Parameter Estimation in Size-Structured Aerosol Populations using Bayesian State Estimation

Matthew Ozon[a], Aku Seppänen[a], Jari Kaipio[a,b], & Kari Lehtinen [a,c]

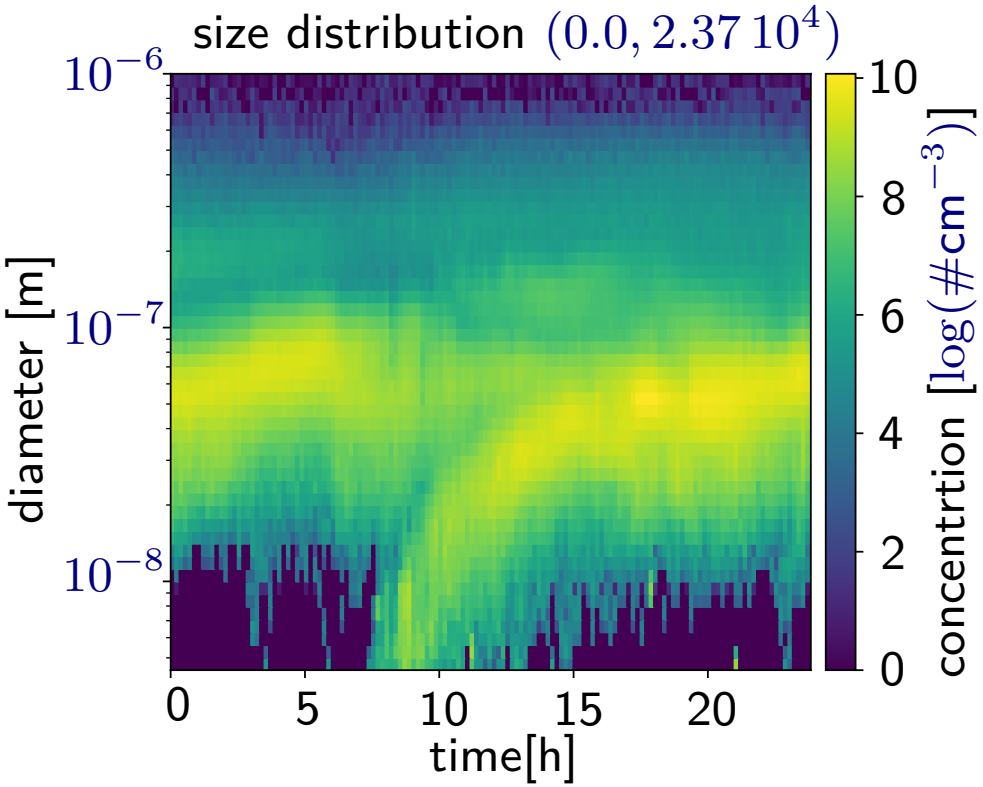
[a] Department of Applied Physics, University of Eastern Finland, Kuopio, Finland
[b] Department of Mathematics, Faculty of Science, University of Auckland, New Zealand
[c] Finnish Meteorological Institute, Kuopio, Finland

UNIVERSITY OF EASTERN FINLAND

SCIENCE
DEPARTMENT OF MATHEMATICS

FINNISH METEOROLOGICAL INSTITUTE

Conservation Principles, Data and Uncertainty in Atmosphere-Ocean Modelling 2019, Potsdam, April 3

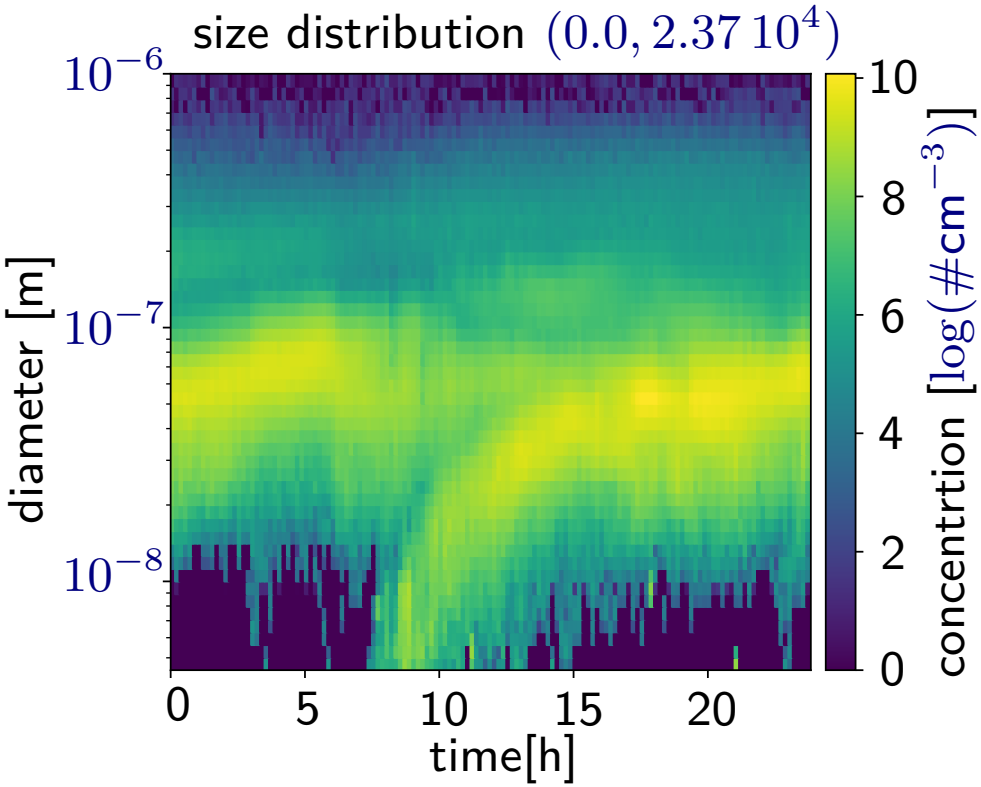


Goal: estimation of the characteristic parameters of aerosol systems along with their uncertainties, e.g.

- condensational growth rate
- nucleation rate
- linear loss rates

Why? Per se, because it is interesting, but also it is relevant, e.g. aerosols play a key role in cloud formation. Also, aerosols are known to disturb

- satellite data
- climate model (IPCC reports)
- radiative forcing



Evolution model of the form:

$$\frac{\partial u}{\partial t} = F(t, s, u; \theta)$$

If it is so important, methods should already exist, right? Well, it's not that simple! Indeed, some methods exist for estimating some parameters... but they are not satisfactory!

- manual recipes,
- no uncertainty estimation.

How to estimate the parameters? We have:

- data: time series of number concentration
- evolution model: GDE...

The perfect framework for an inverse problem.

How to describe the evolution of a population of aerosols whose characteristics depend on size? Size-structured Population Balance Equation (PBE).

$$\frac{\partial u}{\partial t} + \frac{\partial g u}{\partial s} = F(t, s, u; \theta) \tag{1}$$

This PBE is a scalar conservation law where u is the particle size density, g the growth rate and F is the term that describes the mechanisms that make the density evolve — it depends on some parameter θ. We refer to this equation as the General Dynamic Equation for aerosols, or simply GDE.

Note: without going into details, each aerosol particle cannot be described only by a size. By nature, particles are complex objects of different shape, size and chemical composition, but we model them as spherical objects of equivalent volume without considering chemistry. Therefore, it cannot be a perfectly precise model (source of uncertainty, potentially modelled by SPDE).

Condensational growth

Ambient vapor condenses onto the surface of particles, resulting in the growth of particles. Represented in the left hand side of equation (1) by g which is a speed.

Nucleation (and sources)

New particles can be added to an aerosol system either by adding existing particle or by formation of particles from the ambient vapor i.e. nucleation. In either case:

$$F(t, s, u; J) = J \tag{1}$$

is a simple source term.

Linear losses

Particles can be removed from the system by various mechanism such as sedimentation or wall deposition. It is well described by a linear damping term:

$$F(t, s, u; \lambda) = \lambda(t, v)u(t, v) \tag{2}$$

Coagulation

Particle of a given size may be created or removed by coagulation — when two particle collide and stick, they form a new particle.

$$F(t,v,u;\beta) = \underbrace{\int_{v_0}^{v-v_0} \beta_v(s,v-s)u(t,s)u(t,v-s)\mathrm{d}s}_{\text{coagulation source}} - \underbrace{u(t,v)\int_{v_0}^{\infty} \beta_v(v,s)u(t,s)\mathrm{d}s}_{\text{coagulation sink}}$$

where β_v is a collision frequency factor.

The continuous form of the GDE is not really suitable for our purpose (parameter estimation from time series of number concentration), hence we define:

$$\forall i \in [\![0, T]\!], k \in [\![1, K]\!], \quad N_i^k = \int_{\Omega_i} u(s, k\Delta_t)\mathrm{d}s \tag{1}$$

the number of particles in the size range Ω_i per unit of volume. Considering a logarithmic scale, and using the Euler time discretization scheme, we obtain the following time-and-size discrete evolution equations:

$$N_1^{k+1} = N_1^k + \Delta_t^k \left(J^k - \left(\frac{g_1^k}{\Delta_1} + \lambda_1 \right) N_1^k - C_1^{\mathsf{sink}}(N^k) N_i^k \right) + \varepsilon_1^k \tag{2}$$

$$\begin{aligned} N_i^{k+1} = N_i^k + \Delta_t^k \Bigg(& \frac{g_{i-1}^k}{\Delta_{i-1}} N_{i-1}^k - \left(\frac{g_i^k}{\Delta_i} + \lambda_i \right) N_i^k \\ & + C_i^{\mathsf{source}}(N^k) - C_i^{\mathsf{sink}}(N^k) N_i^k \Bigg) + \varepsilon_i^k \end{aligned} \tag{3}$$

Note that both discretization steps — time and size — add errors. The overall errors/uncertainties are encompassed in the terms ε_i^k.
Note that I leave out the parameter evolution for now.

DMA

This device acts as a selector of near monodisperse size distribution around a given size d_i; its size discrimination power determines the sets $\{d_i\}_{i\in[\![1,N]\!]}$ and $\{\Delta_i\}_{i\in[\![1,N]\!]}$. For each channel, we denote the time invariant kernel ψ_i, which models the efficiency of the device. The number concentration at the outlet of the DMA is approximated by:

$$z_i^k = \frac{1}{\Delta_t}\int_{t_0+(k-1)\Delta_t}^{t_0+k\Delta_t}\int_{\omega_i}\psi_i(s)u(s,t)\mathrm{d}s\mathrm{d}t + \iota_i^k = \varphi_i^k + \iota_i^k \tag{4}$$

where ω_i is the support of ψ_i — where it is not null — and ι_i^k accounts for the model uncertainties. The number of particle is then obtained by multiplying by the volume.

CPC

Let the number concentration of particles at the inlet of a CPC is z_i^k (coming from the i^{th} channel of the DMA), the output y_i^k is modeled by:

$$y_i^k = \frac{\tilde{y}_i^k}{V}, \quad \text{with} \quad \tilde{y}_i^k \sim \mathcal{P}\text{oisson}(V z_i^k) \tag{4}$$

where V is the volume of sample used in the CPC for counting. In most cases, the number of particle in the CPC is large enough ($V z_i^k > 20$), thus the Poisson distribution can satisfactorily be approximated by:

$$y_i^k \sim \mathcal{N}(z_i^k, \frac{z_i^k}{V}). \tag{5}$$

Note that from this model it is clear that the quality of the measurement is directly link to the volume of the sample.

SMPS

The full measurement device can be summarized by the model:

$$y_i^k = \frac{\tilde{y}_i^k}{V}, \quad \text{with} \quad \tilde{y}_i^k \sim \mathcal{P}\text{oisson}(V(\varphi_i^k + \iota_i^k)). \tag{4}$$

Assuming that the device operates under normal conditions — that is it is actually counting something — and that the DMA model has no flaw, then the model becomes:

$$y_i^k = \varphi_i^k + \frac{1}{V}\tilde{\iota}_i^k, \quad \text{with} \quad \tilde{\iota}_i^k \sim \mathcal{N}(0, V\varphi_i^k). \tag{5}$$

Note that the N_i^ks used in the GDE correspond to the case:

$$\psi_i(s) = \begin{cases} \frac{1}{|\omega_i|} & \text{if } s \in \omega_i \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

For the sake of clarity, we consider the above case, hence, the measurement model is:

$$y_i^k = N_i^k + \frac{1}{V}\tilde{\iota}_i^k \quad \text{with} \quad \tilde{\iota}_i^k \sim \mathcal{N}(0, VN_i^k) \tag{7}$$

Kalman Filter (KF)

Estimation of the expected state and its uncertainty

$$X^{k|k} = \mathbb{E}[X^k|\mathcal{Y}_k] \quad \text{and} \quad \Gamma^{k|k} = \mathrm{Cov}[X^k|\mathcal{Y}_k], \tag{8}$$

Initialization

Set $X^{0|0} \in \mathbb{R}^N$ *and* $\Gamma^{0|0} \in \mathbb{R}^{N\times N}$ *(Prior knowledge)*

Recursion

While $k \leqslant K$, do

Prediction

$X^{k|k-1} = F^{k-1}\left(X^{k-1|k-1}\right)$ *(state expectation)*

$\Gamma^{k|k-1} = \partial F^{k-1}\Gamma^{k-1|k-1}(\partial F^{k-1})^T + \Gamma_w^{k-1}$ *(state covariance)*

Calculation of Kalman's gain

$K^k = \Gamma^{k|k-1}(H^k)^T(H^k\Gamma^{k|k-1}(H^k)^T + \Gamma_v^k)^{-1}$

Filtering

$\Gamma^{k|k} = (I - K^kH^k)\Gamma^{k|k-1}$

$X^{k|k} = X^{k|k-1} + K^k(y^k - H^kX^{k|k-1})$

Update iterator

$k \Leftarrow k+1$

end(while)

Fixed Interval Kalman Smoother (FIKS)

$$X^{k|K} = \mathbb{E}[X^k|\mathcal{Y}_K] \quad \text{and} \quad \Gamma^{k|K} = \mathrm{Cov}[X^k|\mathcal{Y}_K], \tag{8}$$

Initialization

Run KF and store all variables

Set $X^K_{smo} = X^{K|K}$ *and* $\Gamma^K_{smo} = \Gamma^{K|K}$

$k \leftarrow K-1$

Recursion

While $k \geqslant 1$, do

Compute smoothing gain

$$K^k_{smo} = \Gamma^{k|k}(\partial F^{k+1})^T(\Gamma^{k+1|k})^{-1}$$

Smoothing

$$X^k_{smo} = X^{k|k} + K^k_{smo}\left(X^{k+1}_{smo} - X^{k+1|k}\right)$$

$$\Gamma^k_{smo} = \Gamma^{k|k} + K^k_{smo}\left(\Gamma^{k+1}_{smo} - \Gamma^{k+1|k}\right)(K^k_{smo})^T$$

Update iterator

$$k \Leftarrow k-1$$

end(while)

We have almost all the elements to run the algorithm:

- Evolution model of the number concentration
- Measurement model
- Data: simulation or measurement
- Initial guesses: may depend on the user

however, we still miss one part:

- Evolution model of the parameters

We'll have to create some based on what we know of the system.

Time invariant Some parameter are time invariant, so their time evolution model is a random walk:

$$k \geqslant 1, \quad p^{k+1} = p^k + \eta^k, \quad \text{with} \quad \eta^k \sim \mathcal{N}(0, \Gamma_\eta) \tag{8}$$

where Γ_η is the covariance of the model uncertainty.

Second order If a parameter is known to evolve smoothly with time, it can be modelled as a second order stochastic process such as:

$$G^k = \begin{bmatrix} p^k \\ p^{k-1} \end{bmatrix} = \begin{bmatrix} 2r_p & -r_p^2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} p^{k-1} \\ p^{k-2} \end{bmatrix} + \begin{bmatrix} \eta_k \\ 0 \end{bmatrix} = B(r_p)G^{k-1} + \begin{bmatrix} \eta^k \\ 0 \end{bmatrix} \tag{8}$$

where r_p is the smoothness lever and $\eta^k \sim \mathcal{N}(0, \sigma_\eta^2)$ with σ_η controlling the amplitude of the process. The latter is given by its covariance matrix defined by:

$$\Gamma_G^k = \mathrm{cov}(G^k) = B(r_p)\Gamma_G^{k-1}B(r_p)^T + \begin{bmatrix} \sigma_\eta^2 & 0 \\ 0 & 0 \end{bmatrix}. \tag{9}$$

We are interested in the asymptotic behavior, i. e. $\Gamma_G^k = \Gamma_G^{k-1} = \Gamma_G^\infty = \begin{bmatrix} \sigma_p^2 & c \\ c & \sigma_p^2 \end{bmatrix}$, and how the variance of η controls the variance of p, σ_p^2. By expanding the previous relation, we find that:

$$\sigma_\eta^2 = \sigma_p^2 \left(1 - r_p^2 \left(4 + r_p^2 \left(1 - \frac{8}{1 + r_p^2}\right)\right)\right), \quad c = \frac{2r_p}{1 + r_p^2}\sigma_p^2. \tag{10}$$

Size correlation Some parameters are distributed, yet, the size dependence may be unknown or only approximately known. For most parameter, it is safe to assume that size dependence is continuous, and even rather smooth.
Let $p^k \in \mathbb{R}^N$ follow a random walk described by eq. (8), the covariance Γ_η convey the size dependence information, and it can be constructed as:

$$\Gamma_\eta = \tilde{D}^{\frac{1}{2}} \bar{D}^{-\frac{1}{2}} \bar{\Gamma} \bar{D}^{-\frac{1}{2}} \tilde{D}^{\frac{1}{2}} \tag{8}$$

where $\bar{\Gamma}$ is the Toeplitz matrix build with the sequence $(\bar{\sigma}_i)_{i \in [\![1,N]\!]}$ which determines how the size dependence evolves with the size difference. We choose the sequence

$$\forall i \in [\![1, N]\!], \quad \bar{\sigma}_i = e^{\frac{1-i}{\delta}} \tag{9}$$

with δ so that only the first δ neighboring sizes significantly contribute to the evolution of one variable. The decay is exponential in size index, so it is actually linearly/polynomially decaying in the diameter space. The diagonal matrix $\bar{D} = \bar{\sigma}_1^2 I_N$ normalize the covariance $\bar{\Gamma}$ and the diagonal matrix $\tilde{D} = \mathrm{diag}([\sigma_{\eta,1}^2 \, \sigma_{\eta,2}^2 \ldots \sigma_{\eta,N}^2])$ scales the variance of each size.

What if we know the possible range of a parameter?

Lower bound

$$p = a + \frac{1}{\alpha} \log\left(1 + e^{\alpha\zeta}\right), \tag{10}$$

Range

$$p = a + \frac{b - a}{1 + \frac{1}{\alpha} e^{-\alpha\zeta}} \tag{11}$$

We use simulated data in order to evaluate the performance of the method.

- dense discretization of the size space
- non-approximated measurement model (Gaussian kernel and Poisson noise)

The dense discretization of the size space may lead to spurious oscillation (or diverge) if the following condition on the time step is not met:

$$0 < \Delta_t^k < \frac{1}{\max_i \{ \frac{g_i^k}{\Delta_i} + \lambda_i + C_i^{\mathsf{sink}}(N^k) \}}. \tag{12}$$

Nucleation event

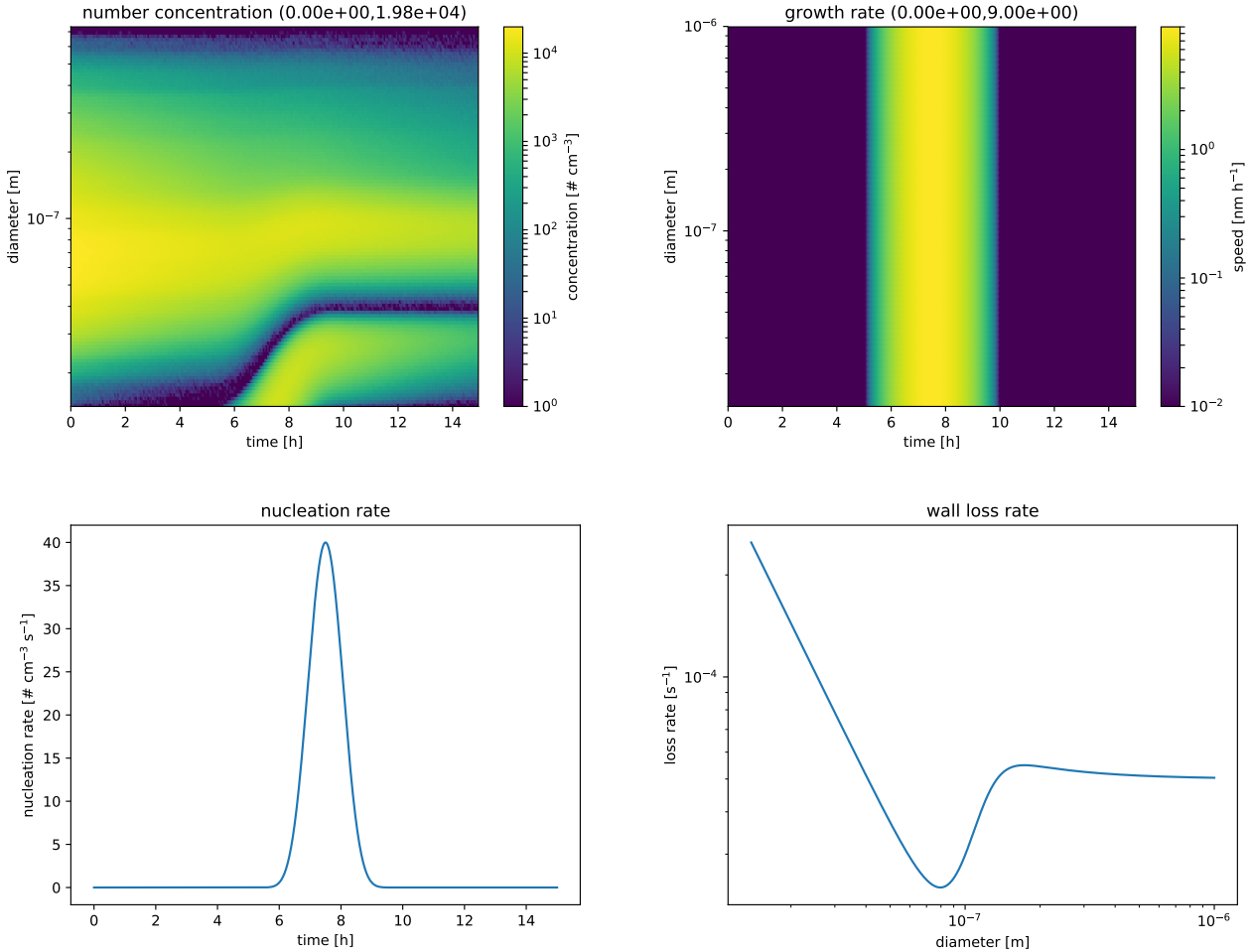


Figure: Simulation of a nucleation event. a) Number concentration contour plot, b) growth rate, c) nucleation rate at 14.1nm, and d) wall loss rate.

CLOUD simulation

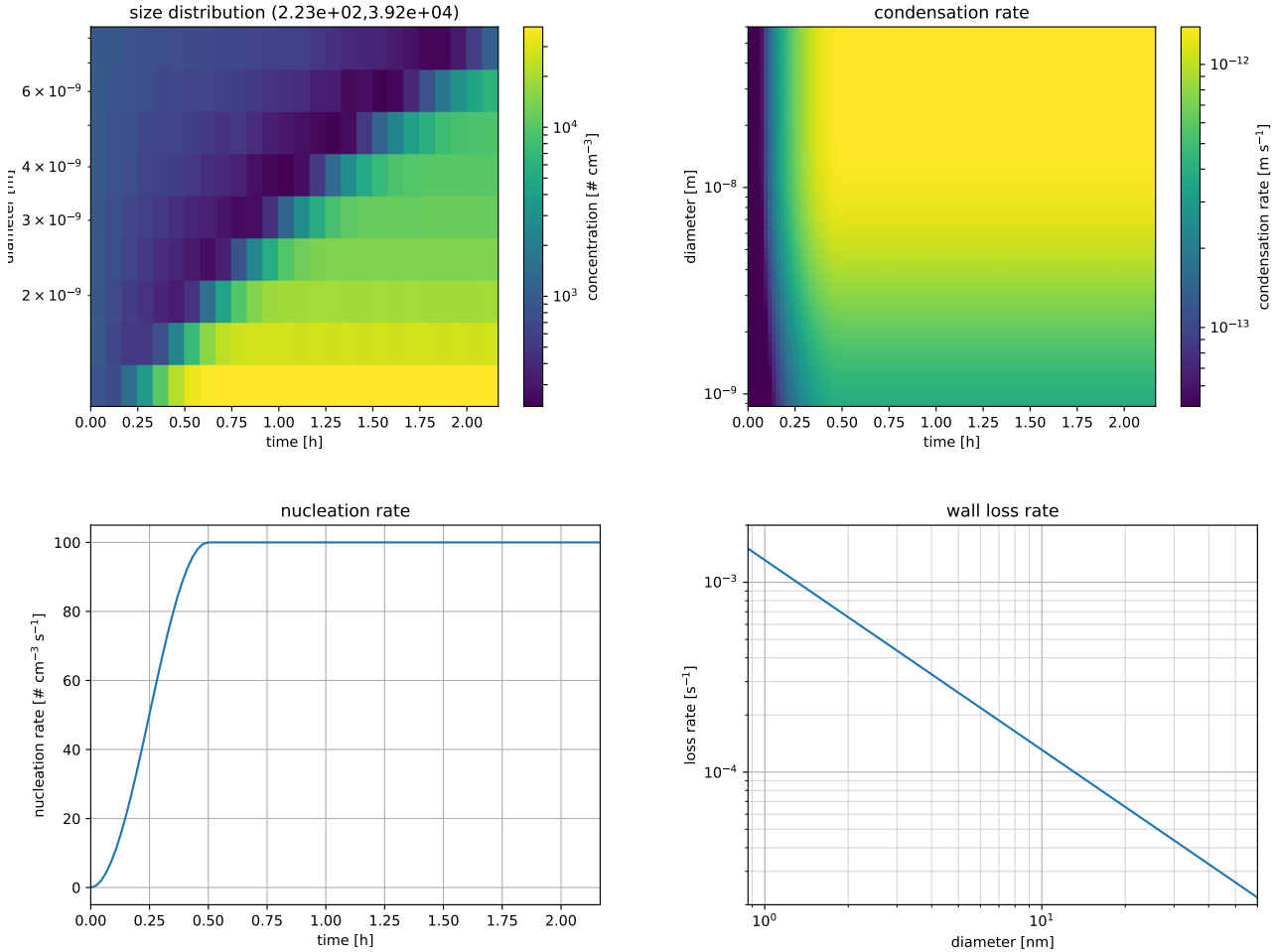


Figure: Simulation of a steady state. 1) Number concentration contour plot, 2) growth rate, 3) nucleation rate, and 4) wall loss rate.

Nucleation event

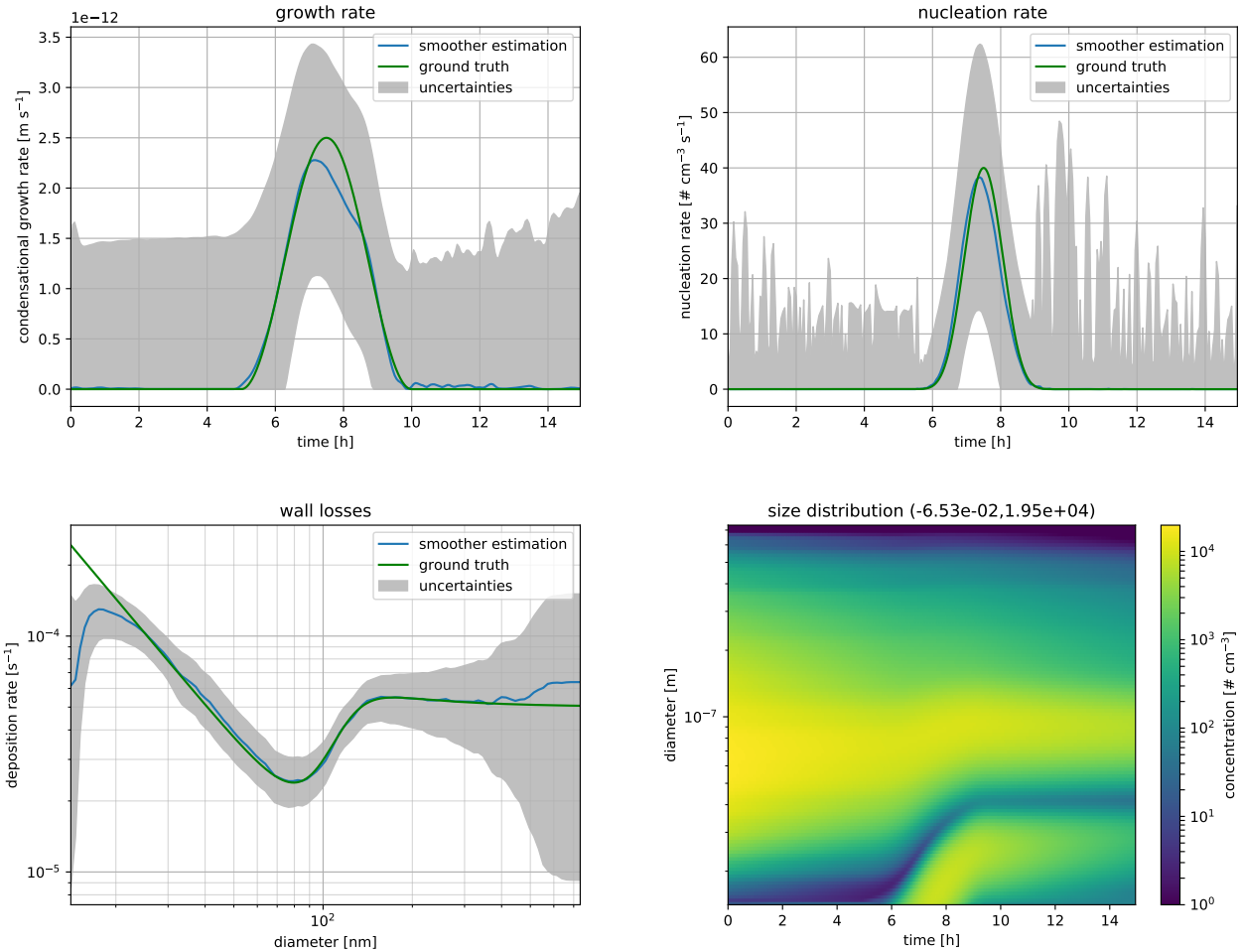


Figure: Estimation of the parameter of the GDE for aerosols from a simulated nucleation event data: 1) growth rate, 2) nucleation rate, 3) loss rate, and 4) number concentration.

CLOUD simulation

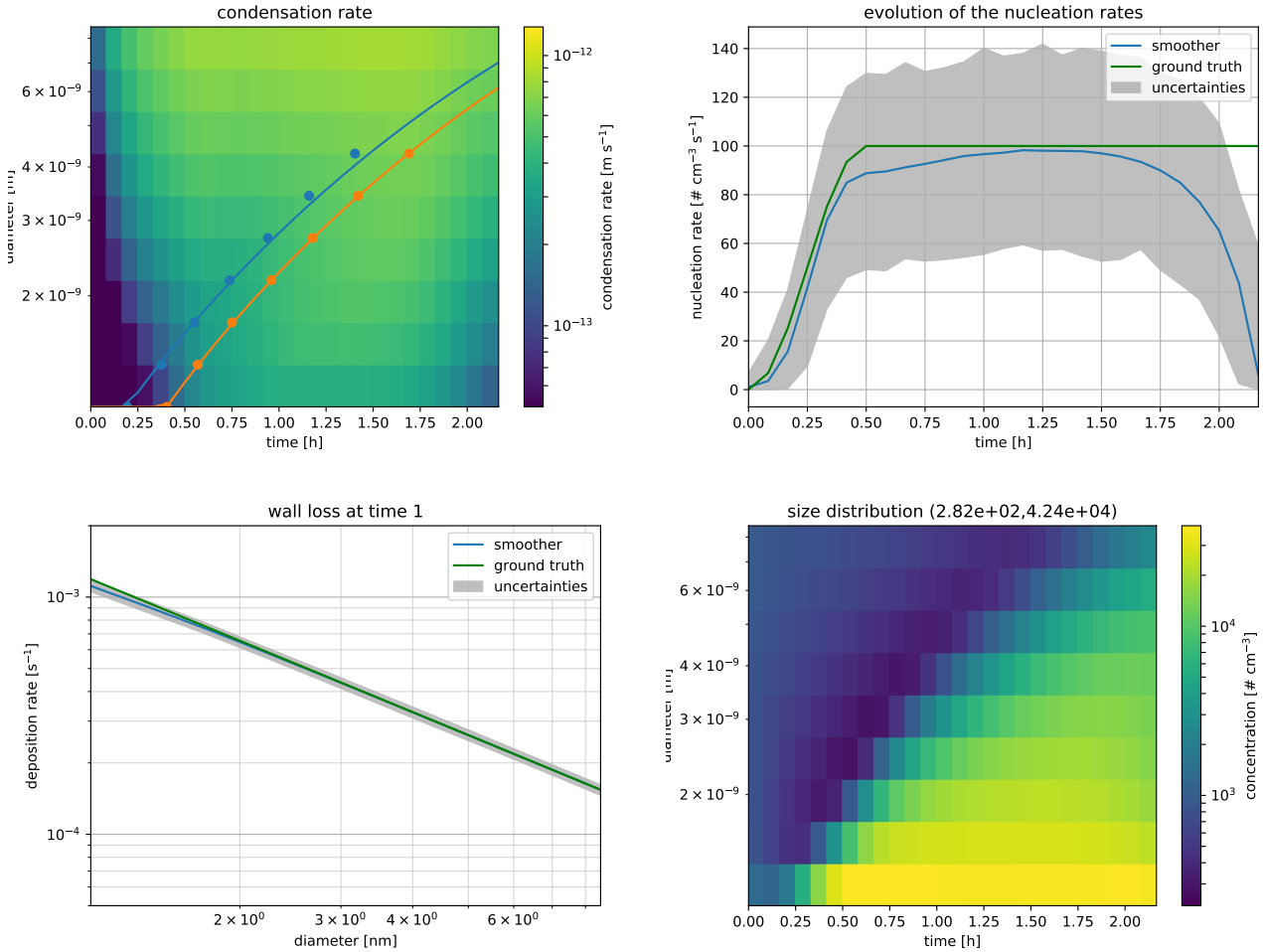


Figure: Estimation of the parameter of the GDE for aerosols from a simulated transition to steady state data: 1) growth rate, 2) nucleation rate, 3) loss rate, and 4) number concentration.

Message to take back home:

- Aerosols can disturb everything... at least the models
- The Fixed Interval Kalman Smoother is a suitable tool for the estimation of the GDE parameters (distributed or not) along with meaningful uncertainties
- The requirements for applying the method are "weak": 1) the model must be well approximated by their Jacobian, and 2) the errors can be approximated as gaussian
- Need surrogate evolution models for the parameters of interest (unless someone comes up with a physically relevant evolution model).

Thanks!