Comparison of ensemble-based data assimilation techniques for epidemiological forecasting and parameter estimation in an age-based compartmental SEIR model

Juan Ruiz [1,2], Santiago Rosa [3], Tadeo Cocucci[3,4], Manuel Pulido[4,2]

jruiz@cima.fcen.uba.ar

Virtual EnKF Workshop (7-11 June 2021)

[1]Centro de Investigaciones del Mar y la Atmósfera (CIMA, UBA-CONICET), Atmospheric and Oceanographic Science Department (DCAO), School of Exact and Natural Sciences (FCEyN), Universidad de Buenos Aires (UBA), Buenos Aires, Argentina

[2]Institut Franco-Argentin d'Estudes sur le Climat et ses Impacts, Unité Mixte Internationale (UMI-IFAECI/CNRS-CONICET-UBA), Buenos Aires, Argentina

[3]School of Mathematics, Physics and Astronomy (FAMAF), Universidad Nacional de Córdoba, Córdoba, Argentina.

[4]Institute of Modeling and Technology Innovation (IMIT, CONICET-UNNE), Department of Physics Exact Science School (FACENA), Universidad Nacional del Nordeste (UNNE), Corrientes, Argentina.

C I M A CONICET U B A

cnrs UMI-IFAECI (3351)

UNIVERSIDAD · DE · BUENOS · AIRES FACULTAD DE CIENCIAS EXACTAS Y NATURALES

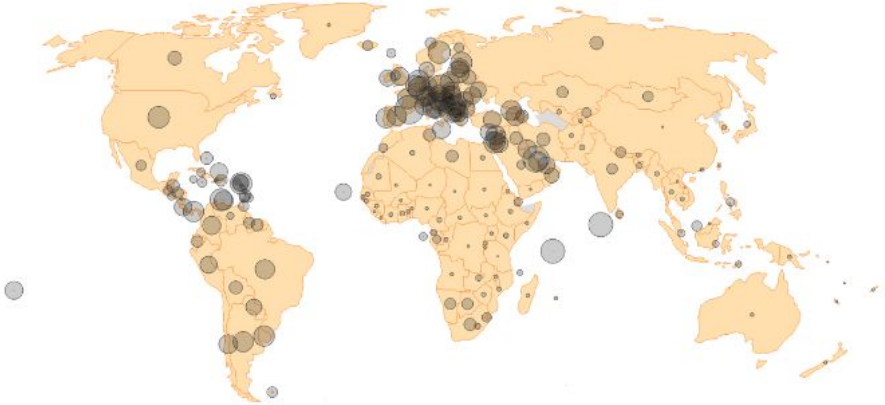
FaCENA - UNNE

UNC

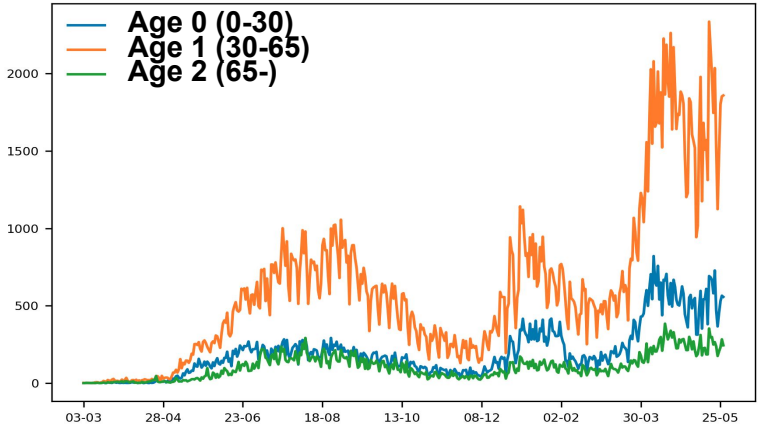
FAMAF

FAMAF Facultad de Matemática, Astronomía y Física

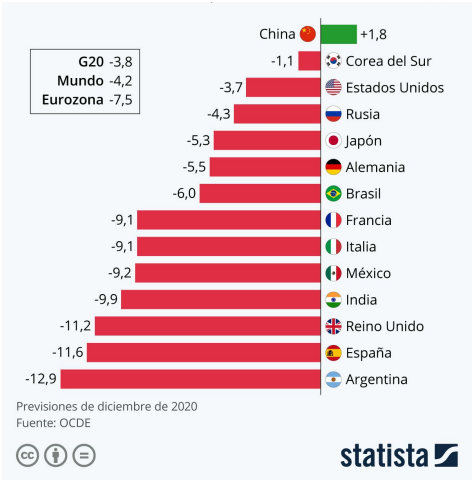
Motivation



Observations (daily infections)



GDP change (%) 2020

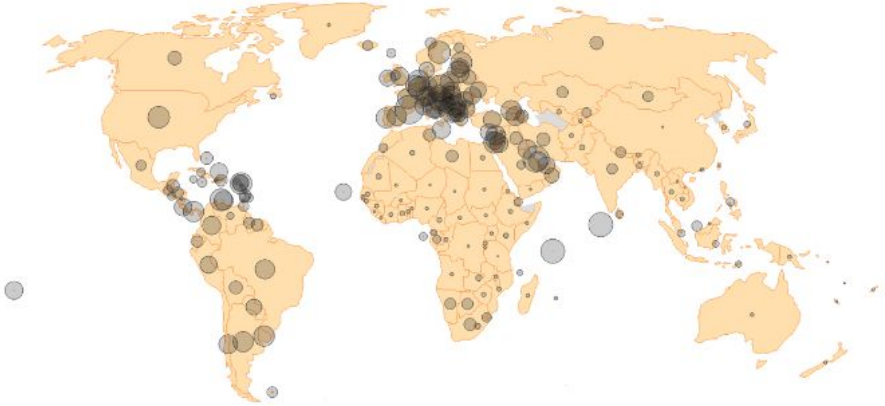


In 2020 the COVID-19 pandemic produce profound impacts worldwide (millions of infected, thousands of deaths, stress and collapse of health systems, long lockdowns, job losses, GDP falls, mobility constraints, among many others).

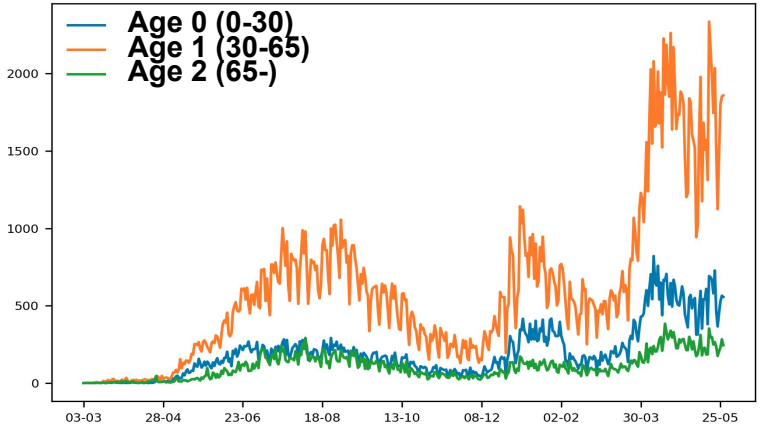
Currently South America is struggling with second/third waves of infection, with a percentage of vaccinated population which is still not enough to prevent these waves and their impact upon the society.

In this work we aim to use a simple epidemiological model and data assimilation techniques to provide a monitoring of epidemiological parameters as well as short range forecasts of the spread of the disease.

Motivation



Observations (daily infections)

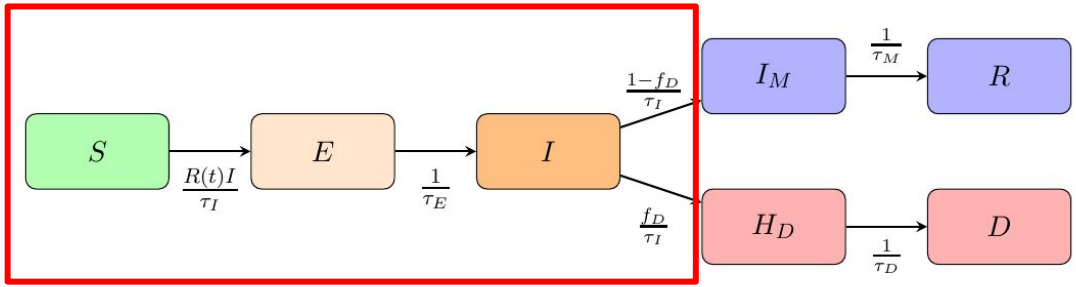


Previous works implemented DA for epidemiological models of diseases like influenza (e.g. Pei et al. 2018, Shaman et al., 2013, Hickmann et al. 2015) and other diseases (e.g. Passeto et al. 2017). Similar approaches has been applied to spread of COVID-19 (e.g. Evensen et al. 2021, Li et al. 2020, Ghostine et al. 2021).

Most works consider homogeneous populations which do not allow to investigate how different population groups interact with each other. Splitting the population into different groups by regions, ages, etc allows to investigate more complex effects like the impact of particular activities (e.g. schools) or the impact of strategies like selective lockdowns. Evensen et al. 2021 uses a priori knowledge of interactions among different age groups, Pei et al. 2018 the impact of the interactions among different regions.

We investigate how the interaction from different age groups can be obtained from the data using joint state and parameter estimation techniques.

Model description: Compartmental SEIRD model



$$\frac{dS_i}{dt} = -\sum_{j=0}^{j=n_a} \frac{C_{i,j}(t)}{\tau_I} \frac{I_j}{N_j} S_i$$

$$\frac{dE_i}{dt} = \sum_{j=0}^{j=n_a} \frac{C_{i,j}(t)}{\tau_I} \frac{I_j}{N_j} S_i - \frac{E_i}{\tau_E}$$

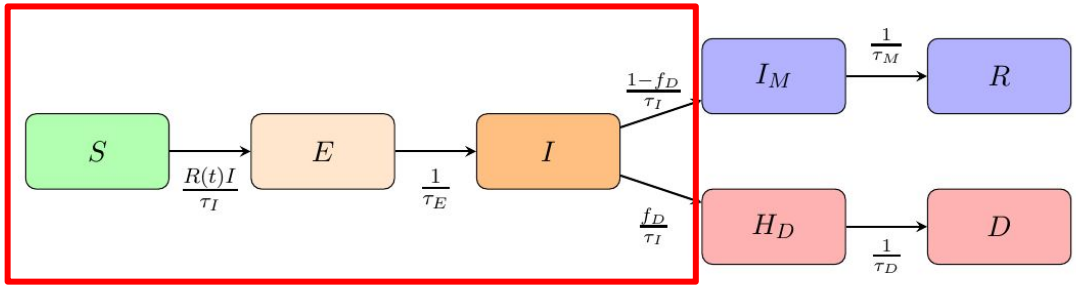
$$\frac{dI_i}{dt} = \frac{E_i}{\tau_E} - \frac{I_i}{\tau_I}$$

S: susceptibles, **E**: exposed, **I**: infected

$C_{i,j}(t)$ are the components of the contact - infection matrix. They depend upon the number of interactions of people within each group and among different groups as well as on the probability of infection (which depends on the disease).

τ_E, τ_I are time scales for the transition from **E** to **I** and from **I** either to mild or severe infections.

Model description: Compartimental SEIRD model



$$\frac{dS_i}{dt} = -\sum_{j=0}^{j=n_a} \frac{C_{i,j}(t)}{\tau_I} \frac{I_j}{N_j} S_i$$

$$\frac{dE_i}{dt} = \sum_{j=0}^{j=n_a} \frac{C_{i,j}(t)}{\tau_I} \frac{I_j}{N_j} S_i - \frac{E_i}{\tau_E}$$

$$\frac{dI_i}{dt} = \frac{E_i}{\tau_E} - \frac{I_i}{\tau_I}$$

The timescales τ_E, τ_I introduce a **time lag** between the time of the infection and time at which infections are detected (observed).

Previous works mostly uses filter approaches for estimating the parameters. Evensen et al. (2021) introduces a smoother approach to deal with this issue.

Another goal of this work is to compare the performance of the filtering approach and the smoother with different time-scales in estimating a time-dependent contact matrix.

Joint state-parameter estimation methods

We use two different joint state and parameter estimation techniques:

- Ensemble Kalman filter (EnKF)
- Ensemble Smoother with Multiple Data Assimilation (ESMDA)

State augmentation in the Ensemble Kalman filter:

In this case we implement a sequential Ensemble Transform Kalman Filter (ETKF) including the state variables (**S**,**E**,**I**,**R** and **D**) and the parameters required to estimate the contact matrix.

On each time step, we are approximately solving based on the Kalman filter hypothesis:

$$p(x_k, \theta_k | y_k) \propto p(y_k | x_k, \theta_k) p(x_k, \theta_k)$$

$$p(y_k | x_k, \theta_k) \approx \mathcal{N}(\overline{\mathcal{H}(x_k)}, R) \qquad x_k, \theta_k = M_{k-1,k}(x_{k-1}, \theta_{k-1})$$

Initial state and parameter ensembles (at k=0) are sampled from Gaussian distributions with known mean and standard deviation.

The filter assimilates observations at a daily frequency (which is the frequency at which observations are usually available).

A **random walk** is used as a dynamical model for the parameters. The random walk avoids the collapse of the filter acting in a similar way as additive inflation in the parameter space.

No additional inflation is performed on the parameters or state variables.

Ensemble Smoother with Multiple Data Assimilations (Evensen et al. 2021)

This is a smoother approach similar to the one used in Evensen et al. 2021, in which longer assimilation windows can be considered. At the end the time dependent parameters as well as the state variables at the beginning of the window are recovered.

$$p(x_0, \theta_{1:K}|y_{1:K}) \propto p(y_{1:K}|x_0, \theta_{1:K})p(x_0, \theta_{1:K})$$

The solution is approximated with a tempered (iterative) localized Kalman smoother approach,
On each iteration we solve

$$p(x_0^{i+1}, \theta_{1:K}^{i+1}|y_{1:K}) \propto p(y_{1:K}|x_0^i, \theta_{1:K}^i)p(x_0^i, \theta_{1:K}^i)$$

This step is solved using a **time-localized 4D-Ensemble Transform Kalman Smoother (4D-LETKS)** and assuming:

$$p(y_{1:K}|x_0^i, \theta_{1:K}^i) \approx \mathcal{N}(\overline{\mathcal{H}(M(x_0^i, \theta_{1:K}^i))}, \gamma_i^{-1}R) \qquad \sum_{i=0}^{i=n_t} \gamma_i = 1$$

Note that this method uses the model as a **strong constraint**. The evolution of the state variables within the assimilation window is determined by the state at the beginning of the window and the estimated parameters.

The prior for the model parameters is set through a random walk with known correlation and noise. The ***γ*** coefficients increases exponentially (the first iteration steps assimilates a smaller fraction of the information contained in the observations).

Estimated parameters

We assume that the contact-infection matrix is unknown and time dependent. We also split the population into 3 age groups. We tested approaches to model the time dependent contact matrix:

Full C

We estimate a total of 6 time-dependent parameters which are the maximum degrees of freedom of a contact matrix for 3 age groups (assuming that the probability of infection is the same for all the groups).

$$C = \begin{bmatrix} C_{1,1}(t) & C_{1,2}(t)N_2 & C_{1,3}(t)N_3 \\ C_{1,2}(t)N_1 & C_{2,2}(t) & C_{2,3}(t)N_3 \\ C_{1,3}(t)N_1 & C_{2,3}(t)N_2 & C_{3,3}(t) \end{bmatrix}$$

Diagonal C

We consider a diagonal C meaning that there are no interaction among groups. In this case we estimate 3 time-dependent parameters corresponding to the effective reproductive numbers for each group.

$$C = \begin{bmatrix} C_{1,1}(t) & 0 & 0 \\ 0 & C_{2,2}(t) & 0 \\ 0 & 0 & C_{3,3}(t) \end{bmatrix}$$

Well mixed C

The structure of C is fixed in time and represents a “well mixed” population in which the number of contacts among different groups depends on their population. The total quantity of contacts is modulated by 1 time-dependent parameter.

$$C = C^*(t) \begin{bmatrix} \frac{N_1}{N_T} & \frac{N_2}{N_T} & \frac{N_3}{N_T} \\ \frac{N_1}{N_T} & \frac{N_2}{N_T} & \frac{N_3}{N_T} \\ \frac{N_1}{N_T} & \frac{N_2}{N_T} & \frac{N_3}{N_T} \end{bmatrix}$$

“A priori” C

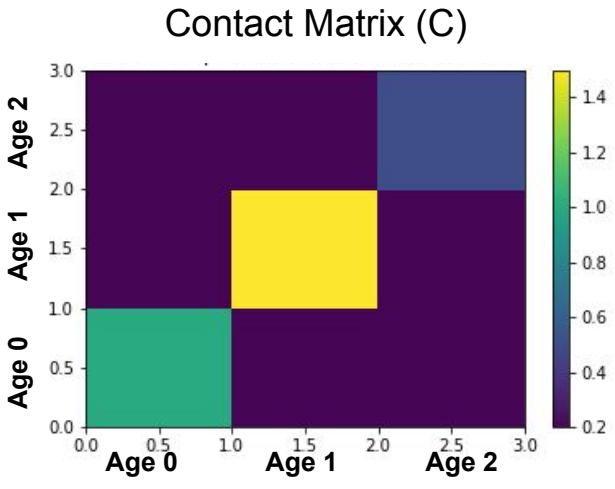
In this case we assume that the structure of C is known a priori, is different from the well-mixed distribution and is constant in time. The number of contacts is modulated by a 1 time-dependent parameter as in Evensen et al. (2021).

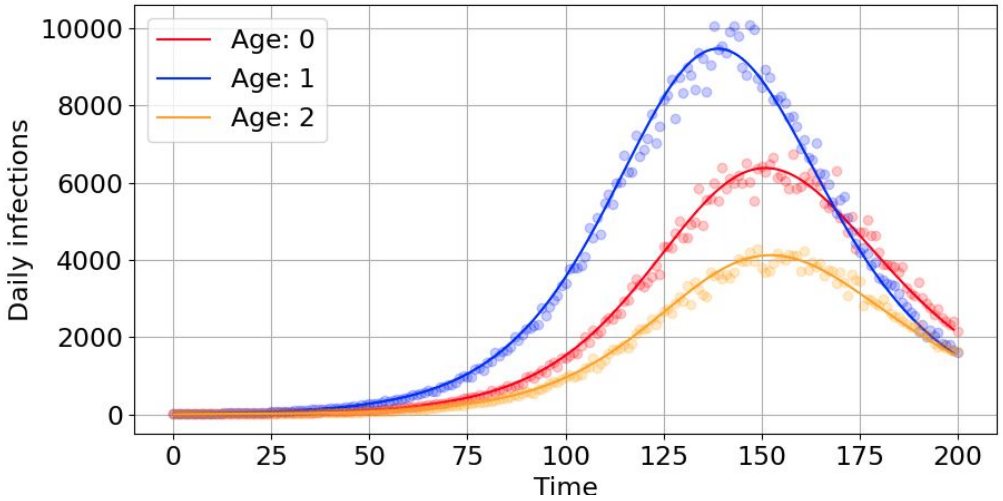
$$C = C^*(t)C_p$$

Experiments with simulated observations

We performed experiments using simulated observations in a twin model settings to evaluate the identifiability of the parameters and the performance of the different techniques.

The nature run - constant contact matrix experiment





- True contact matrix is constant in time.
- We consider 3 age groups ([0-30], [30-65] and [65-]) with 1.0e6 people each.
- Observations are generated adding Gaussian random noise to the model states. The observation error variance is a function of the number of daily infections.
- Only observations of the total number of infected population are assimilated in these experiments.

Parameter estimation experiments: Contact matrix parametrization

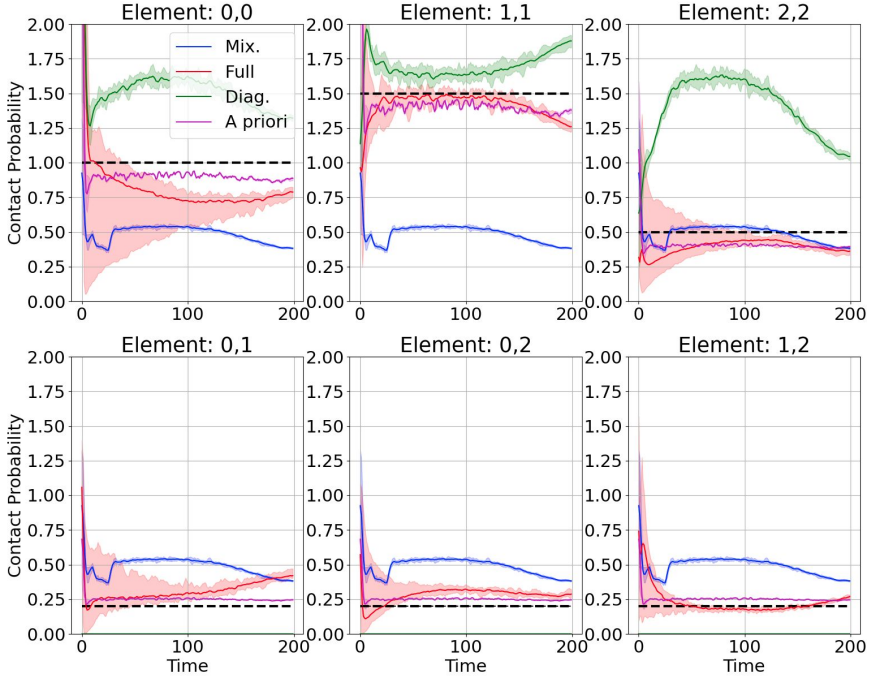
We perform a set of 4 experiments using the 4 contact matrix parametrization approaches:

- Full C
- Diagonal C
- Well mixed
- A priori C

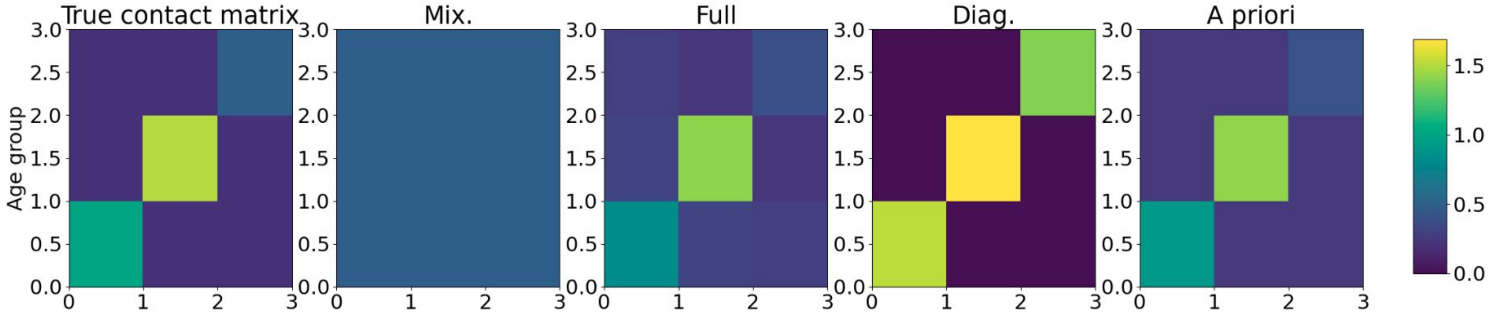
For this set of experiments we use the EnKF:

- Daily observations of total infected people for each age group. Observation error standard deviation equal to 5% of the true value.
- 500 ensemble members.
- Contact matrix parameters randomly initialized.
- Total assimilation period 200 days.
- All the experiments has been repeated 10 times changing the realization of the observation error and the initial values for the parameters.

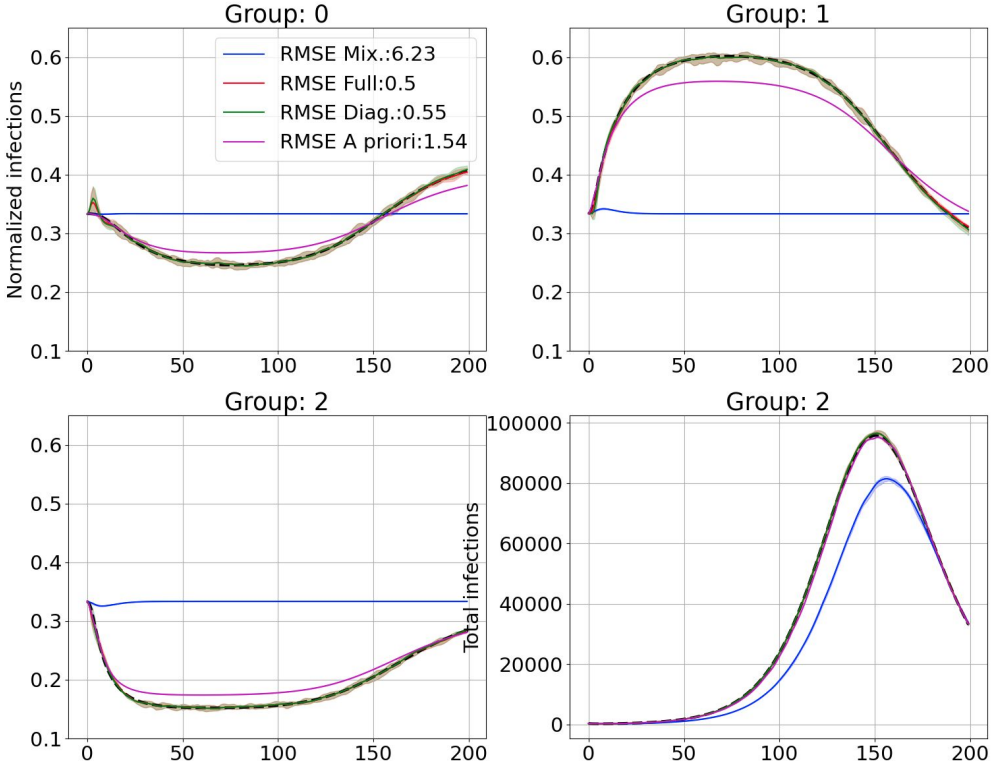
Parameter estimation experiments: Contact matrix parametrization

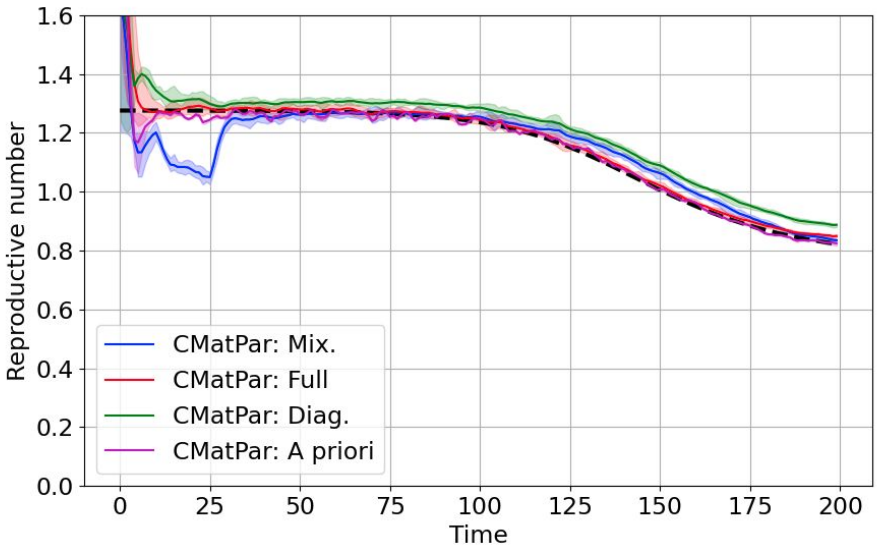


- The full matrix estimation experiment captures the main properties of the true contact matrix.
- The full matrix experiment exhibits the largest variability among experiment suggesting that the parameters are less identifiable.
- During the period of exponential growth the elements of the **Diag** parametrization converge to the same value.
- The “a priori” estimation is (as expected) the closest to the true parameters.



Parameter estimation experiments: Contact matrix parametrization

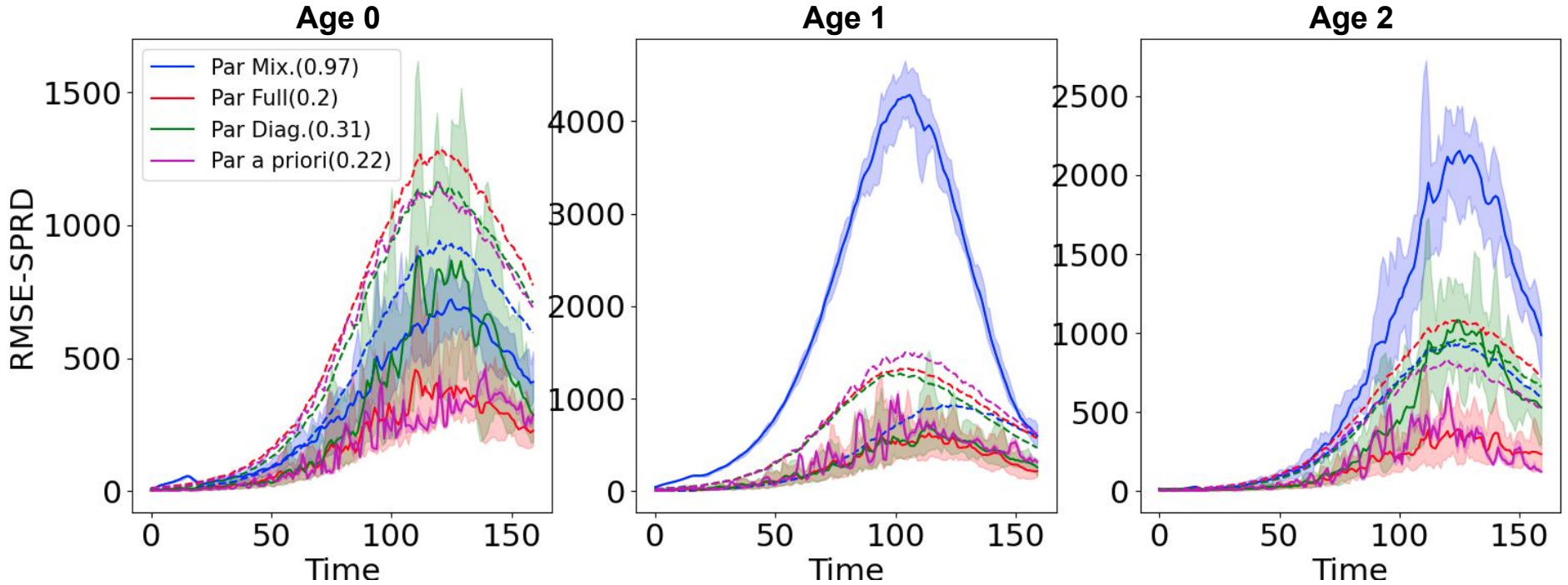




All the parametrizations did a good job in estimating the **effective reproductive number** (maximum eigenvalue of the NGM).

Full and **Diag.** parametrizations better capture the distribution of infected people among the different age groups. Although the **a priori** estimation is close to the true contact matrix, the distribution of infected people among the different groups is biased (this is much worse in the case of the well mixed parametrization). We believe this bias is a consequence of a wrong specification of the leading eigenvector.

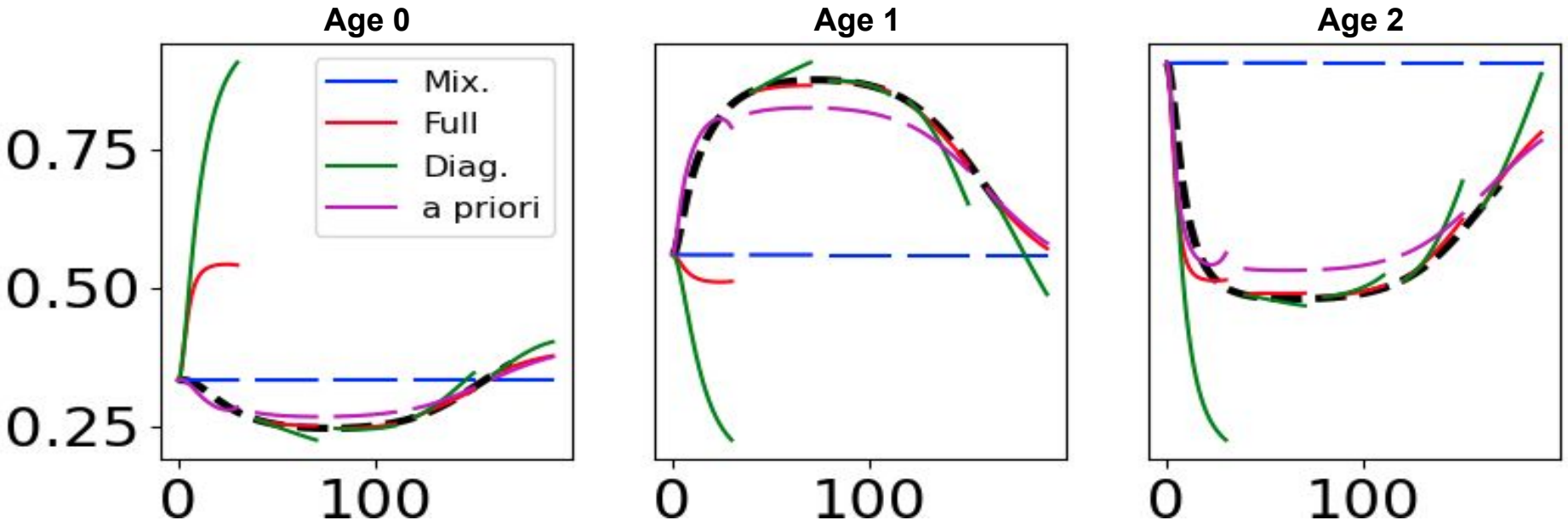
Parameter estimation experiments: Contact matrix parametrization



- Forecast produced by **Full** estimation are the best, followed closely by the **a priori** and **Diag.** parameterizations. The **well mixed** case produce a much worse forecast.

- Differences among parameterizations are larger after the maximum of infections is reached. Probably because in this part the orientation of the infection vector changes faster. Also differences among the parametrizations are larger for Age 0 and 2 since are the ones most affected by off-diagonal interactions.

Parameter estimation experiments: Contact matrix parametrization



- The full parametrization is the one that better captures the changes in the distribution of the infection among different groups during the forecast. The advantage is clearer during periods of rapid change in this distribution.

- While diag can synchronize this distribution based on the observation, it fails to predict future changes in this distributions probably due to not considering inter group interactions (off-diagonal terms of the contact matrix).

- The a priori specification produces a good forecast but with an initial bias in how infections are distributed among different groups.

Parameter estimation experiments: Parameter estimation techniques

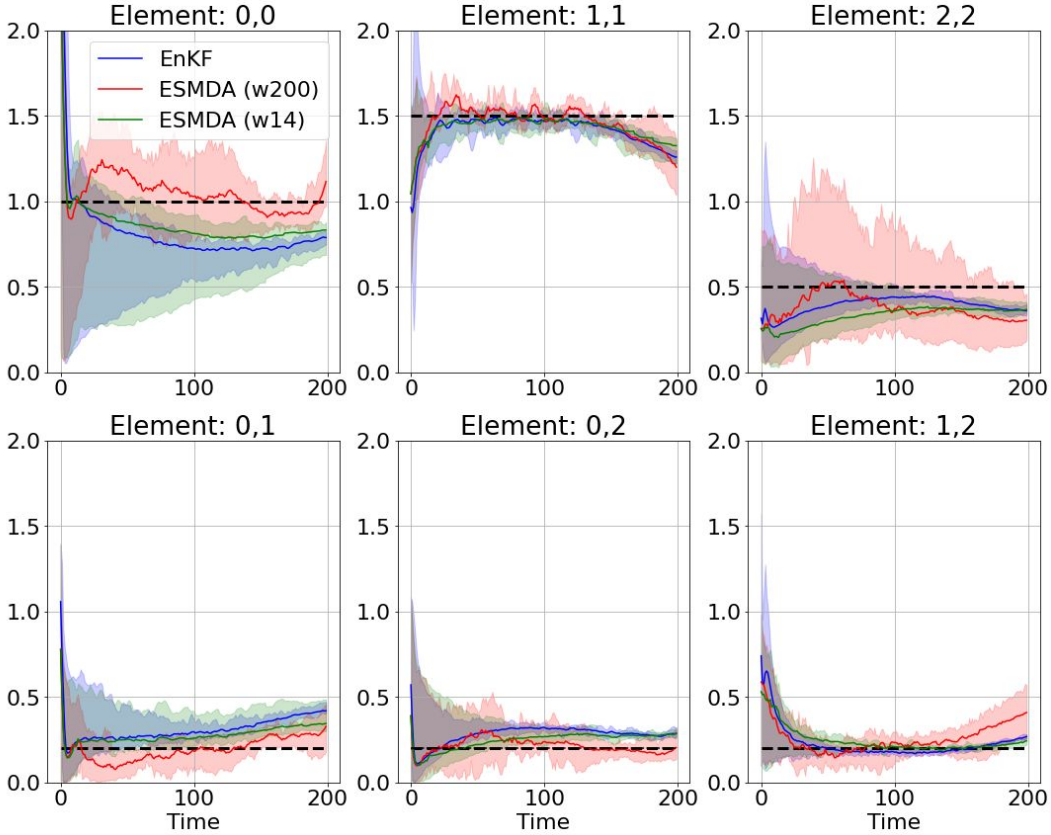
We perform a set of 3 experiments using 3 different data assimilation techniques:

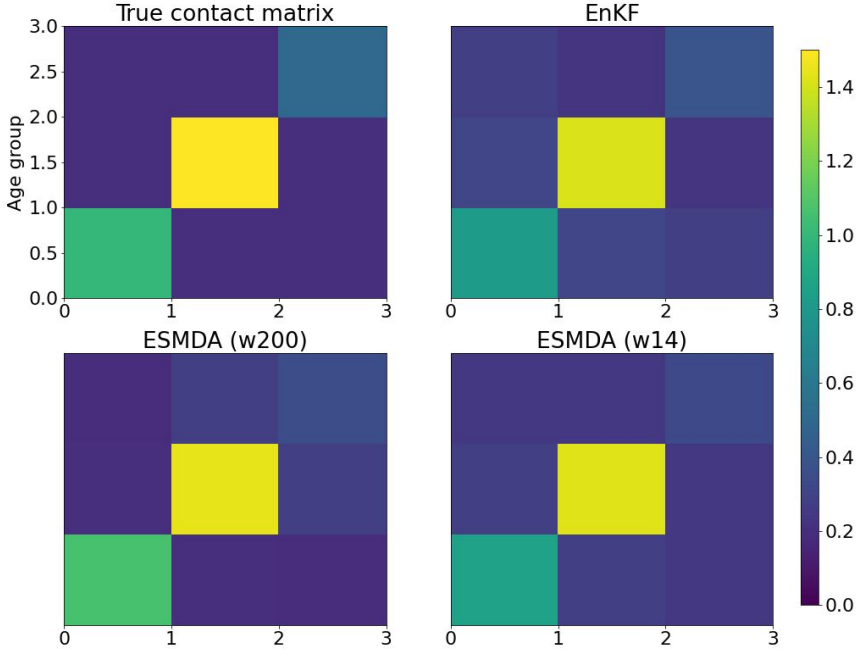
- **EnKF** (as in the previous experiments)
- **ESMDA** with 200 day window.
- **ESMDA** with 14 day window.

For this set of experiments we use:

- Daily observations of total infected people for each age group. Observation error standard deviation equal to 5% of the true value.
- 500 ensemble members.
- Contact matrix parameters randomly initialized and a **Full** matrix parametrization.
- 50 iterations in the **ESMDA**. Also **LETKS** is implemented with a 24 days localization in both cases.
- All the experiments has been repeated 10 times changing the realization of the observation error and the initial values for the parameters.

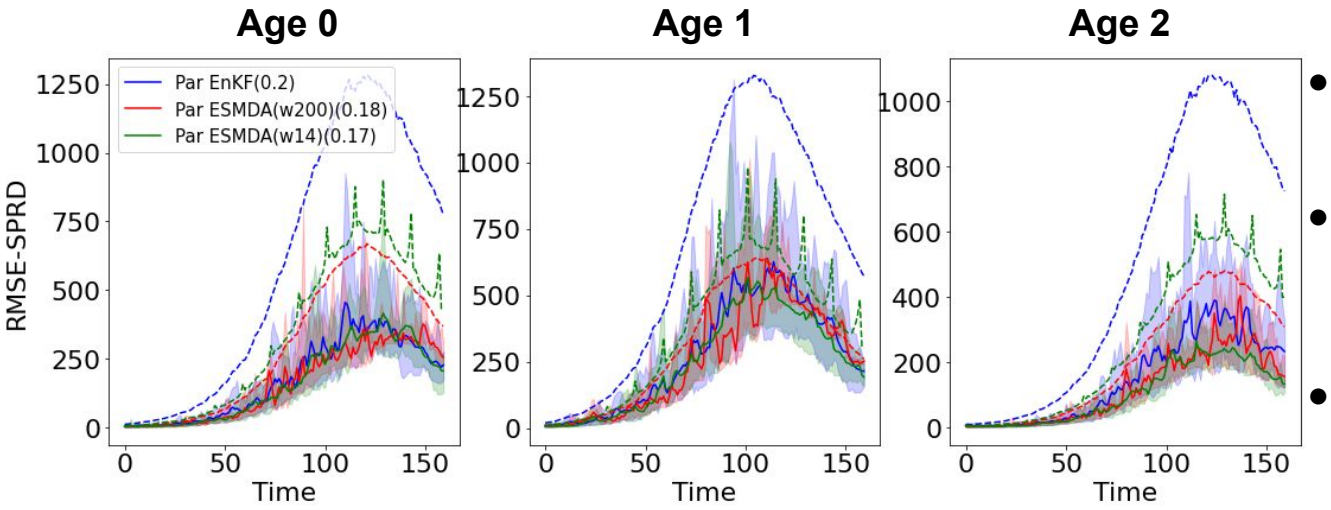
Parameter estimation experiments: Parameter estimation techniques

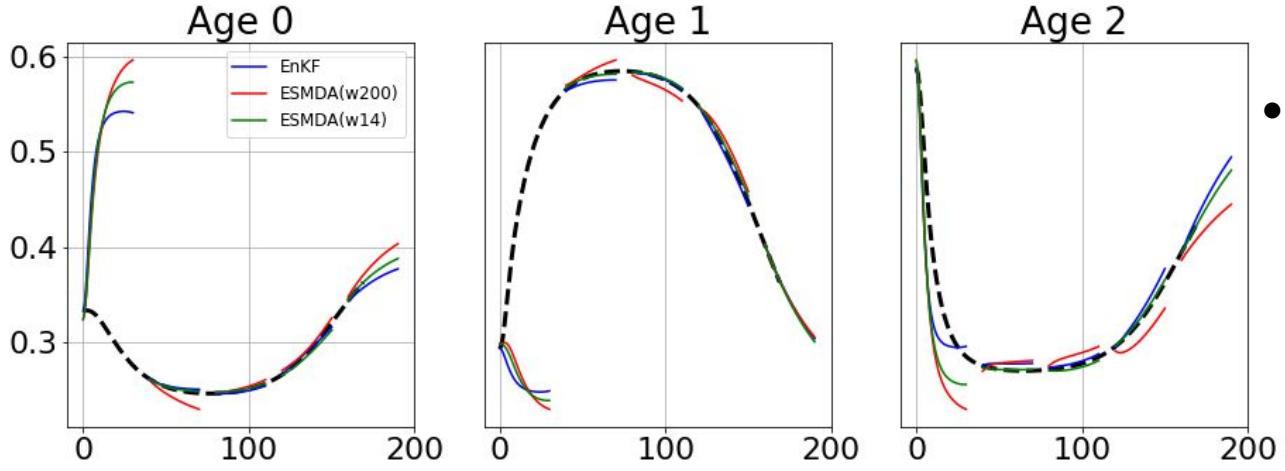




- ESMDA(w200) produces estimation with larger inter-experiment spread. For this long windows the problem becomes highly non-linear which may degrade the estimation of the parameters.
- Differences between EnKF and ESMDA(w14) are small in terms of the estimated parameters.

Parameter estimation experiments: Parameter estimation techniques



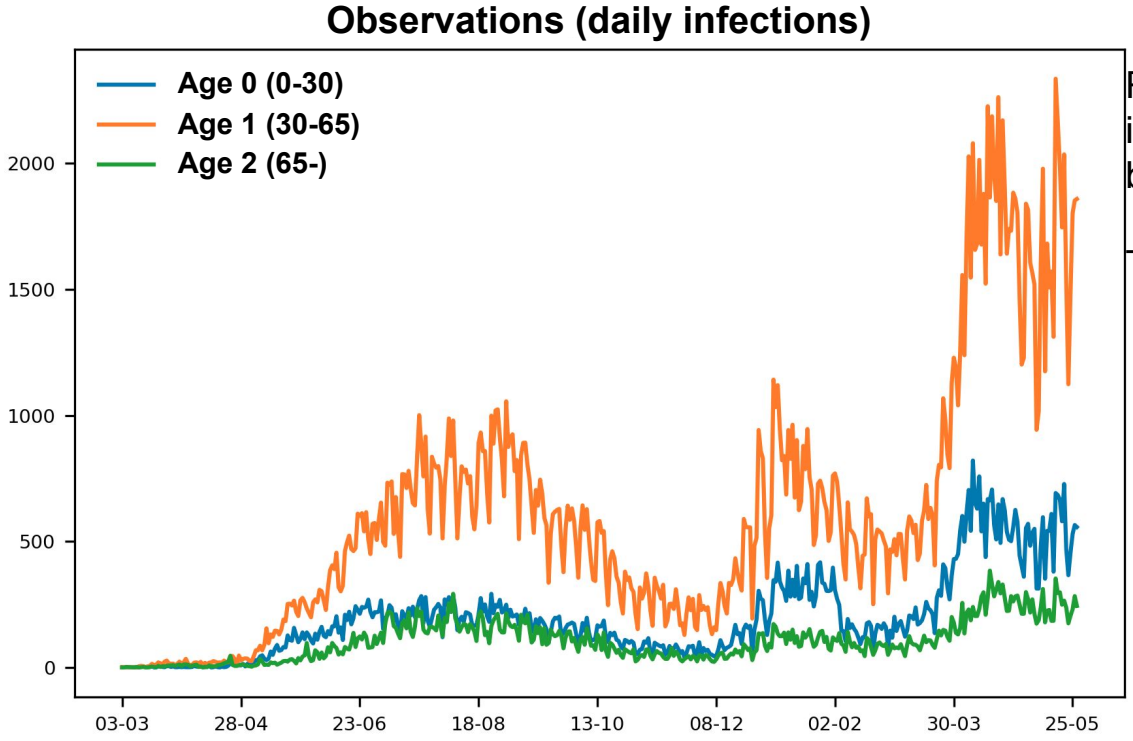


- The forecast error maximizes immediately before the peak.
- The ensemble spread is higher than the error, particularly for the EnKF.
- ESMDA(w200 and 14) performs better than the EnKF. This is probably due to the use of future observations.
- All the estimation schemes properly reproduce the evolution of the distribution of infection among the different groups (with a slight advantage of ESMDA with 14 days window).

Real case experiments

We performed experiments using data from the city of Buenos Aires in the period March 2019 - May 2021.

Real case experiment:

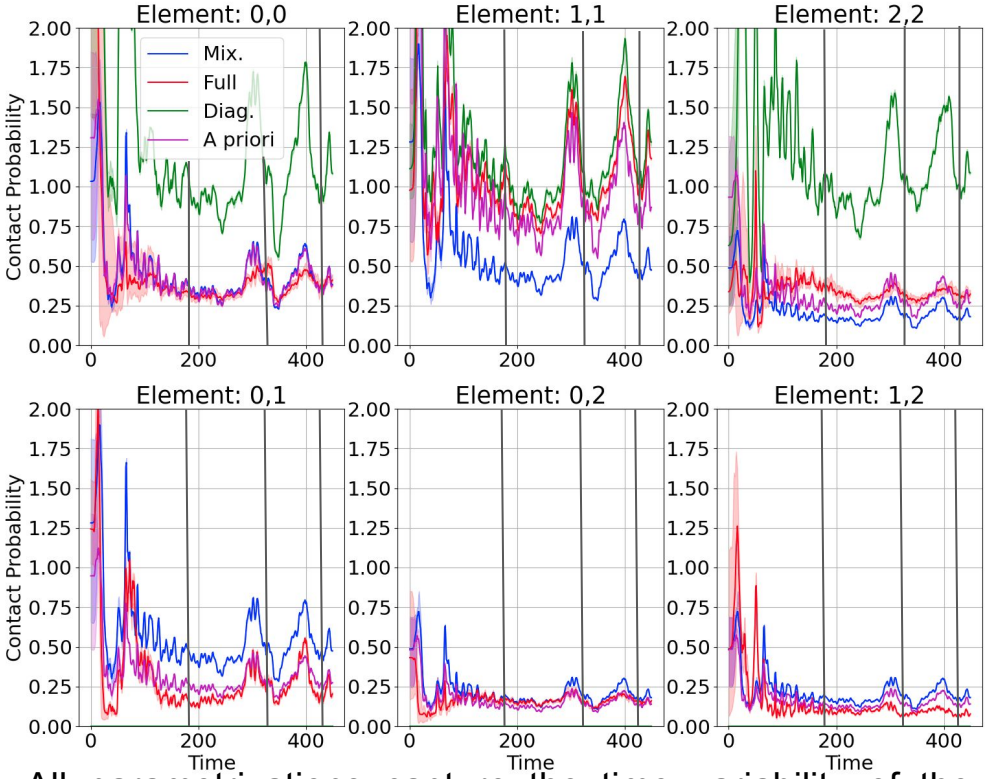


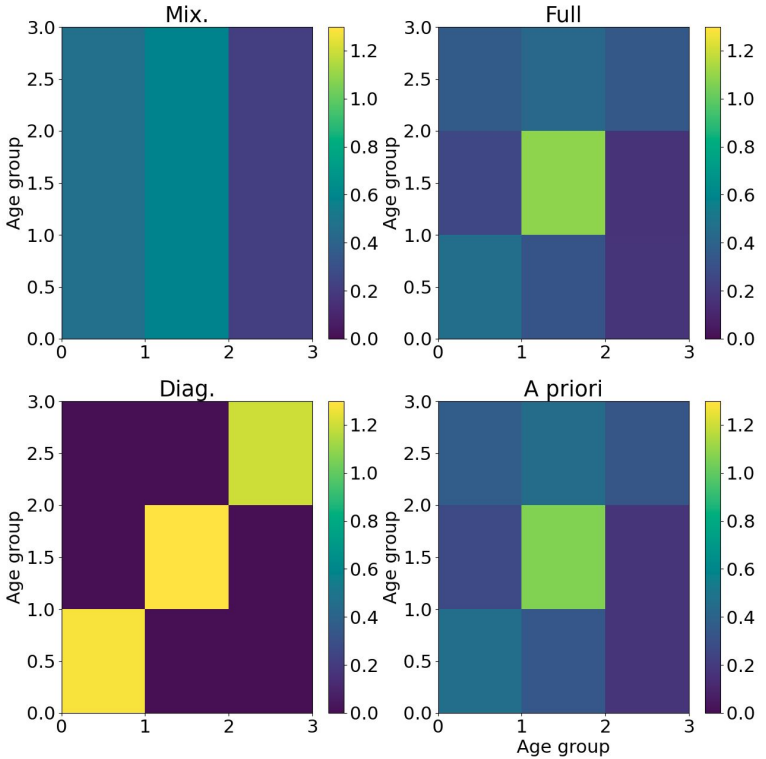
Real time observations of detected infections at the city of Buenos Aires between March 2020 to May 2021.

There are three main waves:

- The first and slowly developing wave during a long lock-down started on March 2020.
- A second wave during Christmas-New year holidays.
- A third rapid wave starting on March 2021 (still ongoing).

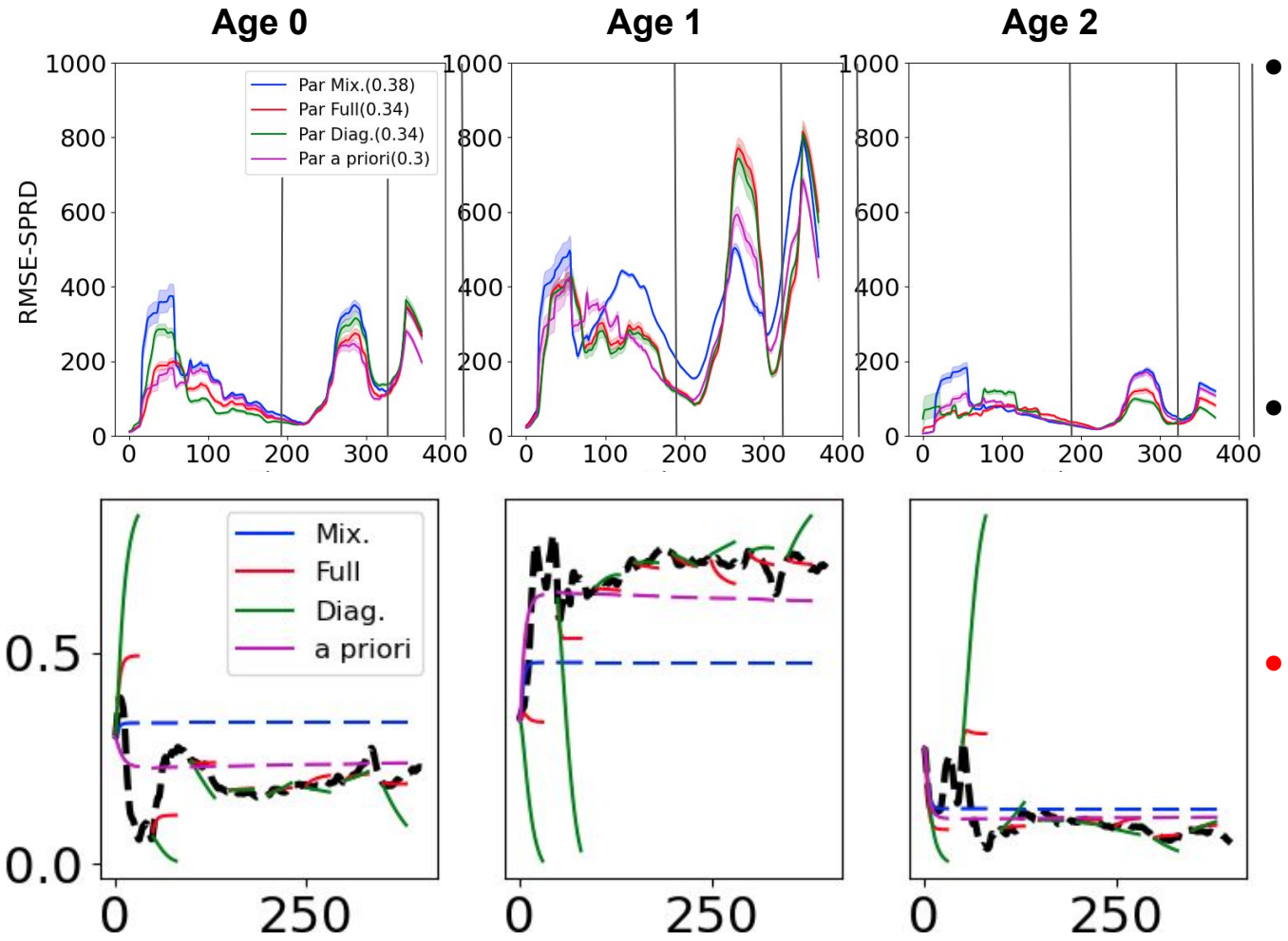
Parameter estimation experiments: Contact matrix parametrization





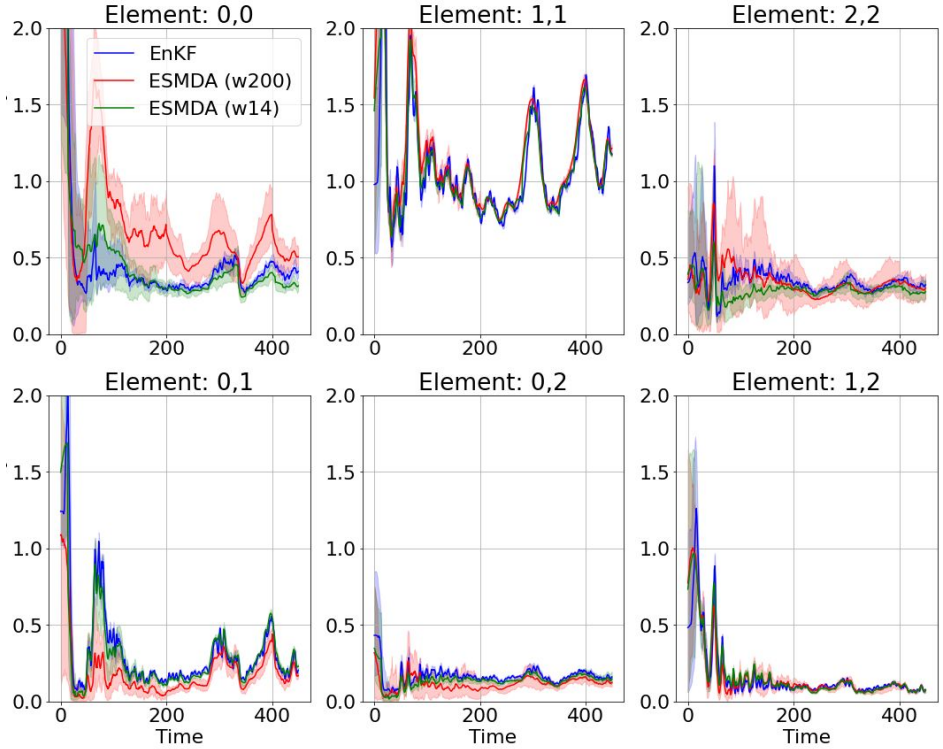
- All parametrizations capture the time variability of the parameters leading to the three distinct waves identified in the data.
- There is a good convergence to the parameters as indicated by the low spread among different realizations.
- The **Full** parametrization differs from the well-mixed scenario and identifies contacts in Age group 1 as the most significant source of growth (larger number of intra-group interactions).

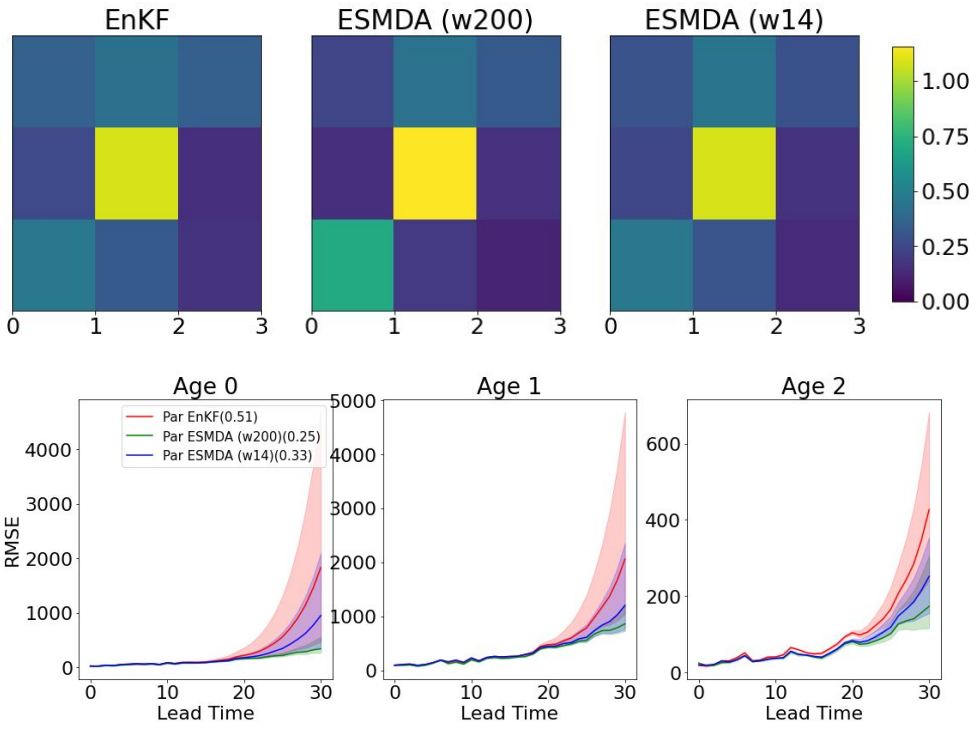
Parameter estimation experiments: Contact matrix parametrization



- Largest errors in the forecast are associated with local maxima. Particularly on the second and third outbreaks when a sharp change in the contact parameters is observed.
- All parametrizations of the contact matrix perform well in this scenario, with a small advantage of the "**a priori**" parametrization.
- From the perspective of the cases distributed among different groups, the **Full** and **Diag.** parametrizations provide the best results.

Parameter estimation experiments: Assimilation techniques





- Broadly speaking the estimated parameters by the different techniques are similar. Also the structure of the time average contact matrix is quite similar.
- The local maximum in the diagonal components of the contact matrix tends to occur earlier in the ESMDA (w200) and ESMDA (w14) which is probably a consequence of using “future observations”.
- ESMDA (w14) produces better forecasts than the EnKF for the three age groups.

Summary:

- We propose and evaluate four different parameterizations to represent age population inhomogeneity in a metapopulation SEIRD model.

- Parametrizations considering the inhomogeneity perform better than the "**well mixed**" assumption in both OSSE and real data experiments. The estimation of the **full** contact matrix produces better forecasts under OSSE experiments but shows no clear advantage in the real data experiment. In the real case experiment biases can result from the different detection rate associated to different age groups.

- We compare the EnKF and the ESMDA with two different window lengths. These two techniques perform well in the retrieval of the state and time-dependent model parameters.

- The ESMDA with a 14 day window outperforms the EnKF forecasts in both the OSSE and real data experiments. We believe this is mainly because ESMDA can handle the time lag between the parameter sensitivity and the observations. Also using the model as a strong constraint can be contributing to a better parameter estimation.

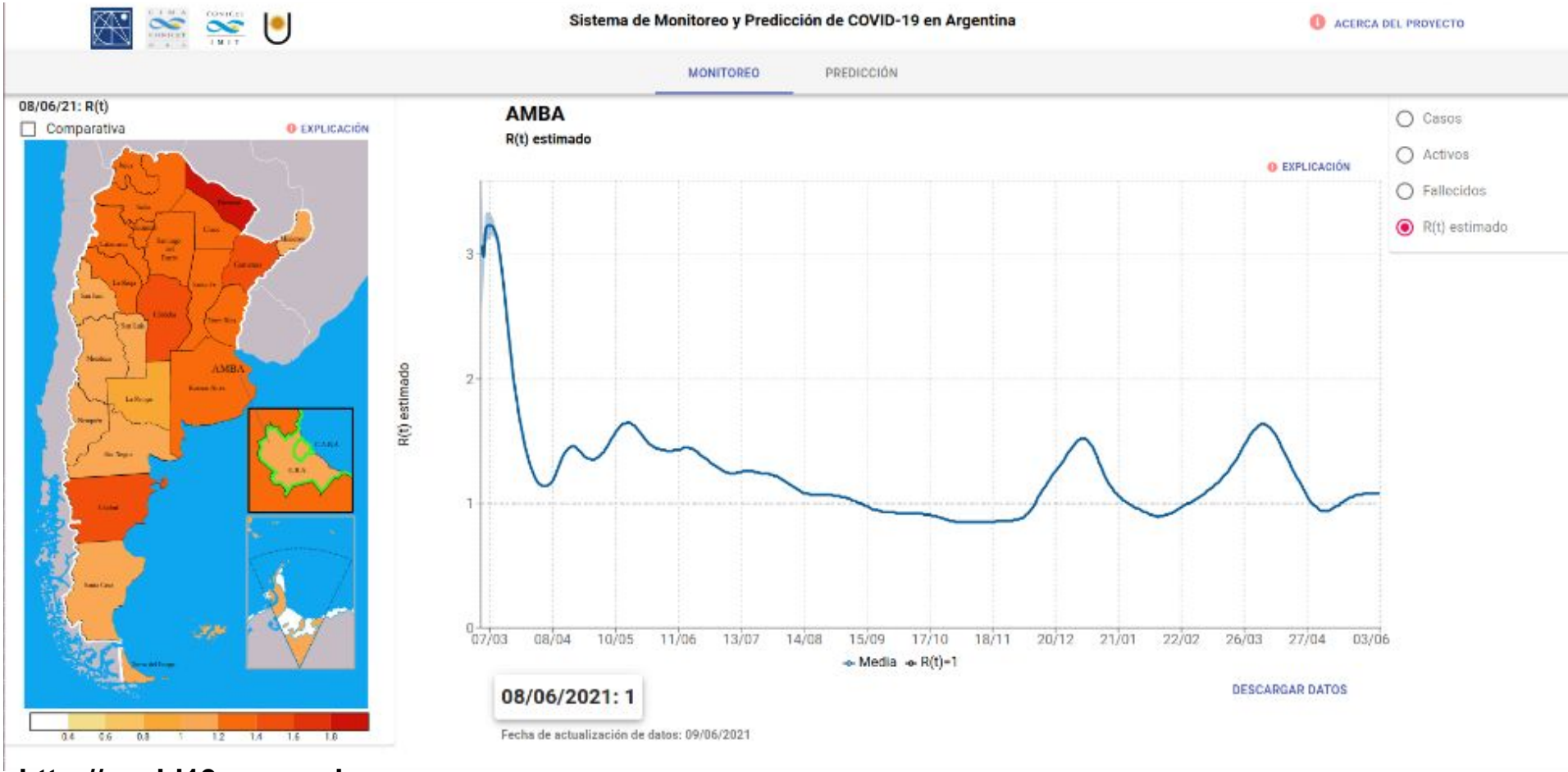
Possible future directions:

- Evaluate the estimation of contact matrices representing the interaction among different regions and study how the optimization of multi-region models helps to forecast the evolution of the outbreaks.

- Optimize the stochastic hyper-parameters of the model (e.g. characteristics of the random walk parameters) in order to better quantify the uncertainty.

- Include additional parameters of the coupling of epidemiological parameters with external forcings (e.g. weather conditions).

Operational COVID-19 monitoring using data assimilation



http://covid19.unne.edu.ar

Thank you

References

Evensen G., Amezcua J., Bocquet M., Carrassi A., Farchi A., Fowler A., Houtekamer P. L., Jones C. K., de Moraes R., Pulido M., Sampson C., Vossepoel F. 2021: An international initiative of predicting the SARS-CoV-2 pandemic using ensemble data assimilation. *Foundations of Data Science*, doi: 10.3934/fods.2021001

Hickmann KS, Fairchild G, Priedhorsky R, Generous N, Hyman JM, Deshpande A, et al. (2015) Forecasting the 2013–2014 Influenza Season Using Wikipedia. PLoS Comput Biol 11(5): e1004239. https://doi.org/10.1371/journal.pcbi.1004239

Ghostine, R.; Gharamti, M.; Hassrouny, S.; Hoteit, I.: 2021. "An Extended SEIR Model with Vaccination for Forecasting the COVID-19 Pandemic in Saudi Arabia Using an Ensemble Kalman Filter" *Mathematics* 9, no. 6: 636. https://doi.org/10.3390/math9060636

Li X., Zhao X., Liu F., Big data assimilation to improve the predictability of COVID-19, Geography and Sustainability, 1, 2020, 317-320, https://doi.org/10.1016/j.geosus.2020.11.005.

Pasetto D., Finger F., Rinaldo A. and Bertuzzo E., Real-time projections of cholera outbreaks through data assimilation and rainfall forecasting, *Adv. Water Res.*, 108 (2017), 345-356. doi: 10.1016/j.advwatres.2016.10.004.

Pei S., Kandula S., Yang W., Shaman J. 2018: Forecasting the spatial transmission of influenza in the United States. Proc Natl Acad Sci U S A;115(11):2752-2757. doi: 10.1073/pnas.1708856115. PMID: 29483256; PMCID: PMC5856508.

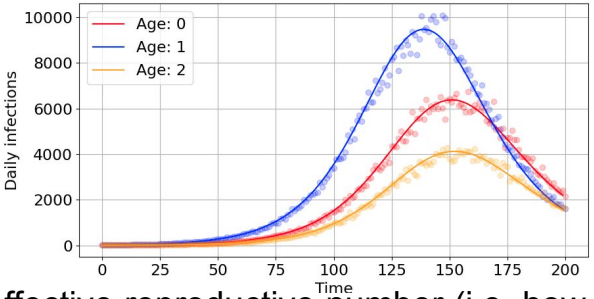
Shaman J., Karspeck A., Yang W., Tamerius J. and Lipsitch M. 2013: Real-time influenza forecasts during the 2012–2013 season, *Nature Commu.*, 4, 1-10. doi: 10.1038/ncomms3837.

The nature run - constant contact matrix experiment

We computed the **next generation matrix (NGM)** for the compartimental SEIRD model following Heffernan et al. 2005.

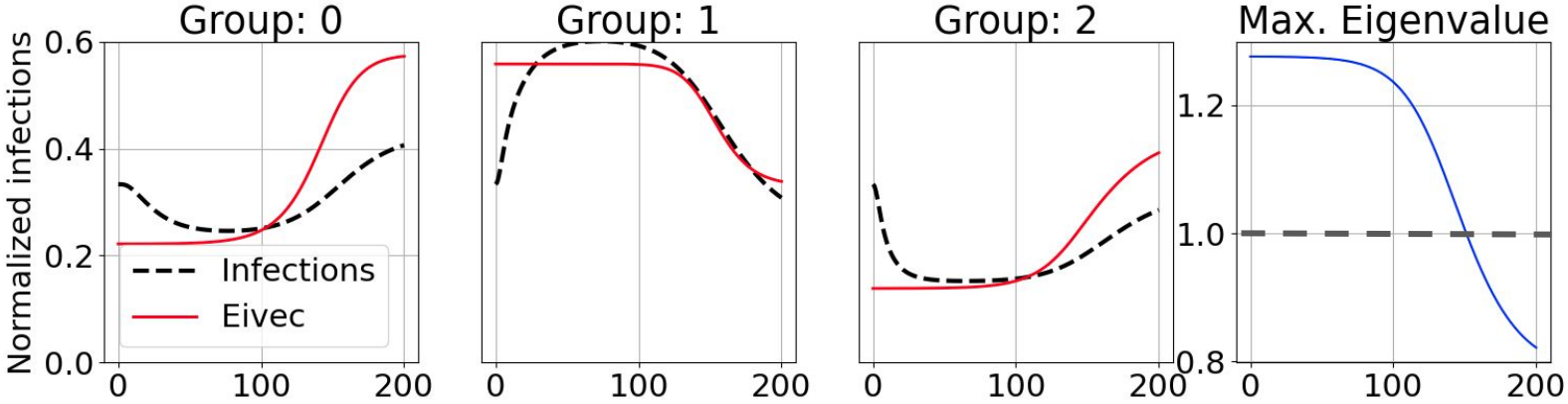
$$NGM = FV^{-1}$$

$$F = \begin{bmatrix} 0^{n_a \times n_a} & \frac{C(t))}{\tau_I} \circ (\vec{S} \otimes (1/\vec{N})) \\ \frac{I_d^{n_a \times n_a}}{\tau_E} & 0^{n_a \times n_a} \end{bmatrix} \qquad V = \begin{bmatrix} \frac{I_d^{n_a \times n_a}}{\tau_E} & 0^{n_a \times n_a} \\ 0^{n_a \times n_a} & \frac{I_d^{n_a \times n_a}}{\tau_I} \end{bmatrix}$$

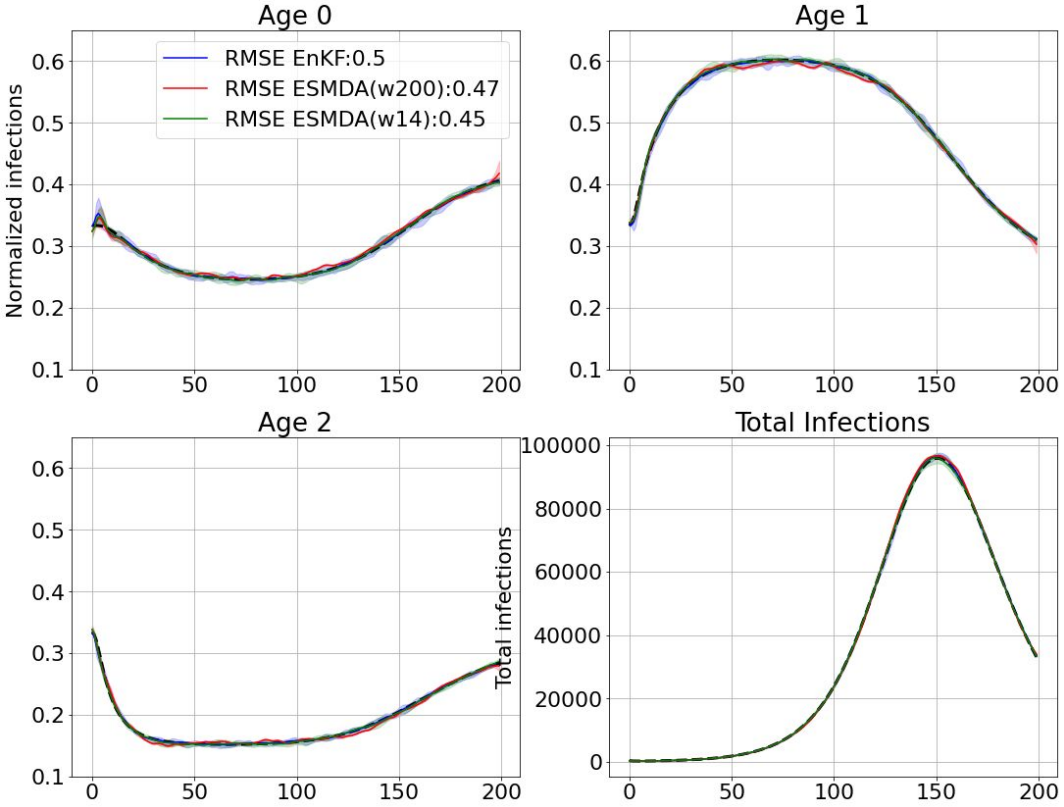


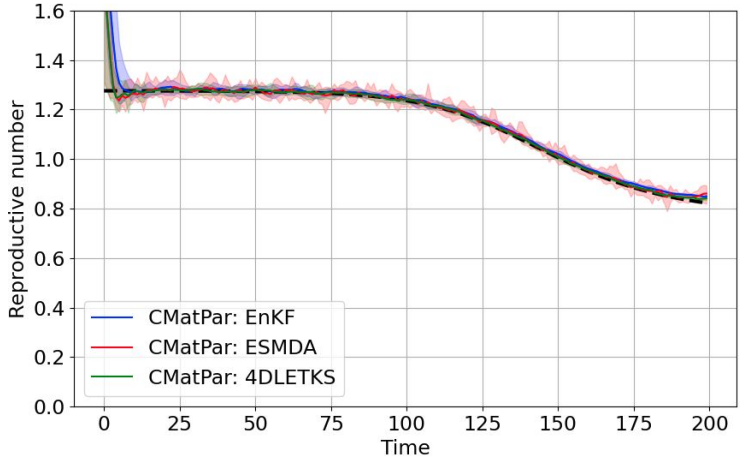
The maximum eigenvalue of the NGM can be associated with the time-dependent effective reproductive number (i.e. how many new infections starts from an infected person in a time period equal to τ_I). If this number is greater than one, then the number of infected people will grow exponentially.

The leading eigenvector is also an interesting property, showing how infections will be distributed among different groups.



Parameter estimation experiments: Parameter estimation techniques





- The three techniques provide a good representation of the number of active infections and its distribution among the different age groups, they also accurately capture changes in the reproductive number.
- **ESMDA(w14)** is slightly more accurate than **EnKF** and **ESMDA(w200)**.