Ensemble Forecast with Machine Learning Algorithms and Application to Air Quality

Vivien Mallet[1,2], Gilles Stoltz[3,4]

SAMA, IPSL
16 May 2008

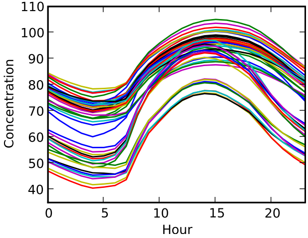
[1] INRIA
[2] CEREA, laboratoire commun ENPC - EDF R&D, Université Paris-Est
[3] CNRS & DMA, École normale supérieure
[4] HEC Paris

[1,2] Research carried out in the INRIA-ENPC joint project-team CLIME
and in the ENPC-EDF R&D joint laboratory CEREA

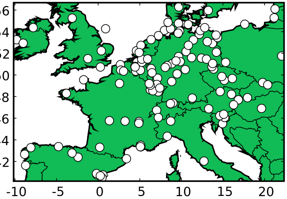
Support of the French Agence Nationale de la Recherche (ANR), under grant ATLAS (JCJC06_137446) “From Applications to Theory in Learning and Adaptive Statistics”

Ensemble Forecasting in Short

How to Take Advantage of Ensemble Simulations?



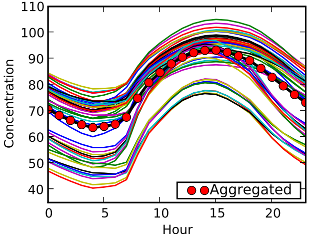
Ozone daily profiles of
a 48-member ensemble



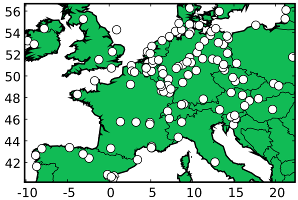
One ozone monitoring network

Ensemble Forecasting in Short

How to Take Advantage of Ensemble Simulations?



Ozone daily profiles of
a 48-member ensemble



One ozone monitoring network

Sequential aggregation of models, based on past observations and model predictions

Sequential Aggregation to Improve the Forecasts

Notations

Output of model m at time t and position x: $M_{m,t,x}$

Linear combination: $E_{t,x} = \sum_m \alpha_{m,t} M_{m,t,x}$ Observation: $O_{t,x}$

Notes

1. The weights do not depend on the location, to permit aggregation around the stations
2. The weights may or may not be constrained
3. Obviously, the weight $\alpha_{m,t}$ may depend on $O_{1,x}, \ldots O_{t-2,x}, O_{t-1,x}$ and $M_{m',1,x}, \ldots M_{m',t-1,x}, M_{m',t,x}$ (for all x and m')

Sequential Aggregation to Improve the Forecasts

Notations

Output of model m at time t and position x: $M_{m,t,x}$

Linear combination: $E_{t,x} = \sum_m \alpha_{m,t} M_{m,t,x}$ Observation: $O_{t,x}$

Best Model: the Reference to Beat

- $EB_{t,x} = M_{\widehat{m},t,x}$ where $\widehat{m}$ minimizes
  $\mathrm{RMSE}(M_{\widehat{m}}, O)^2 = \frac{1}{N_{\mathrm{step}} N_{\mathrm{station}}} \sum_{t,x} (M_{\widehat{m},t,x} - O_{t,x})^2$

The Simplest Methods

- $EM_{t,x} = \frac{1}{N_{\mathrm{model}}} \sum_m M_{m,t,x}$ (ensemble mean)
- The ensemble median

Least-Squares Methods

See also "superensembles" (Krishnamurti et al., 2000)

Best Combination from the Previous Forecasted Step

- $ELS_{t,x} = \sum_m \alpha_{m,t} M_{m,t,x}$,
  $\forall t \quad \alpha_{\cdot,t} = \mathrm{argmin}_\beta \sum_x \left(O_{t-1,x} - \sum_m \beta_m M_{m,t-1,x}\right)^2$

Smoothing Weights Time Evolution

- $ELS^P_{t,x} = \sum_m \alpha^P_{m,t} M_{m,t,x}$ where
  $\forall t \quad \alpha^P_{\cdot,t} = \mathrm{argmin}_\beta \sum_{t-P \le T < t,x} \left(O_{T,x} - \sum_m \beta_m M_{m,T,x}\right)^2$

Notes

1. Best *unconstrained* weight in a P-step learning period
2. Efficient, but no mathematical framework

Mallet and Sportisse, *Ensemble-based air quality forecasts: A multimodel approach applied to ozone*, JGR, 2006

Machine Learning Algorithms

Example: Exponentiated Gradient (Kivinen and Warmuth, 1997)

- Combination $EG_{t,x} = \sum_m \alpha_{m,t} M_{m,t,x}$
- Loss function: $L(EG_{t,\cdot}, O_{t,\cdot}) = \sum_x (EG_{t,x} - O_{t,x})^2$
- Weights update:

$$\forall m \quad \alpha_{m,t} = \exp\left(-\eta \sum_{T<t} \frac{\partial L(EG_{T,\cdot}, O_{T,\cdot})}{\partial \alpha_{m,T}}\right) \times \text{normalization}$$

- Choice: learning rate η
- Bound on the regret
  $\sum_{t \le N_{\text{step}}} L(EG_{t,\cdot}, O_{t,\cdot}) - \min_\alpha \sum_{t \le N_{\text{step}}} L(\sum_m \alpha_m M_{m,t,\cdot}, O_{t,\cdot}) \le \gamma N_{\text{station}} \sqrt{N_{\text{step}} \ln N_{\text{model}}}$
- $\text{RMSE}(EG, O)^2 - \min_\alpha \text{RMSE}(E_\alpha, O)^2 \le \gamma \sqrt{\frac{\ln N_{\text{model}}}{N_{\text{step}}}} \longrightarrow_{N_{\text{step}} \to \infty} 0$

Machine Learning Algorithms

Another Example: Ridge Regression

- Combination $RR_{t,x} = \sum_m \alpha_{m,t} M_{m,t,x}$
- Loss function: $L(RR_{t,\cdot}, O_{t,\cdot}) = \sum_x (RR_{t,x} - O_{t,x})^2$
- Weights:

$$\alpha_{\cdot,t} = \mathrm{argmin}_\alpha \left[\lambda \|\alpha\|_2^2 + \sum_{T<t,x} \left(\sum_m \alpha_m M_{m,T,x} - O_{T,x} \right)^2 \right]$$

- Choice: penalization λ
- Bound on the regret, for any $\alpha \in \mathbb{R}^{N_{\mathrm{model}}}$,
  $\sum_{t \le N_{\mathrm{step}}} L(RR_{t,\cdot}, O_{t,\cdot}) - \sum_{t \le N_{\mathrm{step}}} L(\sum_m \alpha_m M_{m,t,\cdot}, O_{t,\cdot}) \le \frac{\lambda}{2} \|\alpha\|_2^2 + K \sum_m \ln\left(1 + \frac{\mu_m}{\lambda}\right) \simeq \mathcal{O}(\ln N_{\mathrm{step}})$
- $\mathrm{RMSE}(RR, O)^2 - \min_\alpha \mathrm{RMSE}(E_\alpha, O)^2 \lesssim \mathcal{O}\left(\frac{\ln N_{\mathrm{step}}}{N_{\mathrm{step}}}\right) \xrightarrow[N_{\mathrm{step}} \to \infty]{} 0$

Machine Learning Algorithms: Three Variants

Example: Ridge Regression

$$\alpha_{\cdot,t} = \mathrm{argmin}_{\alpha} \left[\lambda\|\alpha\|_2^2 + \sum_{T<t,x} \left(\sum_m \alpha_m M_{m,T,x} - O_{T,x}\right)^2\right]$$

Windowing

- Restricted learning period
- Loses the theoretical bounds!

Discounting

- Vanishing influence of the past
- Delivers the best forecasts: this is the recommended method

Hybrid Methods (Meta-Learning)

- Including aggregated models in the ensemble
- Applying a learning method on the new ensemble
- Works well and preserves the theoretical bounds

Ensembles: Motivations and Strategy

Limitations of Deterministic Approaches

- High uncertainties: input data, parameterizations, numerical resolution, even bugs and user mistakes
- State dimension, 10^6–10^7, versus number of observations, 10^2
- (Over?)tuning
- A single forecast, even the best one, is uncertain

Technical Framework

- From *all-in-one* models to a *platform of models*
  - Fragmented model with alternative physical formulations, alternative numerical schemes and alternative input data
  - A single change or multiple changes in a reference model
- Polyphemus air quality modeling system

  `http://cerea.enpc.fr/polyphemus/`

Mallet, Quélo, Sportisse, Ahmed de Biasi, Debry, Korsakissok, Wu, Roustan, Sartelet, Tombette and Foudhil, *Technical Note: The air quality modeling system Polyphemus*, ACP, 2007

Building a Multimodel Ensemble (Polyphemus)

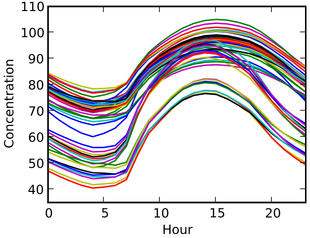
#	*Parameterization*	*Reference*	*Alternative(s)*
		Physical parameterizations	
1.	Chemistry	RACM	RADM 2
2.	Vertical diffusion	Troen & Mahrt	Louis
3.			Louis in stable conditions
4.	Deposition velocities	Zhang	Wesely
5.	Surface flux	Heat flux	Momentum flux
6.	Cloud attenuation	RADM method	Esquif
7.	Critical relative humidity	Depends on σ	Two layers
		Input data	
8.	Emissions vertical distribution	All in the first cell	All in the two first cells
9.	Land use coverage (dep.)	USGS	GLCF
10.	Land use coverage (bio.)	USGS	GLCF
11.	Exponent p in Troen & Mahrt	2	3
12.	Photolysis rates	JPROC	Depends on zenith angle
		Numerical issues	
13.	Time Step	600 s	100 s
14.			1800 s
15.	Splitting method	First order	Strang splitting
16.	Horizontal resolution	0.5°	0.1°
17.			1.0°
18.	Vertical resolution	5 layers	9 layers
19.	First layer height	50 m	40 m

Ensemble Simulations Bring Useful Information

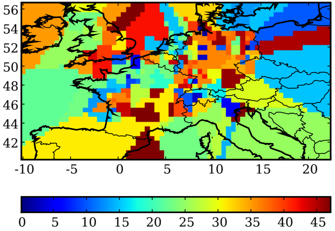
Study

- Ozone simulations at European scale during 4 months
- Resolution of 0.5°, 5 layers up to 3000 m
- ECMWF meteorological fields
- 48 members in the ensemble

Mallet, Sportisse, *Uncertainty in a chemistry-transport model due to physical parameterizations and numerical approximations: An ensemble approach applied to ozone modeling* , JGR, 2006



Ozone daily profiles from 48 members



Best model index

Sequential Aggregation Results (RMSE, $\mu g\, m^{-3}$)

Ozone Peaks

	Network 1	Network 2	Network 3
Best model	22.43	21.90	23.87
RR (disc.)	**19.45**	**18.12**	20.88
Hybrid *RR* (disc.)	19.62	18.26	**20.70**
E_α	19.24	18.16	20.26
E_{α_t}	11.99	8.47	12.46

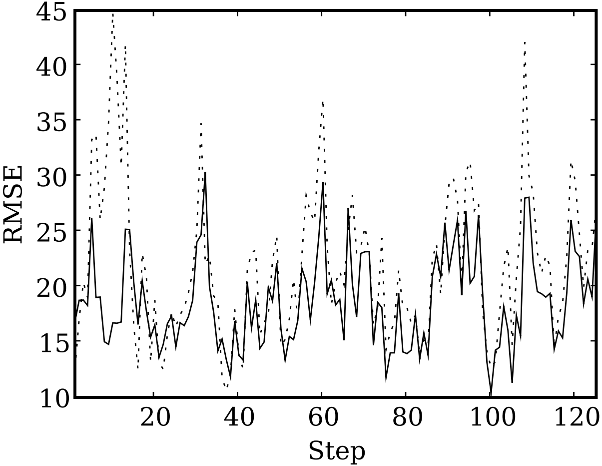
Ozone Hourly Concentrations

	Network 1	Network 2	Network 3
Best model	26.68	25.98	28.45
RR (disc.)	**22.02**	**22.82**	**22.27**
Hybrid *RR* (disc.)	22.26	23.02	22.45
E_α	22.80	23.52	23.19
E_{α_t}	14.88	12.03	15.32

Mallet, Stoltz, Mauricette, *Ozone ensemble forecast with machine learning algorithms*, submitted to JGR

Sequential Aggregation Results

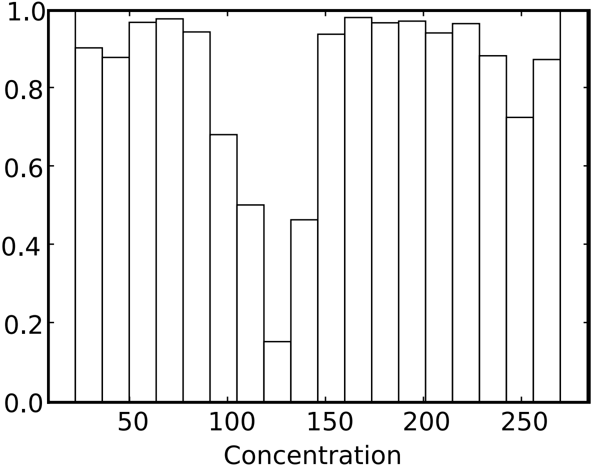
Robustness



RMSE ($\mu g\, m^{-3}$) against time

Sequential Aggregation Results

Robustness: Do Statistics Miss Extreme Events?



Frequency with which *RR* (discounted) beats the best model

Comparison with Classical Data Assimilation

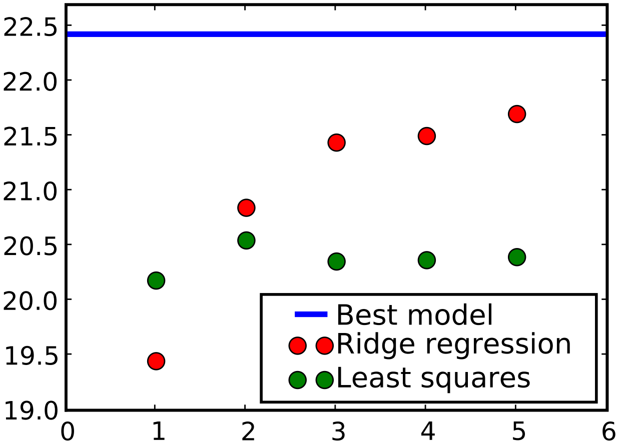
Two Advantages over Classical Data Assimilation

1. Parameters/inputs to the methods
   - Learning: very few inputs (learning rate, . . .), easy to determine
   - DA: error modeling
2. Support from theory
   - Learning: theory does hold in practice (whatever the sequence of predictions and observations may be)
   - DA: unmet assumptions (wrong error modeling, biased models, suboptimal algorithms)

One Drawback over Classical Data Assimilation

1. Spatial extent
   - Learning: weights only available at observation locations
   - DA: handles *n*-dimensional fields

Comparison with Classical Data Assimilation



D+1, D+2, . . . forecasts

Credit: Lauriane Saunier

Conclusion

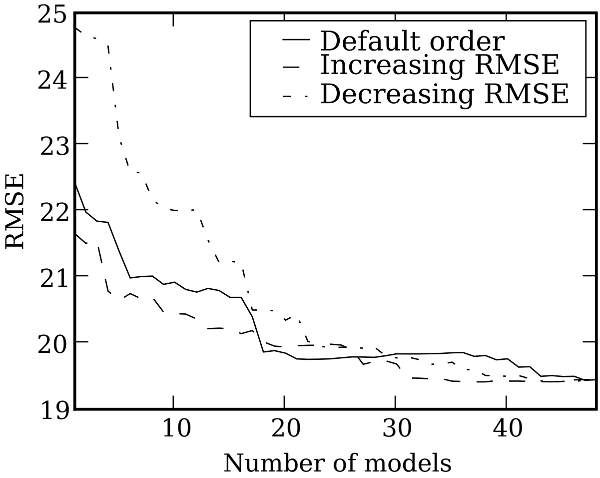
Sequential Aggregation in a Few Words

- Light assumptions and theoretical results
- Efficient and compatible with other approaches (through meta-learning)

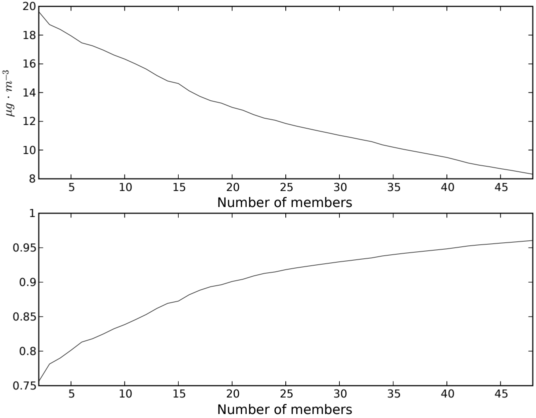
A Few Other Questions for Ensembles in Air Quality

- Spatial and multi-pollutant aggregation
- Which models to include or to build?
- Big issue: ensemble calibration for uncertainty estimation (Damien Garaud's PhD thesis)
- Probabilistic forecast

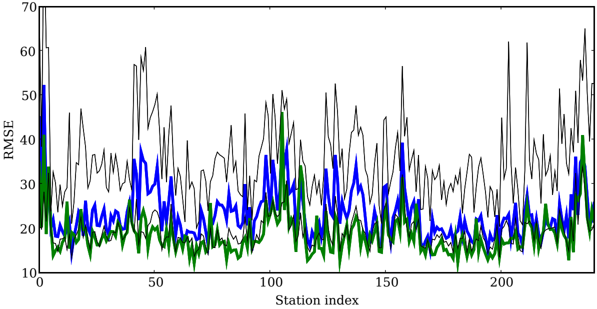
Number of Models (*RR* with Discount)



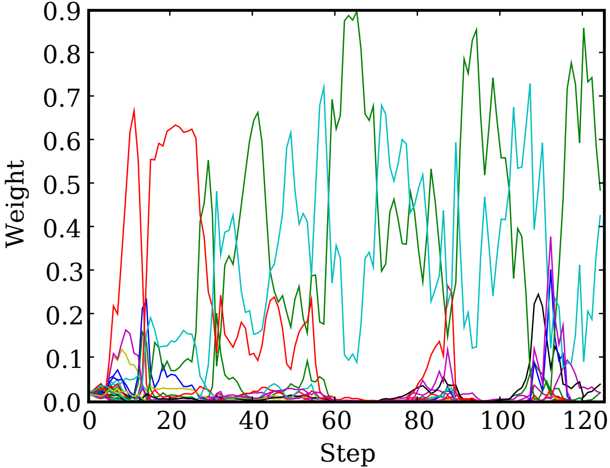
Number of Models (*ELS* a posteriori)



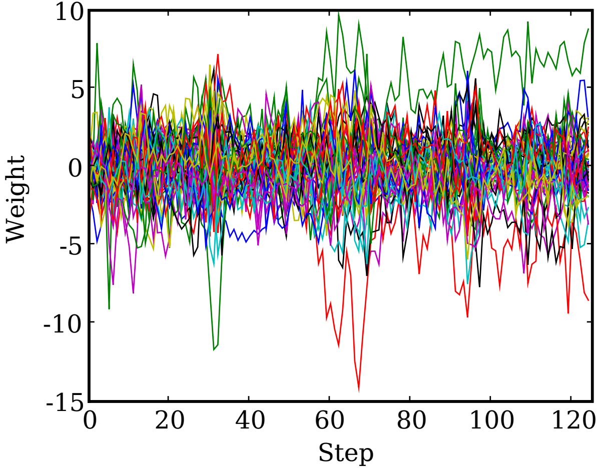
Per Station



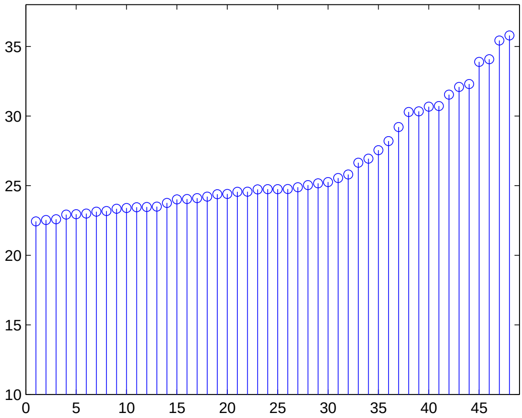
EG Weights



RR (with Discount) Weights



Models RMSEs



Ensemble Spread