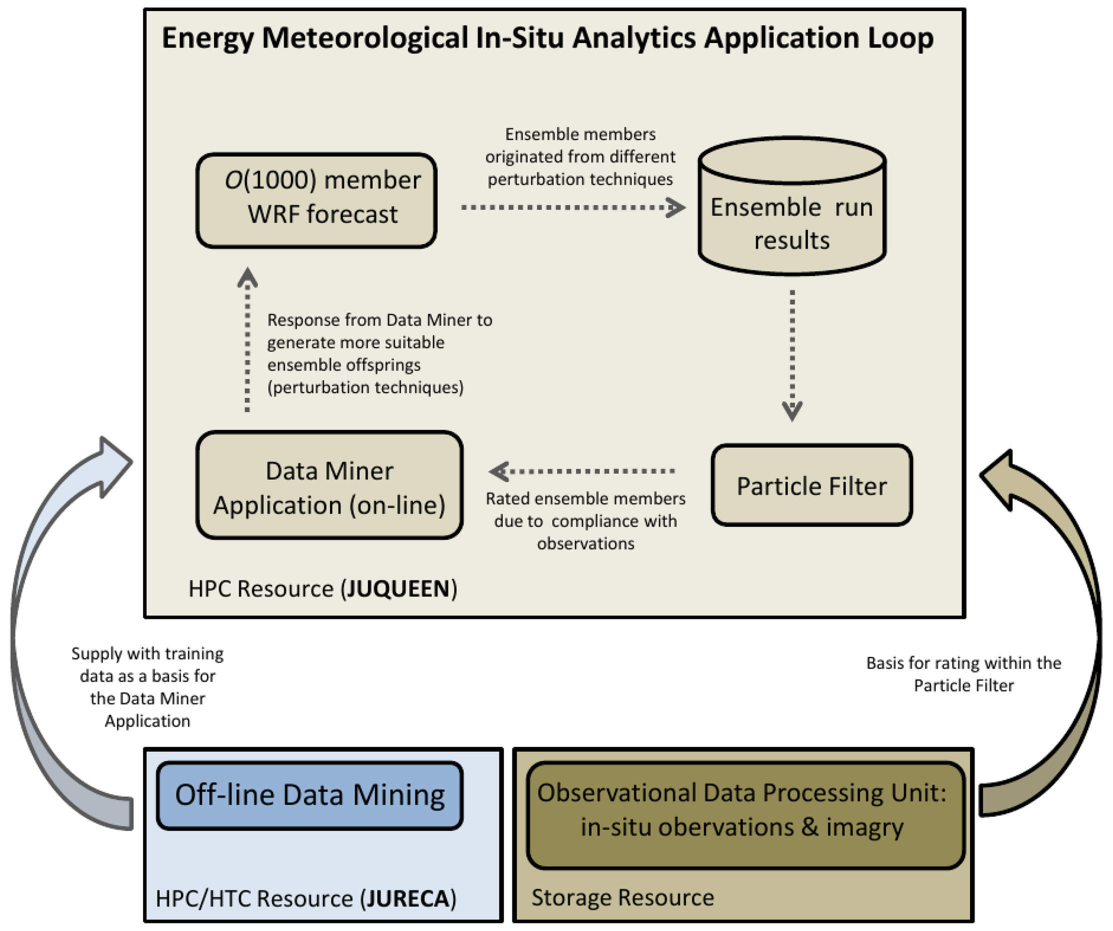
Particle Filter



Particle Smoother



Coupled Forecast-Analysis System



Computational Resources

The amount of computational resources is proportional to the ensemble member number. As an elemantary example, 1024 members require 65536 cores on the IBM BlueGene Q system (JUQUEEN) to accomplish a 24 hour simulation in approximately 1.5 hours (the NWP comprises the majority, with I/O tasks and on-line data mining being of secondary relevance).

Amount of data to be managed

.
.
.
.
.
.

In-Situ Analytics Application Loop

The coupled forecast-analysis system combines the meteorological forecast, particle filtering and data miner in one application loop. Due to high computational demands, special focus is set on computational efficiency. The particle filter as well as the piSVM1.2.1 shall be embedded in the WRF ensemble setup, with inter-member communication exclusively at particle filter resampling steps.

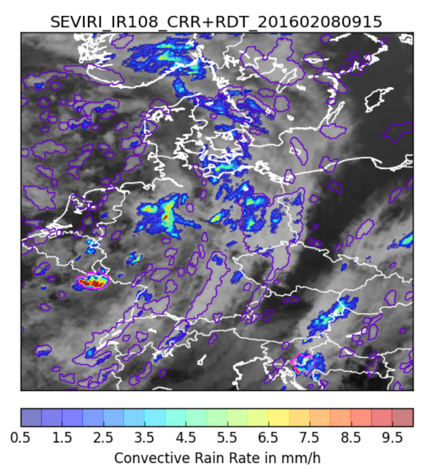
Off-line Data Mining

A medium ensemble size of 256 members is classified and featured due to temporal compliance with measurment tower observation (target forcast variable) over historical periods of different atmospheric conditions. This training phase of the SVM serves as the basis for the on-line data miner. Beforehand, the realtionship between the particle filter resampling interval and forecast skill has to be determined empirically (connected to feature engeneering), a yet unresolved issue in atmospheric data assimilation.

Observational data processing unit

Observations of high temporal frequency are processed to capture the temporal evolution of the atmosphere within an appropriate particle filter resampling period (off-line and on-line).

- **In-situ**: pressure observations as an indication of the time evolution of synoptic conditions + various measuring masts evaluated at hub height for a sufficiant boundary compliance
- **Satellite imagry**: to be included for the most sensitive phenomena of frontal system movement and convective system delevlopment



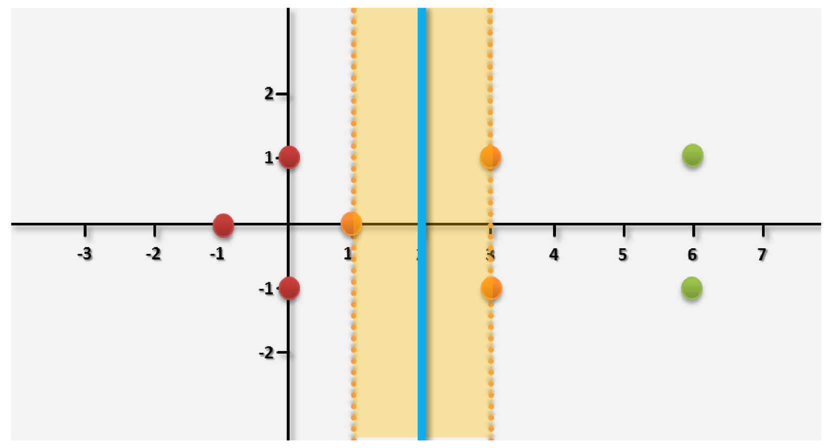
Satellite imagry from SEVIRI instrument mounted on the Meteosat Second Generation (MSG) satellite (images from 7th Feb 201601:15 and 8th Feb 2016 09:15 UTC, channel IR10.8). Automatic detection and tracking of cloud systems using SAFNWC software (pink: convective cells, violet: non-convective cells).

Data Mining Methodology

- Classification methodology trains a model of the data given training set *T*

$$T = (x_1, y_1), \ldots, (x_n, y_n)$$

- Supervised classification problem: Experts provide labels y_i data of WRF ensembles x_i quality
- Multi-class design enabling scientists to label with an increasing range of quality classes
- The trained model is then used with unseen WRF data to assign it to a quality class
- Depending on the quality class predicted by the model WRF, ensembles are canceled/continued
- Chosen algorithm to create a model are Support Vector Machines (SVM) with kernel methods



In this simplified 2D example of a two class problem (red = bad WRF ensemble members, greed = good WRF ensemble members), SVM achieve the optimal decision boundary between both classes. While many lines will separate both classes in this example, SVM will automatically learn via the training set the blue line as shown in the illustration. The interesting property of this blue line is that is offers the best generalization out of sample. In other words, once the training data has been used to train the model, the model will work quite well with unseen WRF ensemble members.

- Train a model with support vectors (cf. orange data in figure) is computationally complex
- SVM needs to find the best decision boundary (aka points most far away from existing points)
- It is a constraint optimization problem solved inherently with sequential minimal optimization
- The optimization problem aims to maximize the margin (above orange background color)

$$\min_{w,\xi_i,b}\left\{\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i\right\}$$

subject to

$$y_i(\langle\phi(\mathbf{x}_i), \mathbf{w}\rangle + b) \geq 1 - \xi_i \quad \forall i = 1, \ldots, n$$

$$\xi_i \geq 0 \qquad \forall i = 1, \ldots, n.$$

- The above formula is the dual notation of SVMs by performing minimization with constraints
- The generalization parameter C steers how much errors in the training process we allow
- This approach is a soft margin classifier with slack variables EPS as violations of margin
- The optimization algorithm identified the support vectors that define the decision boundary

Data Preparation & Initial Studies

- The first dataset available is providing wind tower features with overall 3584 (100%) samples
- We follow on approach to take 1/5 (717, 20%) for validation and 1/5 (717,20%) for testing
- This leaves training data with 2150 (60%) samples in order to create a model of the data
- Cross-validation is performed to optimize parameters, detailed accuracy results will follow

Acknowledgements

The authors gratefully acknowledge the Gauss Centre for Supercomputing (GCS) for providing computing time through the John von Neumann Institute for Computing (NIC) on the GCS share of the supercomputer JUQUEEN at Jülich Supercomputing Centre (JSC). GCS is the alliance of the three national supercomputing centres HLRS (Universitat Stuttgart), JSC (Forschungszentrum Julich), and LRZ (Bayerische Akademie der Wissenschaften), funded by the German Federal Ministry of Education and Research (BMBF) and the German State Ministries for Research of Baden-Wurttemberg (MWK), Bayern (StMWFK) and Nordrhein-Westfalen (MIWF).