A mechanistic and data-driven reconstruction of the time-varying reproduction number: Application to the COVID-19 epidemic

Bernard Cazelles[a,b,c], Clara Champagne[d,e], Benjamin Nguyen Van Yen[c,f], Catherine Comiskey[g], Elisabeta Vergu[b,☯], Benjamin Roche[h,☯]

a. Sorbonne Université, UMMISCO, Paris, France
b. INRAE, Université Paris-Saclay, MaIAGE, Jouy-en-Josas, France
c. Eco-Evolution Mathématique, IBENS, UMR 8197, CNRS, Ecole Normale Supérieure, Paris, France
d. Swiss Tropical and Public Health Institute, Bassel, Switzerland
e. Universty of Basel, Basel, Switzerland
f. Institut Pasteur, Unité de Génétique Fonctionnelle des Maladies Infectieuses, Paris, France
g. School of Nursing and Midwifery, Trinity College Dublin, The University of Dublin, Dublin, Ireland
h. MIVEGEC, IRD, CNRS and Université de Montpellier, Montpellier, France
☯ These authors contributed equally to this work.

Abstract

The effective reproduction number R_{eff} is a critical epidemiological parameter that characterizes the transmissibility of a pathogen. However, this parameter is difficult to estimate in the presence of silent transmission and/or significant temporal variation in case reporting. This variation can occur due to the lack of timely or appropriate testing, public health interventions and/or changes in human behavior during an epidemic. This is exactly the situation we are confronted with during this COVID-19 pandemic.
In this work, we propose to estimate R_{eff} for the SARS-CoV-2 (the etiological agent of the COVID-19), based on a model of its propagation considering a time-varying transmission rate. This rate is modeled by a Brownian diffusion process embedded in a stochastic model. The model is then fitted with Bayesian inference (PMCMC) using multiple well-documented hospital datasets from several regions in France and in Ireland.
This mechanistic modeling framework enables us to reconstruct the temporal evolution of the transmission rate of the COVID-19 based only on the available data, without any specific hypothesis on its evolution. This approach allows us to follow both the course of the COVID-19 epidemic and the temporal evolution of its $R_{eff}(t)$. In this way, we can appropriately assess the effect of the mitigation strategies implemented to control the epidemic waves in France and in Ireland.

Introduction

In the last months of 2019, clustered pneumonia cases were described in China [1]. The etiological agent of this new disease, a betacoronavirus, was identified in January and named SARS-CoV2. This new coronavirus disease (COVID-19) spread rapidly worldwide, causing millions of cases, killing hundreds of thousands of people and causing socio-economic damage. Until vaccination campaigns are widely implemented, the expansion of COVID-19 continues to threaten to overwhelm the

healthcare systems of many countries, despite a wide range of public health strategies using different non-pharmaceutical interventions (NPI).

In the early stages of each new epidemic, one of the first steps to design a control strategy is to estimate pathogen transmissibility in order to provide information on its potential to spread in the population. This is crucial to understand the likely trajectory of the new epidemic and the level of intervention that is needed to control it. Among the various indicators that quantify this transmissibility, the most commonly used is the reproduction number, which measures how many amore people on average can be infected by one infected individual. In the initial phase of the epidemic, when the entire population is susceptible, this quantity is referred to as R_0, the basic reproduction number, and is defined as the average number of secondary cases caused by one infected individual in an entirely susceptible population [2-3]. As the epidemic develops and the number of infected individuals increases (and the number of susceptible individuals decreases), the effective reproduction number R_{eff}, characterizing the transmission potential according to the immunological state of the host population, is used. It can be estimated as a function of time, the instantaneous effective reproduction number $R_{eff}(t)$ quantifying the number of secondary infections caused by an infected person at a specific time-point of an epidemic. The epidemic is able to spread when $R_{eff}(t)>1$ and is under control when $R_{eff}(t)<1$. $R_{eff}(t)$ can be used to monitor changes in transmission in near real time and, for instance, the effect of control or mitigation measures.

Classically, $R_{eff}(t)$ is estimated by the ratio of the number of new infections generated at time t to the total infectiousness of individuals in the infected state at time t. This latter quantity is defined based on the generation time distribution (the time between infection and transmission) or on the serial interval (the time between the onset of symptoms of a primary case and the onset of symptoms of secondary cases) [4-5].

For the COVID-19 epidemic, considering the data needed for the estimation of the reproductive numbers, caution is urged when interpreting the values obtained and the short-term fluctuations in these estimates due to both data quality and data uncertainty, which must be taken into account [6-9]. In a recent study by O'Driscoll et al [6], it was concluded, comparing different methods, that there are many important biases in the $R_{eff}(t)$ estimates and that this can easily lead to erroneous conclusions about changes in transmissibility during an epidemic. These biases are mainly due to the uncertainty in incidence data that can arise due to both the transmission characteristics of this virus (asymptomatic and pre-symptomatic transmission) and the quality and preparedness of the public health system. For COVID-19, it has been shown that the number of observed confirmed cases significantly underestimates the actual number of infections [10-11]. For instance, during the initial rapid growth phase of the COVID-19 epidemic, the number of confirmed case underestimated the actual number of infections by 50 to 100 times [10]. In France, it has been estimated that the detection rate increased from 7% in mid-May to 40% by the end of June, compared to well below 5% at the beginning of the epidemic [12]. In addition, these biases can be amplified by the combination of the high proportion of asymptomatic cases [13] and low health-seeking behaviors. Although the estimation of the reproductive number is robust to underreporting [14], this is only true if the reporting rate (among other characteristics) is constant over the observation period. This was not verified for the COVID-19 epidemic, mainly due to fluctuations in the capacity of testing and the

availability of information. The uncertainties in incidence data are not the only one, there is also uncertainty in the generation time and in the serial interval [15-16].

Other studies have highlighted these issues. Gostic et al [7] quantified the effects of data characteristics on $R_{eff}(t)$ estimates: reporting process; imperfect observation of cases; missing observations of recent infections; estimation of the generation interval. Moreover, Pitzer et al [8] showed that biases in $R_{eff}(t)$ are amplified when reporting delays have fluctuated due to the availability and changing practices of testing. These two studies concluded that changes in diagnostic testing and reporting processes must be monitored and taken into account when interpreting estimates of the reproductive number of COVID-19. Nevertheless these changes are extremely difficult to quantify. In these contexts, as testing capacity and reporting delays evolve, the use of hospital admission or death data may be preferable for inferring reproduction numbers [9,17]. However, the delays in the time from infection to hospitalization and/or death are also uncertain and, overall, it is difficult to incorporate the uncertainty related to all of these delays [9,17,18].

In addition to the classical methods, a complementary approach consists in inferring changes in transmission using mathematical models, and computing R_{eff} based on its proportionality with the transmission rates [19-24]. The time-variation of R_{eff} is computed indirectly by simply fitting the model to different time periods (before or after the lockdown in the simplest cases) or by using exponential decay models [21], but see Lemaitre et al [20].

To overcome all these numerous weaknesses in the data available for computing R_{eff}, we propose using a framework that has been already implemented to tackle non-stationarity in epidemiology [25-27]. This framework uses diffusion models driven by Brownian motion to model time-varying key epidemiological parameters embedded in a stochastic state-space framework coupled with Bayesian inference methods. The advantages of this approach are the possibilities of (i) considering the mechanisms of the pathogen transmission and then its particularities (asymptomatic), (ii) using multiple datasets (incidence and hospital data), (iii) accounting for all the uncertainty associated with the data used and especially (iv) following the evolution over time of some of the model parameters. The main advantage of this approach is that it is data-driven. Indeed, in this framework, the reconstruction of the time-varying parameters is done only under the weak assumption that they follow a basic stochastic process and it is estimated solely based on the available observations [27].

Applied to COVID-19, this framework makes it possible to monitor the evolution of disease transmission over time under non-stationary conditions such as those that prevailed during this epidemic. We have chosen to illustrate our approach with data from certain regions of France and Ireland.

Modeling a time-varying R_{eff}

Our framework is based on three main components: a stochastic epidemiological model embedded in a state-space framework, a diffusion process for each time-varying parameter and a Bayesian inference technique based on adaptive PMCMC (see *Supporting Information*). The main advantage

of the state-space framework is to explicitly consider the observation process. This makes it possible to take into account the unknowns and uncertainty in the partial observation of the disease. The proposed epidemiological model accounts for the transmission characteristic of the COVID-19 and the feature of the data. Figure 1 illustrates the model used, the compartments and flows between them. The temporal variation in the transmission rate $\beta(t)$ was modeled by making the assumption that it is not driven by mechanistic terms but evolves randomly. We consider that $\beta(t)$ follows a continuous diffusion process:

$$d\log(\beta(t)) = \nu.dB(t) \tag{1}$$

where ν is the volatility of the Brownian process (dB) to be estimated. Intuitively, the higher the values of ν the larger the changes in $\beta(t)$. The logarithmic transformation avoids negative values, which have no biological meaning.

Based on the SEIR model structure that accounts for the asymptomatic states (see eqs S2), the R_{eff} can be computed as:

$$R_{eff}(t) = \left(\frac{(1+q_1)}{2}.(1-\tau_A) + q_2.\tau_A\right).\frac{\beta(t)}{\gamma}.\frac{S(t)}{N} \tag{2}$$

where $1/\gamma$ is the infection duration, τ_A is the fraction of asymptomatic individuals in the population, $(1-\tau_A)$ the proportion of symptomatic infectious individuals and q_i the reduction in the transmissibility of some infected (I_2) and asymptomatics (A_i) (see Fig. 1 and eqs S2 in the *Supporting Information*).

For parameter estimation we used particle Markov chain Monte Carlo methods (Andrieu et al, 2010) suitable for partially observed stochastic non-linear systems (see *Supporting Information*). The implementation is done with the SSM software [26].

Our estimations of R_{eff} were compared, based on data from Ireland. with those obtained with two other methods, The first method, proposed by Cori et al [14] and frequently used in recent studies analyzing COVID-19 data, relies on the number of new infections generated during a given time period and the serial interval distribution and is implemented in the EpiEstim package. In the second method new infections and hence R_{eff} are generated using a simple discrete SIR model fitted with Kalman-filtering tools [28].

Data used

In the first wave of the COVID-19 epidemic the number of cases reported was very low and associated with large uncertainties [10,12]. This was due, on the one hand, to the testing capacity (RT-PCR laboratory capacity) that was limited and varied greatly during this epidemic and, on the other hand, to the features of this new virus, such as transmission before symptom onset and substantial asymptomatic transmission, which resulted in a low fraction of infected people attending health facilities for testing. This suggests that hospitalization data were likely to be the most accurate COVID-19 related data [9,17]. Thus, we focused on multiple hospital datasets in France and in Ireland. We used incidence data to avoid all the shortcomings associated with the use of cumulative data (see [29]), *i.e.*: daily hospital admissions, daily ICU admissions, daily hospital deaths and daily hospital discharges. We also used cases both in hospital and in ICU. Taking account of the large

variability in the daily observations, since the 1st of May for French data and the 1st of June for Irish data, we have used a weekly average of the observed daily values.
In France, since the beginning of the epidemic, the French regional health agencies (ARS – Agences Régionales de Santé https://www.ars.sante.fr/) have been reporting a number of aggregated statistics. We have also used data made available on the open platform for French public data: https://www.data.gouv.fr/fr/datasets/donnees-hospitalieres-relatives-a-lepidemie-de-covid-19/.
In Ireland, this data are published by the Health Protection Surveillance Centre, HPSC. We have used the daily reports (https://www.hpsc.ie/a-z/respiratory/coronavirus/novelcoronavirus/casesinireland/epidemiologyofcovid-19inireland/) and also the data hub launched last June (https://covid19ireland-geohive.hub.arcgis.com/).

Since these hospital multiple datasets were only available after the implementation of the first mitigation measures in France and in Ireland and that our aim was to estimate R_{eff} before, during and after the NPI measures, we also used reported incidence data for the first period.

Results

The temporal evolution of both the transmission rate $\beta(t)$ and $R_{eff}(t)$ (Fig. 2A) and the fit of the model to the observed data are displayed in Fig. 2. Fig. 3 shows the time evolution of the $R_{eff}(t)$ (Fig. 3A) and the dynamic of the model's unobserved compartments. Another important characteristic of this epidemic is the fact that the peak of daily hospital admission and daily ICU admission are concomitant (Figs 2G-2H and Figs S3,S6,S9,S12,S15,S18). This feature has been incorporated in the model by allowing hospitalization or admission to ICU for each stage of the symptomatic infection (Fig. 1 and eqs S2).

The capacity of our framework to describe the different types of data, among which some of them are characterized by large noise (within the Paris region for instance) is also illustrated in Fig. 2. The main advantage of this framework consists in its ability to reconstruct the time variation of the transmission rate $\beta(t)$ (Fig. 2A). Based on the estimation of $\beta(t)$, we can compute the time-variation of $R_{eff}(t)$ (Fig. 2A) and thus reconstruct the observed dynamics of the COVID-19 epidemic. Figure 3 provides the results for each component of the epidemic dynamic model for the Paris region. Our main result is described in Fig. 4 displaying the time evolution of the R_{eff} in the five French regions considered and in Ireland. Our estimates of the initial value of R_{eff} lie in the range [3.,3.5] in agreement with other estimates [11,30]. The peak of $R_{eff}(t)$ just before the start date of the first mitigation measures is presumably an effect of the model to accommodate diverging trends between reported case data and hospital data. Then one can note a decrease of about 80% in R_{eff} between the 1st of March and the 1st of May in all the regions considered (Fig. 4 and Table 1). The reduction in the transmission following the second lockdown was smaller, between 40% and 55% in France but less than 25% in Ireland. In both cases, given the temporality of the decline compared to the timing of the NPIs, these sharp decreases seem to be the result of the implementation of the mitigation measures.

We can also notice that there is a larger uncertainty at the end of the observation period. As the dynamics of the model are mainly driven by the hospitalization data, these latter determine the

transmission rate (and then R_{eff}) during the interval $[t-\delta t,t]$, with δt of the order of the average delay between infection and hospitalization, of around 2-3 weeks (Figs 4A-4B).
Fig. 4 also illustrates the robustness and but also the sensitivity of our method. Firstly, Fig. 4A shows that whether hospital discharge data are used or not during the inference process, similar medians and CI are found. Secondly, Figs 4A-4B highlight the sensitivity to asymptomatic transmission: when asymptomatic transmission increases (i.e. large q_A value) lower values are inferred for $\beta(t)$ hence leading to lower R_{eff} (the opposite happens when q_A decreases).

Table 1. Estimated R_{eff} reduction and seroprevalence for Ile-de-France region, Ireland, Provence Alpes Côte d'Azur region (PACA), Occitanie region (OC), Nouvelle-Aquitaine region (NA), Auvergne Rhône Alpes region (ARA).

Regions or Country	**R_{eff} reduction** between 01/03/2020 and 01/05/2020	**R_{eff} reduction** between peak of R_{eff} and 01/05/2020	**R_{eff} reduction** between 15/10/2020 and 27/11/2020	**Seroprevalence 15-05-2020*** ** Median, [95%CI]	**Seroprevalence 01-22-2021** Median, [95%CI]
Ile-de-France*	81.7.%	86.4%	40.3%	12.8%, [6.3%-24.1%]	25.3%, [12.7%-47.8%]
Ile-de-France**	82.7%	86.3%	42.1%	13.5%, [7.3%-26.4%]	26.4%, [14.3%-51.7%]
Ireland	79.2%	81.6%	23.9%	0.8%, [0.5%-1.6%]***	2.3%, [1.4%-5.1%]
PACA	76.3%	87.4%	46.8%	4.4%, [2.7%-9.2%]	19.7%, [12.1%-43.0%]
OC	83.2%	86.1%	50.1%	1.9%, [1.0%-4.3%]	8.6%, [4.6%-18.7%]
NA	77.7%	82.3%	47.0%	1.4%, [0.7%-3.6%]	6.5%, [3.2%-15.6%]
ARA	78.4%	81.6%	56.6%	2.9%, [2.1%-7.7%]	12.6% [9.1%-32.4%]]

*using hospital discharge data; **not using hospital discharge data; ***01-07-2020 for Ireland.

Figs 3C-3D show that the number of asymptomatic infectious is of the same order of magnitude as the number of symptomatic infectious but with a larger uncertainty due to lack of information in the data. Indeed, the data used contain very little information on asymptomatics and we observe identical prior and posterior distributions for the rate of asymptomatic, τ_A (Figs S1,S5,S8). It can also be noted that the number of removed individuals is small and that the seroprevalence calculated as of 15 May for France or the 1 July for Ireland was very low but in full agreement with the published results of the seroprevalence surveys (Tables 1,S3). These estimated seroprevalence show that herd immunity is far from being reached even after the second waves.

The comparison of the performances of our estimations of R_{eff} with those of two others methods is summarized in Fig. 5. It is difficult to compare the absolute values of the estimates obtained with the 3 methods because the true values are unknown and speculative, but we can limit ourselves to comparing the trends given by these methods. The trajectory of R_{eff} over time computed by other

methods fall within the range of our 95% CI, which is large because of uncertainties in the transmission rate, asymptomatic transmission and the different delays needed to describe hospital data. We can also note that the 95% CI of R_{eff} obtained with the EpiEstim method is very narrow and its width is smaller than the variability of fluctuations in its median. The main differences between the three estimates relate to the time-lags of the effects of the lockdown, the peaks of R_{eff} and the date of crossing the 1 threshold. These lags range from one week to more than one month and do not correspond to the early or late lags of the date of crossing 1, when comparing to the estimations provided by our method.

A final important point concerns the observed incidence. We have used the incidence data until the 22 March (Fig. 2B, black points; Figs S6-S9-S12-S15-S18) in the inference process leading to a median of the posterior distribution of the reporting rate of 2.7% (95% CI: 1.9%-4.1%) for the Paris region (Table S2). The plotted dynamics of the estimated observed incidence uses this value for the whole trajectory. The values of the reporting rate for the other regions are in Table S2. The comparison of the observed incidence when available and the simulated incidence (Figs 2,S3,S6,S9,S12,S15,S18) clearly illustrates that the reporting rate has greatly evolved during the course of the epidemic. It is relatively easy to see that during the first wave the observed peak of incidence comes after the peak of hospitalization, whereas for the second wave the opposite happens, in agreement with what can be expected. Moreover, if we compare the intensity of the observed incidence waves, we can see that the second observed wave is 5 to 10 times greater than the first one, whereas model-based simulations suggest that they have similar magnitudes. This difference cannot only be explained by the fact that, as the frequency of testing increases, it is also increasingly likely that some of people tested positive are asymptomatic, whereas in the model the people tested are considered symptomatic. In any case, it appears crucial to be able to take into account a reporting delay that seems relatively large for the COVID-19 epidemic using models for now-casting [31-32] or a detailed observation model [18].

Discussion

For each epidemic, it is always essential to estimate pathogen transmissibility in order to provide information on its potential diffusion in the population. This is commonly done using the effective reproduction number $R_{eff}(t)$. Its time evolution allows us to follow, almost in real time, the evolution of the course of the epidemic, and get information regarding the potential effects of mitigation measures taken.

Here we propose a mechanistic approach based on time-varying parameters embedded in a stochastic model coupled with Bayesian inference. This enables us to reconstruct the temporal evolution of the transmission rate of COVID-19, on the basis of the only data available without specific hypothesis on this evolution. With this approach we describe both the temporal evolution of the COVID-19 epidemic and its $R_{eff}(t)$. Thus, we can quantify the effect of mitigation measures on the epidemics waves in France and in Ireland. It is important to note that this methodology overcomes many of the biases associated with estimates of $R_{eff}(t)$ with more traditional methods in the case of COVD-19 [6-7]. Indeed, there are many biases related to under-reporting of infectious cases, uncertainties in the

generation time and in the serial interval, and also, to the importance of silent transmission. Moreover these biases are amplified by the fact that reporting delays have fluctuated widely over the course of the epidemic [8].

Within our framework, the estimates of $R_{eff}(t)$ represent the values required to reproduce the well-documented hospital data. Such hospital data are clearly of a higher quality than the reported number of cases [17-18]. Moreover, this estimation accounts for transmission mechanisms and the different delays describe the process of transmission (even asymptomatic transmission) and the different processes evolving in the hospital (Fig. 1). Using our approach, by visualizing the evolution of $R_{eff}(t)$ we can follow the course of the COVID-19 epidemic in five French regions and in Ireland. Therefore, we can quantify the effect of the mitigation measures during and between epidemics waves. For the first lockdown we estimated a decrease of around 80% in transmission while between 45 to 55% in French regions and less than 25% in Ireland (Table 1) during the second lockdown, which was less restrictive. This may reflect the nature of the mitigation measures introduced in both countries for the second wave. In Ireland, the mitigation measures introduced in the second wave were less restrictive than the first, and this may explain the smaller reduction in the effective reproductive number. We also found other interesting results such as a significant high correlation between the trend of mobility and transmission between the epidemic waves (see Fig. S21 and [32]), highlighting the importance of following the evolution of mobility when relaxing mitigation measures to anticipate the future evolution of the spread of the SARS-CoV-2.

We have also compared the trend of our estimations of R_{eff} with those of two other methods from the literature [14,28] on data from Ireland (Fig. 5). The main difference between our estimations and the two others consists in asynchronous peaks of R_{eff}. For instance, the peak of R_{eff} occurs at the beginning of August for the two other methods, while our estimates suggest a peak in mid-September (Fig. 5). These differences may be explained by the fact that the two other methods computed their estimations based on the new cases only.
The peak of R_{eff} in early August could be explained by the increase in testing at during the vacation period while our estimates peaked later due to the increase in hospitalization generated by higher values of new infections in early September when the economy restarted. We therefore believe that our estimates based on admissions in hospital and ICU and deaths were more consistent with the peak of the second waves of hospitalization that arrived in late October in Ireland (Figs S6-S7). This convinces us of the value of our framework in presenting a more coherent picture of the course of this epidemic.

Clearly the weakness of our approach is the dependence of the $R_{eff}(t)$ estimations on the relevance of the model used and the accuracy and completeness of the available data. In our case, inference was based on hospital data that are clearly of a higher quality and accuracy compared to the observed number of infected cases. Moreover, despite its relative simplicity, the model incorporating time varying transmission rate is able to accurately describe the multiple hospital datasets that included daily hospitalized admission for COVID-19, daily ICU admission, daily deaths at hospital and daily hospital discharged and also the number of beds used each day both in hospital and ICU. Furthermore our model can be partially validated by the fact that our predicted seroprevalence in the French regions and in Ireland are in complete agreement with the results published from seroprevalence

surveys in these settings (see Tables 1,S3). To highlight our model results, we can see that the asymptomatic infectious are as numerous as symptomatic ones, but are characterized by a larger uncertainty, due to the lack of information in the data (Figs 3C-3D and figures in the *Supporting Information*). This is in agreement with recent papers [11,24,34,35], which emphasize that the growth of the COVID-19 epidemic is driven by silent infections. This has also been highlighted in Ireland where it has been estimated that in the second epidemic wave the ratio of silent infections to known reported cases was approximately 1:1 [36]. It is indeed interesting to note that our model estimates for asymptomatic cases in Ireland lead to similar ratios but with a large 95% CI (see Fig. S7).

Our study is not without limitations. The model used here is, like all complex SEIR models developed for COVID-19, non-identifiable which means that it is likely that several solutions exist and we only present one of the most likely ones. This point is overlooked very open but see [37]. One limitation is the use of the classical homogeneous mixing assumption in which all individuals are assumed to interact uniformly and ignores heterogeneity between groups by sex, age, geographical region. However this kind of data are not easily available and where mixing patterns among age groups are available at the individual level in contact tracing databases they are only accessible following extensive ethical reviews. Another weakness is to not consider an age-structure in the model in order to simulate age-based predictions. In all cases taking an age structure and mixing matrix appears insufficient and heterogeneity of contact is important (see Britton et al, 2020). Nevertheless, in our opinion, these limitations are more than balanced by the fact that we take into account the non-stationarity of the epidemic data and that our results are mainly driven by hospital data, which is more accurate and timely than the number of infected cases. As our main objective was to infer global R_{eff}, and not to explore age-specific mitigation strategies, simplification of the age structure appears justified. The corroboration of our Irish findings on proportions asymptomatic with those of others provides further evidence of this [36].

As demonstrated previously, modeling with time varying parameter is an interesting framework for modeling the time evolution of an epidemic even if the knowledge about disease transmission is either incomplete or uncertain [27]. Indeed a large part of all the unknowns is put in the time-varying parameters described by a diffusion process but driven by the observed data. This is exactly what we have been confronted during the COVID-19 pandemic, as the data are uncertain, as are the transmission mechanisms of SARS-CoV-2. We therefore proposed to model the spread of this disease using a stochastic model with a time-varying transmission rate inferred using well-documented hospital multiple datasets. The knowledge of the transmission rate makes it possible to easily calculate the $R_{eff}(t)$, which is a key parameter of the epidemic, in order to monitor the effects of public health policies on the course the COVID-19 epidemic. We believe, therefore, that this framework could be particularly useful for analyzing this new phase of worrying expansion and for helping to refine possible future mitigation measures after the second wave, pending the imminent launch of the vaccines.

Authors' contributions

BC, EV and BR contributed to conception and design of the study. BC, ClC and BNVY constructed and ran the model. All authors analyzed the simulations. The first draft of the manuscript was written by BC and CaC, all authors provided comments, edited and approved the final manuscript.

Acknowledgments

We thank Una Ni Mhaoldhomhnaigh who extracted and managed the data during the lockdown periods and edited some of the previous version of the manuscript.

Funding

B. C. and B. R. are partially supported by a grant ANR Flash Covid-19 from the "Agence Nationale de la Recherche" (DigEpi).

References

1. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, Ren R, Leung KSM, Lau EHY, Wong JY, Xing X, Xiang N, Wu Y, Li C, Chen Q, Li D, Liu T, Zhao J, Liu M, Tu W, Chen C, Jin L, Yang R, Wang Q, Zhou S, Wang R, Liu H, Luo Y, Liu Y, Shao G, Li H, Tao Z, Yang Y, Deng Z, Liu B, Ma Z, Zhang Y, Shi G, Lam TTY, Wu JT, Gao GF, Cowling BJ, Yang B, Leung GM, Feng Z. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. N Engl J Med. 2020;382:1199-1207.

2. Heesterbeek, J. A. P., & Dietz, K. (1996). The concept of Ro in epidemic theory. Statistica Neerlandica. 1996:50;89-110.

3. Diekmann O, Heesterbeek JA, Roberts MG. The construction of next-generation matrices for compartmental epidemic models. J R Soc Interface. 2010;7:873-885.

4. Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. Proc R Soc B Biol Sci. 2007; 274(1609):599–604

5. Lehtinen S, Ashcroft P, Bonhoeffer S. On the relationship between serial interval, infectiousness profile and generation time. J R Soc Interface. 2021;18:20200756.

6. O'Driscoll M, Harry C, Donnelly CA, Cori A, Dorigatti I, A comparative analysis of statistical methods to estimate the reproduction number in emerging epidemics with implications for the current COVID-19 pandemic, Clinical Infectious Diseases. 2020; ciaa1599.

7. Gostic KM, McGough L, Baskerville EB, Abbott S, Joshi K, Tedijanto C, et al. Practical considerations for measuring the effective reproductive number, Rt. PLoS Comput Biol. 2020;16:e1008409.

8. Pitzer, V. E., Chitwood, M., Havumaki, J., Menzies, N. A., Perniciaro, S., Warren, J. L., ... & Cohen, T. (2020). The impact of changes in diagnostic testing practices on estimates of COVID-19 transmission in the United States. medRxiv. 2020.04.20.20073338.

9. Sherratt, K., Abbott, S., Meakin, S., Hellewell, J., Munday, J. D., Bosse, N., ... & CMMID Covid-19 working group. (2020). Evaluating the use of the reproduction number as an epidemiological tool, using spatio-temporal trends of the Covid-19 outbreak in England. medRxiv. 2020.10.18.20214585;

10. Richterich, P. Severe underestimation of COVID-19 case numbers: effect of epidemic growth rate and test restrictions. MedRxiv. 2020; 2020.04.13.20064220.

11. Subramanian, R., He, Q., & Pascual, M. (2020). Quantifying Asymptomatic Infection and Transmission of COVID-19 in New York City using Observed Cases, Serology and Testing Capacity. MedRxiv:2020; 2020.10.16.20214049.

12. Pullano, G., Di Domenico, L., Sabbatini, C. E., Valdano, E., Turbelin, C., Debin, M., ... & Colizza, V.(2020). Underdetection of COVID-19 cases in France in the exit phase following lockdown. medRxiv 10.1101/2020.08.10.20171744.

13. Oran DP, Topol EJ. Prevalence of Asymptomatic SARS-CoV-2 Infection: A Narrative Review. Ann Intern Med. 2020;173:362-367.

14. Cori A, Ferguson NM, Fraser C, Cauchemez S. A new framework and software to estimate time-varying reproduction numbers during epidemics. Am J Epidemiol. 2013;178:1505-1512.

15. Ali, S. T., Wang, L., Lau, E. H., Xu, X. K., Du, Z., Wu, Y., ... & Cowling, B. J. (2020). Serial interval of SARS-CoV-2 was shortened over time by nonpharmaceutical interventions. Science. 2020;369:1106-1109.

16. Ganyani, T., Kremer, C., Chen, D., Torneri, A., Faes, C., Wallinga, J., & Hens, N. Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, 2020. Eurosurveillance 2020;25:2000257.

17. Abbott, S., Hellewell, J., Thompson, R. N., Sherratt, K., Gibbs, H. P., Bosse, N. I., ... & Chan, Y. W. D. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. Wellcome Open Res. 2020;5:112.

18. Wilder, B., Mina, M. J., & Tambe, M. (2020). Tracking disease outbreaks from sparse data with Bayesian inference. arXiv preprint arXiv:2009.05863.

19. Hao X, Cheng S, Wu D, Wu T, Lin X, Wang C. Reconstruction of the full transmission dynamics of COVID-19 in Wuhan. Nature. 2020;584:420-424.

20. Lemaitre, J. C., Perez-Saez, J., Azman, A., Rinaldo, A., & Fellay, J. (2020). Assessing the impact of non-pharmaceutical interventions on SARS-CoV-2 transmission in Switzerland. Swiss Med Wkly. 2020;150: w20295.

21. Dehning J, Zierenberg J, Spitzner FP, Wibral M, Neto JP, Wilczek M, Priesemann V. Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. Science. 2020;369:eabb9789.

22. Di Domenico L, Pullano G, Sabbatini CE, Boëlle PY, Colizza V. Impact of lockdown on COVID-19 epidemic in Île-de-France and possible exit strategies. BMC Medicine. 2020;18:240.

23. Prague M, Wittkop L, Clairon Q, Dutartre D, Thiébaut R, Hejblum BP. Population modeling of early COVID-19 epidemic dynamics in French regions and estimation of the lockdown impact on infection rate. MedRxiv. 2020;2020.04.21.20073536.

24. Aguiar M, Van-Dierdonck JB, Stollenwerk N. Reproduction ratio and growth rates: Measures for an unfolding pandemic. PLoS One. 2020;15:e0236620.

25. Cazelles B, Chau, NP. Using the Kalman filter and dynamic models to assess the changing HIV/AIDS epidemic. Mathematical Biosciences. 1997;140:131-154.

26. Dureau J, Kalogeropoulos K, Baguelin M. Capturing the time-varying drivers of an epidemic using stochastic dynamical systems. Biostatistics. 2013;14:541-555.

27. Cazelles B, Champagne C, Dureau J. Accounting for non-stationarity in epidemiology by embedding time-varying parameters in stochastic models. PLoS Computational Biology. 2018;14:e1006211.

28. Marioli, F. A., Bullano, F., Kučinskas, S., & Rondón-Moreno, C. (2020). Tracking R of COVID-19: A New Real-Time Estimation Using the Kalman Filter. MedRxiv. 2020.04.19.20071886.

29. King AA, Domenech de Cellès M, Magpantay FM, Rohani P. Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. Proceedings of the Royal Society B: Biological Sciences. 2015;282:20150347.

30. Flaxman S, Mishra S, Gandy A, Unwin HJT, Mellan TA. Coupland H, et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. Nature. 2020;584:257–61.

31. Höhle M., an der Heiden M. Bayesian nowcasting during the STEC O104: H4 outbreak in Germany, 2011. Biometrics. 2014;70:993-1002.

32. Bird S, Nielsen B. Now-casting of COVID-19 deaths in English hospitals. 2020; Nuffield College; (preprint) (available from: http://users.ox.ac.uk/~nuff0078/Covid/)

33. Cazelles, B., Comiskey, C., Nguyen Van Yen, B., Champagne, C., Roche, B. (2021). Identical trends of SARS-Cov-2 transmission and retail and transit mobility during non-lockdown periods. medRxiv. 2020.12.29.20248990.

- Britton T, Ball F, Trapman P. A mathematical model reveals the influence of population heterogeneity on herd immunity to SARS-CoV-2. Science. 2020;369:846-849.

34. Moghadas, S. M., Fitzpatrick, M. C., Sah, P., Pandey, A., Shoukat, A., Singer, B. H., & Galvani, A. P. (2020). The implications of silent transmission for the control of COVID-19 outbreaks. Proceedings of the National Academy of Sciences, 117, 17513-17515.

35. Fox, S. J., Pasco, R., Tec, M., Du, Z., Lachmann, M., Scott, J., & Meyers, L. A. (2020). The impact of asymptomatic COVID-19 infections on future pandemic waves. MedRxiv:2020; 2020.06.22.20137489.

36. Comiskey C, Snel A, Banka S. The second wave: Estimating the hidden asymptomatic prevalence of Covid-19 in Ireland as we plan for imminent immunisation. HRB Open Research under review. 2021.

37. Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, Shaman J. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science. 2020;368:489-493.

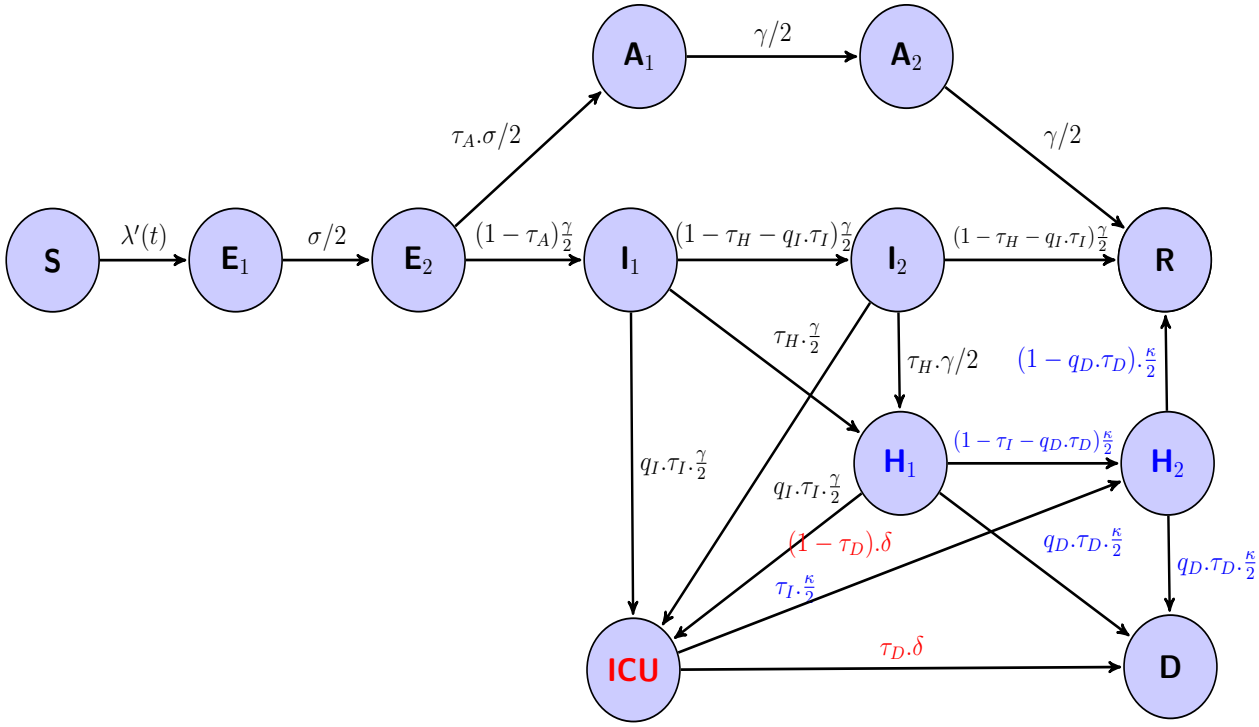

FIGURE 1 – Flow diagram of the model, a generalized SEIR model accounting for asymptomatic transmission and a simple hospital system (see eqs A2). The variables are : the susceptibles S, the infected non-infectious E, the infectious symptomatic I, the infectious symptomatic A, the removed people R, and the hospital variables : hospitalized individuals H, individuals in intensive care unit ICU, hospital discharge G, and deaths at hospital D. The subscripts 1 and 2 are for the 2 stages of the Erlang distribution of the stage durations. The hospital discharge is the flow from H_2 to R. Flows in blue are from hospital (H_i) and flow in red from ICU.

$\lambda'(t) = \beta(t).(I_1 + q_1.I_2 + q_2.(A_1 + A_2))/N$ then the force of infection is $\lambda(t) = \lambda'(t).S(t)$, $\beta(t)$ is the time-varying transmission rate (1), σ the incubation rate, γ the recovery rate, $1/\kappa$ average hospitalized period, $1/\delta$ average time spent in ICU, τ_A the fraction of asymptomatics, τ_H the fraction of infectious hospitalized, τ_I the fraction of ICU admission, τ_D death rate, q_1 and q_2 reduction of transmissibility of I_2 and A_i, q_I a reduction of the fraction of people admitted in ICU and q_D reduction of the death rate.

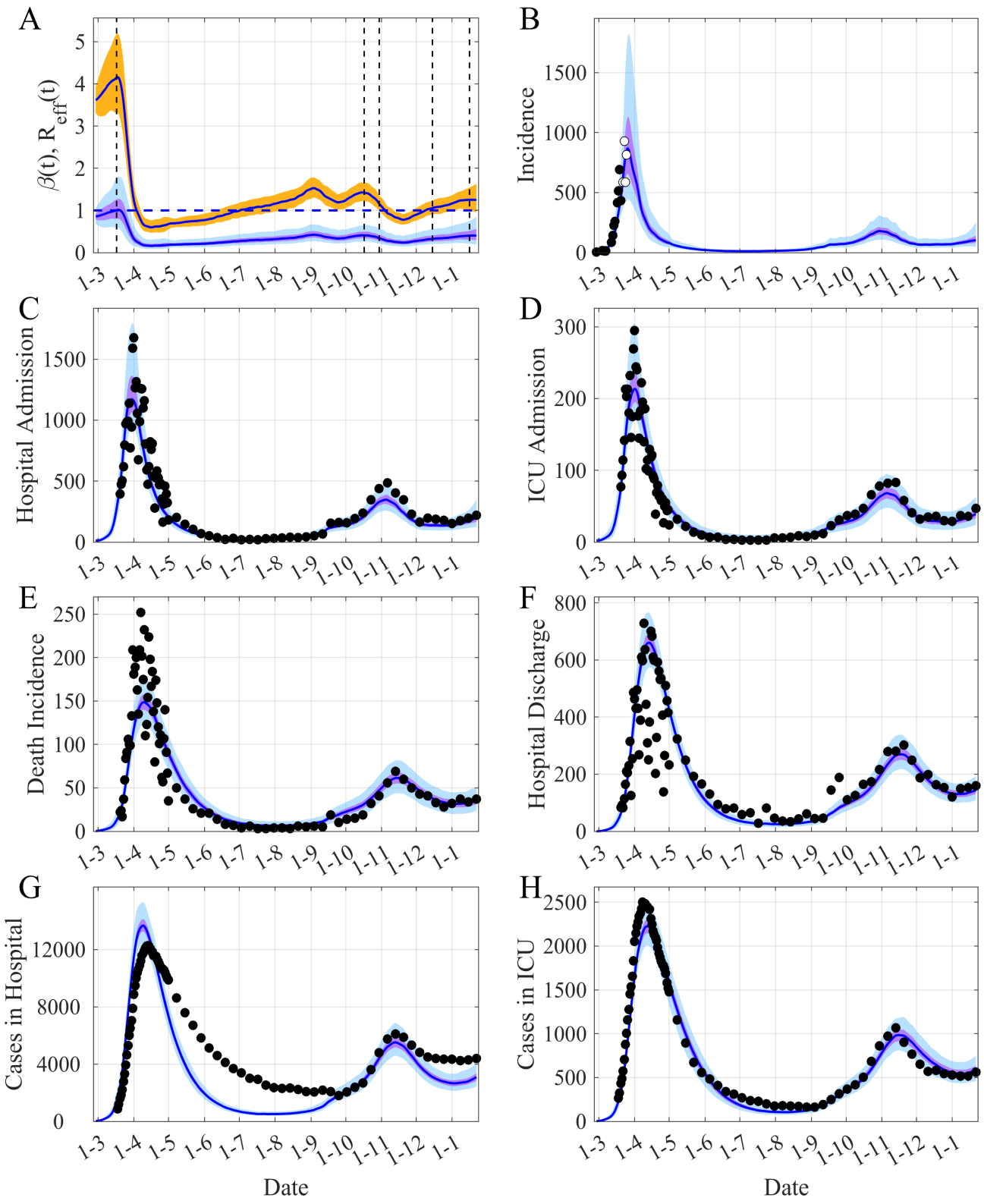

FIGURE 2 – Flow diagram of generalized SEIR model accounting for asymptomatic transmission and a simple hospital system (see eqs A2). The variables are : the susceptibles S, the infected non-infectious E, the infectious symptomatic I, the infectious asymptomatic A, the removed people R, and the hospital variables : the hospitalized individuals H, the individuals in intensive care unit ICU, and the deaths at hospital D. The subscripts 1 and 2 stand for the 2 stages of the Erlang distribution of the sojourn times in E, A, I, H. The hospital discharge is the flow from H_2 to R. Flows in blue are from hospital (H_i) and flow in red from ICU. $\lambda'(t) = \beta(t).(I_1 + q_1.I_2 + q_2.(A_1 + A_2))/N$ then the force of infection is $\lambda(t) = \lambda'(t).S(t)$, $\beta(t)$ is the time-varying transmission rate (1), σ the incubation rate, γ the recovery rate, $1/\kappa$ average hospitalized period, $1/\delta$ average time spent in ICU, τ_A the fraction of asymptomatics, τ_H the fraction of infectious hospitalized, τ_I the fraction of ICU admission, τ_D death rate, q_1 and q_2 the reduction in the transmissibility of I_2 and A_i, q_I the reduction in the fraction of people admitted in ICU and q_D the reduction in the death rate.

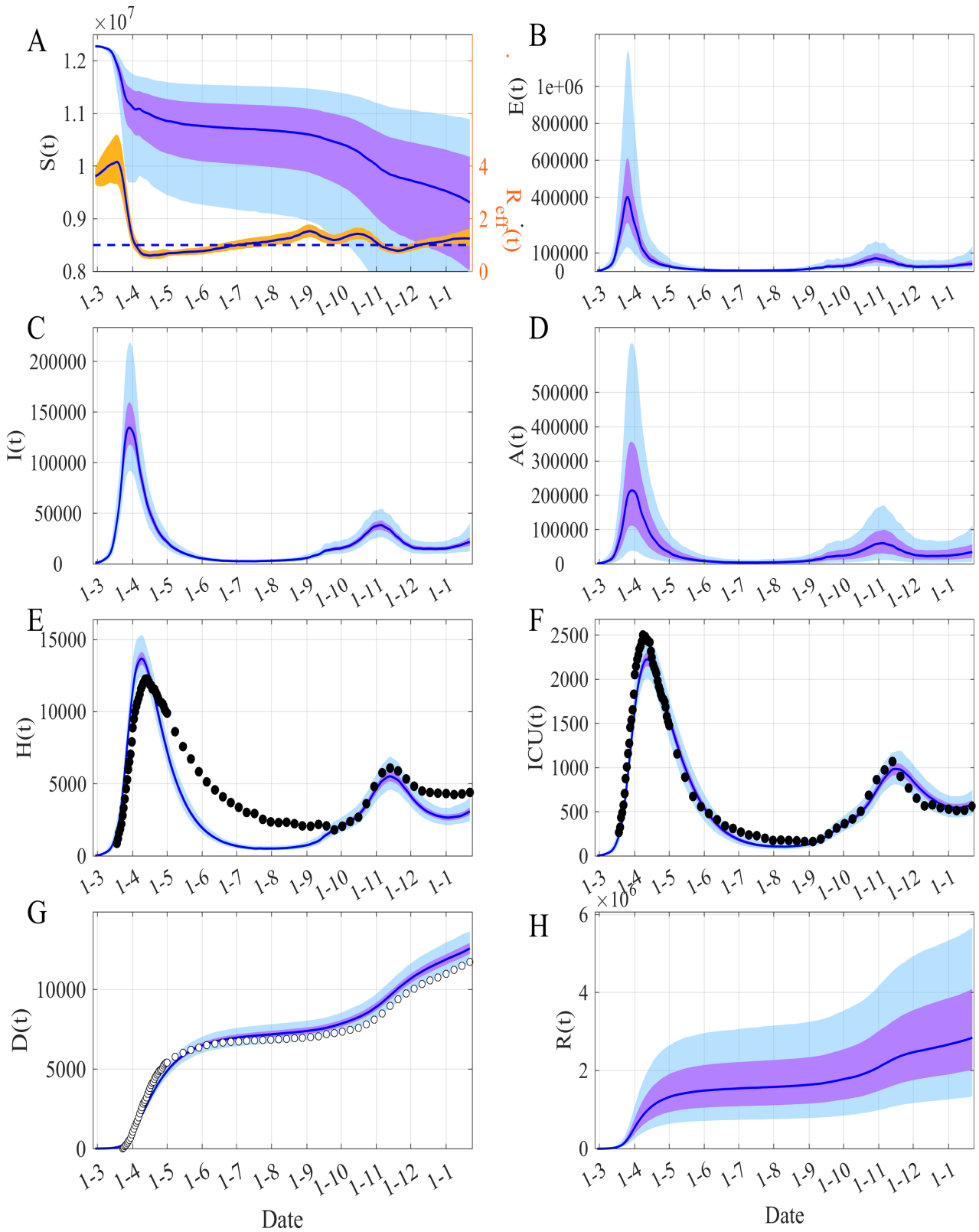

FIGURE 3 – Model dynamics of COVID-19 in **Ile-de-France, the Paris region**. (A) Time evolution of susceptibles $S(t)$ and $R_{eff}(t)$. (B) Infected non infectious, $E(t) = E_1(t) + E_2(t)$. (C) Symptomatic infectious $I(t) = I_1(t) + I_2(t)$. (D) Asymptomatic infectious $A(t) = A_1(t) + A_2(t)$. (E) Hospitalized people $H(t) = H_1(t) + H_2(t) + ICU(t)$. (F) Individuals in ICU, $ICU(t)$. (G) Cumulative death $D(t)$. (H) Removed $R(t)$. The blue lines are the median of the posterior estimates of the simulated trajectories, the mauve areas are the 50% Credible Intervals (CI) and the light blue areas the 95% CI. In (A) the orange area corresponds to the 50% CI of R_{eff}. The black points are observations used in by the inference process, the white points are the observations not used.

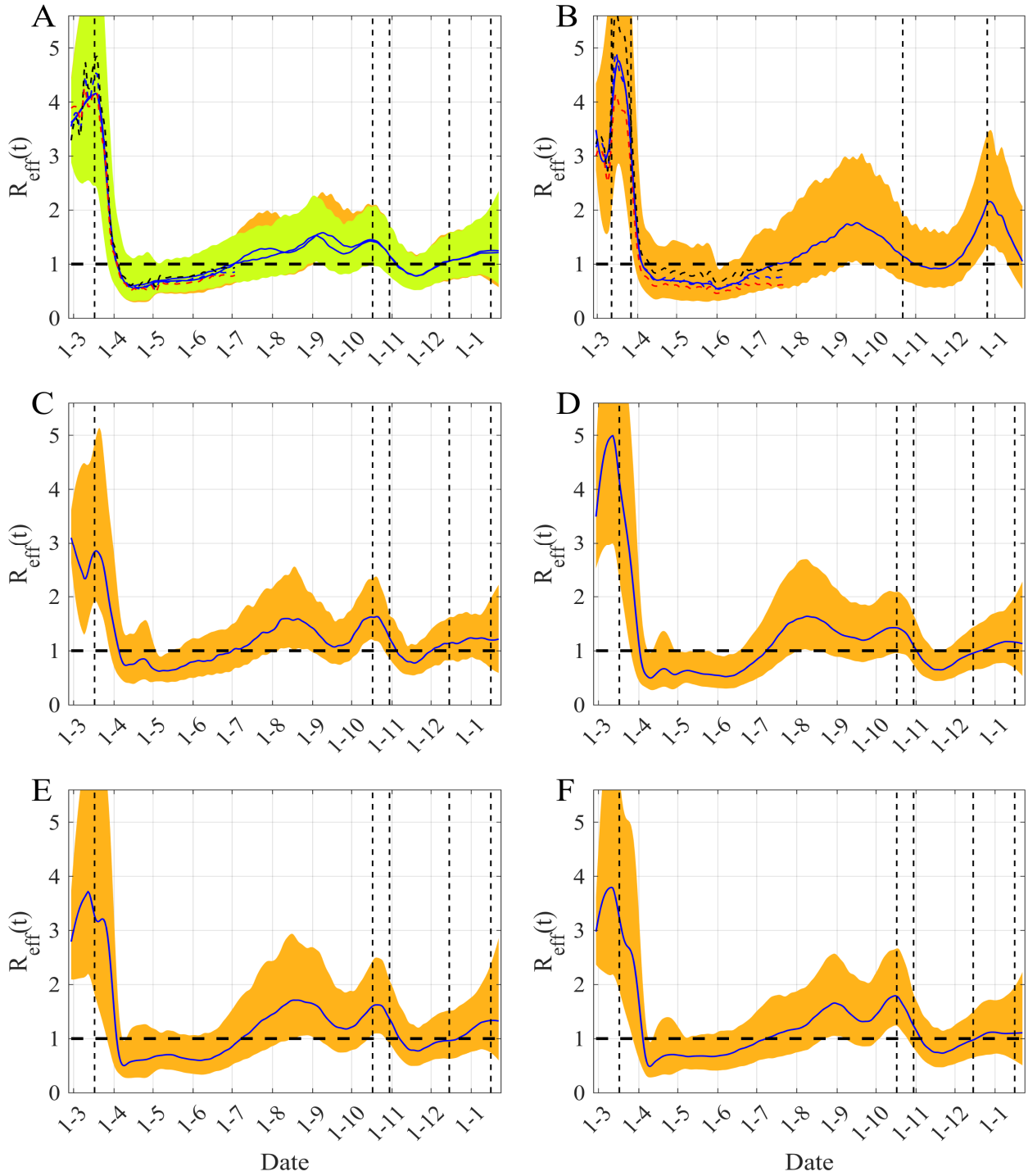

FIGURE 4 – Time varying $R_{eff}(t)$ in five regions of France and in Ireland. (A) Ile-de-France ; (B) Ireland ; (C) Provence Alpes Côte d'Azur ; (D) Occitanie ; (E) Nouvelle Aquitaine (F) Auvergne Rhône Alpes. The blue lines are the median of the posterior estimates of $R_{eff}(t)$ and orange and yellow areas are the 95% CI of R_{eff}. In (A) the orange area corresponds to the case where hospital discharges where included in the inference process, whereas the yellow area corresponds to the model that does not account for them. In (A) and (B) the dashed curves represent the median of $R_{eff}(t)$ for a preceding time period (blue), or computed with lower transmissibility of the asymptomatics, $q_A = 0.40$ (black) or computed with higher transmissibility, $q_A = 0.70$ (red). The vertical black dashed lines correspond to the start dates of the main mitigation measures.

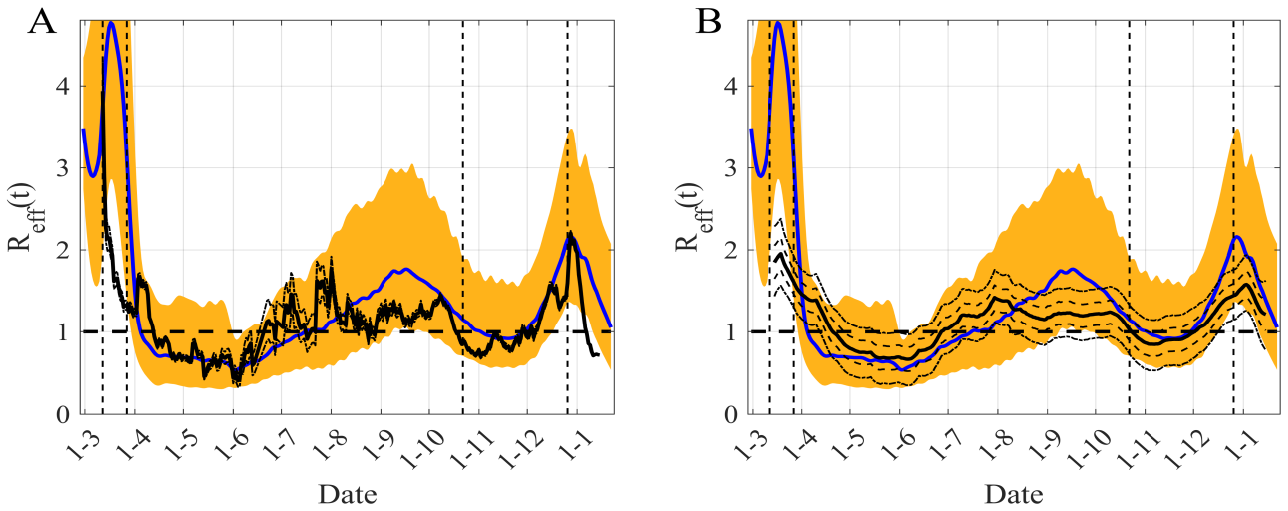

FIGURE 5 – Comparison between our $R_{eff}(t)$ estimation and those obtained with two other methods based on Irish multiple datasets. A/ Comparison with the method implemented in EpiEstim R package (http ://metrics.covid19-analysis.org/). B/ Comparison with the method proposed by Marioli et al (2020) (http ://trackingr-env.eba-9muars8y.us-east-2.elasticbeanstalk.com/). Blue lines are the median of the posterior of our estimates of $R_{eff}(t)$ and orange areas are the corresponding 95% CI of our R_{eff} estimates. The black lines represent the median of $R_{eff}(t)$ for the other methods and the black dotted-dashed lines delimit the corresponding 95% CI associated. In B/ the black dashed lines delimited the 65% CI. The vertical black dashed lines correspond to the start dates of the main mitigation measures.

A mechanistic and data-driven reconstruction of the time-varying reproduction number: Application to the COVID-19 epidemic

Bernard Cazelles[a,b,c], Clara Champagne[d,e], Benjamin Nguyen Van Yen[c,f], Catherine Comiskey[g], Elisabeta Vergu[b,☯], Benjamin Roche[h,☯]

a. Sorbonne Université, UMMISCO, Paris, France
b. INRAE, Université Paris-Saclay, MaIAGE, Jouy-en-Josas, France
c. Eco-Evolution Mathématique, IBENS, UMR 8197, CNRS, Ecole Normale Supérieure, Paris, France
d. Swiss Tropical and Public Health Institute, Bassel, Switzerland
e. Universty of Basel, Basel, Switzerland
f. Institut Pasteur, Unité de Génétique Fonctionnelle des Maladies Infectieuses, Paris, France
g. School of Nursing and Midwifery, Trinity College Dublin, The University of Dublin, Dublin, Ireland
h. MIVEGEC, IRD, CNRS and Université de Montpellier, Montpellier, France
☯ These authors contributed equally to this work.

Supporting Information

State-space framework

We have used a state-space framework defined by two sets of equations: the first set describes the propagation of the disease in the population and the second corresponds for the observation process. This allows to consideration of unknowns and uncertainty both in the epidemiological process and in the observation of epidemic dynamics:

$$\begin{cases} \dot{x}(t) = g\left(t, x(t), \theta(t), u(t)\right) \\ y(t) \mid x(t) \sim f\left(h(x(t)), \phi(t)\right) \end{cases} \tag{S1}$$

The first equation of (S1) corresponds to the epidemiological model, with $x(t)$ representing the state variables (for instance, $S(t)$ the susceptibles, $E(t)$ the infected non-infectious, $I(t)$ the infectious and $R(t)$ the removed for the classical SEIR model) and $\theta(t)$ the epidemiological parameters, among which some are time-dependent. The second equation is associated to the observational process defined by the probabilistic distribution f with parameters depending on a functional of some of the epidemiological model coordinates, $h(x(t))$ and some observational and/or measurement error, $\phi(t)$. $y(t)$ corresponds to observed epidemic dynamics (partial and noisy observations of $x(t)$) and $u(t)$ is the process noise describing intrinsic or environmental stochasticity other than the observational noise is included in f. In our applications, $h(x(t))$ will represent the daily new cases.

Epidemiological Model – hidden state

Our objective was to propose a simple but relevant model that incorporates the available information to examine the dynamics of Covid-19 epidemics before, during and after the lock-down.

Our model is an extended stochastic SEIR model also accounting for asymptomatic transmission and the hospital course of the hospitalized infectious. It is similar to other models that have already been proposed to model and forecast the Covid-19 epidemic (Li et al, 2020; Prague et al, 2020, Di Domenico et al, 2020; Liu et al, 2020; Prem et al, 2020). It includes the following variables the susceptibles *S*, the infected non-infectious *E*, the symptomatic infectious *I*, the asymptomatic infectious *A*, the removed people *R*, and the hospital-related variables: hospitalized people *H*, people in intensive care unit *ICI*, cured people *G*, and deaths at hospital *D*. We have also introduced Erlang-distributed stage durations (with a shape parameter equal to 2) in *E*, *I*, *A* and *H* compartments, to relax the mathematically convenient but unrealistic assumption of exponential stage durations. For sake of simplicity, we present here below the differential equations describing the deterministic version of our model, whereas its stochastic version was used in this study:

$$
\begin{aligned}
\frac{dS}{dt} &= -\beta(t).\left(I_1 + q_1.I_2 + q_2.(A_1 + A_2)\right) \\
\frac{dE_1}{dt} &= \beta(t).\left(I_1 + q_1.I_2 + q_2.(A_1 + A_2)\right) - \frac{\sigma}{2}.E_1 \\
\frac{dE_2}{dt} &= \frac{\sigma}{2}.(E_1 - E_2) \\
\frac{dI_1}{dt} &= (1-\tau_A).\frac{\sigma}{2}.E_2 - \frac{\gamma}{2}.I_1 \\
\frac{dI_2}{dt} &= \left(1-\tau_H - q_I.\tau_I\right)\frac{\gamma}{2}.I_1 - \frac{\gamma}{2}.I_2 \\
\frac{dA_1}{dt} &= \tau_A.\frac{\sigma}{2}.E_2 - \frac{\gamma}{2}.A_1 \\
\frac{dA_2}{dt} &= \frac{\gamma}{2}.(A_1 - A_2) \\
\frac{dR}{dt} &= \frac{\gamma}{2}.\left(\left(1-\tau_H - q_I.\tau_I\right).I_2 + A_2\right) + (1-q_D.\tau_D).\frac{\kappa}{2}.H_2 \qquad \text{(S2)} \\
\frac{dH_1}{dt} &= \tau_H.\frac{\gamma}{2}.\left(I_1 + I_2\right) - \frac{\kappa}{2}.H_1 \\
\frac{dH_2}{dt} &= \left(1-\tau_I - q_D.\tau_D\right).\frac{\kappa}{2}.H_1 + (1-\tau_D).\delta.ICU - \frac{\kappa}{2}.H_2 \\
\frac{dICU}{dt} &= q_I.\tau_I.\frac{\gamma}{2}.\left(I_1 + I_2\right) + \tau_I.\frac{\kappa}{2}.H_1 - \delta.ICU \\
\frac{dG}{dt} &= \left(1-q_D.\tau_D\right).\frac{\kappa}{2}.H_2 \\
\frac{dD}{dt} &= q_D.\tau_D.\frac{\kappa}{2}.\left(H_1 + H_2\right) + \tau_D.\delta.ICU
\end{aligned}
$$

The main originality of this model is the time-varying transmission rate $\beta(t)$ that follows a Brownian diffusion process, equation (1) in the main text. The other parameters are: σ is the incubation rate, γ the recovery rate, $1/\kappa$ the average hospitalizion period, $1/\delta$ the average time spent in ICU, τ_A the fraction of asymptomatics, τ_H the fraction of infectious hospitalized, τ_I the fraction of ICU admission, τ_D the death rate, q_1 and q_2 the reduction in transmissibility of I_2 and asymptomatics (A_1+A_2), respectively. As the peaks of hospitalized and admitted to ICU individuals are concomitant, we consider that a small fraction, $q_I.\tau_I$ of infectious with severe symptoms enter directly the ICU. Even if the majority of deaths occur in the ICU, a fraction, $q_D.\tau_D$, can occur in the hospital, outside the intensive care. Then, q_I and q_D are the reduction in admission in ICU and in death rate, respectively. Most of the parameters are estimated in our inference process and the remaining ones set to plausible values, in agreement with the literature (see Table S1).

Our inference approach is mainly based on data on new entries in several model compartments. These observed variables used are described in the following equations:

$$
\begin{aligned}
C_I(t) &= \int_t^{t+1} (1-\tau_A).\frac{\sigma}{2}.E_2(t).dt \\
C_H(t) &= \int_t^{t+1} \tau_H.\frac{\gamma}{2}.\left(I_1(t)+I_2(t)\right).dt \\
C_{ICU}(t) &= \int_t^{t+1} \left(q_I.\tau_I.\frac{\gamma}{2}.\left(I_1(t)+I_2(t)\right)+\tau_I.\frac{\kappa}{2}.H_1(t)\right).dt \qquad \text{(S3)} \\
C_G(t) &= \int_t^{t+1} \left(1-q_D.\tau_D\right).\frac{\kappa}{2}.H_2(t).dt \\
C_D(t) &= \int_t^{t+1} \left(q_D.\tau_D.\frac{\kappa}{2}.\left(H_1(t)+H_2(t)\right)+\tau_D.\delta.ICU(t)\right).dt
\end{aligned}
$$

where C_I is the incidence of symptomatic cases, C_H the daily admission to hospital, C_{ICU} the daily new admission in ICU, C_D the new daily number of of hospital deaths and C_G the daily new discharge at hospital.

Observational Model

Our model trajectories are related to the data using a negative binomial observation model, which is commonly used (Bretó et al, 2009). The observed incidences, $C_{k,obs}(t)$, are assumed to be drawn from a negative binomial distribution with mean $\rho_k.C_k(t)$ and variance $\rho_k.C_k(t) + \phi_k.(\rho_k.C_k(t))^2$. For each k, $C_k(t)$ represents the number of events occurring during day t simulated by the model, $\rho_k \in [0,1]$ is an observation rate (or reporting rate) and ϕ_k is an over-dispersion parameter.

Current hospital data, $H_{obs}(t)$ (corresponding to H_1+H_2+ICU) and $ICU_{obs}(t)$ (corresponding to ICU), are also available and therefore used to fit the model. We make the assumption that these variables $X_{k,obs}$ follow a discretized normal distribution with fixed standard deviation proportional to its estimated higher value and with mean equal to the predicted values by the

model X_k (equations S2) times an observation rate ρ_k (see Table S1) and with fixed standard deviation proportional to its estimated highest value.

All variables are assumed to be independent conditional on the underlying epidemiological model, and therefore the model likelihood is the product of the likelihoods for all observed variables.

Inference

Stochastic framework

Equations (S1-S3) are considered in a stochastic framework solved with the Euler-Maruyama algorithm (Kloeden and Platen, 1999) implemented in the SSM platform (Dureau et al, 2013b).

Estimation using the particle Markov Chain Monte Carlo (PMCMC) method

The likelihood of our stochastic model being intractable, it was estimated with particle filtering method (Sequential Monte Carlo, SMC). With a given set of parameters, the SMC algorithm reconstructs sequentially the trajectory of the state variables and the time-varying parameters, and allows to compute the associated likelihood. Firstly, the distribution of the initial conditions of the system is approximated by a sample of particles. Then, at each iteration, the particles are projected according to the epidemiological and observation models up to the next observation point. To each of the particles a weight is associated, reflecting the quality of its prediction compared to the observation, and the total likelihood is updated. A resampling step using the weights is performed before the next iteration, in order to discard the trajectories associated with low weight particles.

In order to estimate the parameters of the system, the particle filter is embedded in a Markov Chain Monte Carlo framework, leading to the PMCMC algorithm (Andrieu et al, 2010). More precisely, the likelihood estimated by the SMC method is used within a Metropolis Hasting scheme (particle marginal Metropolis Hastings) (Andrieu et al, 2010). The proposal distribution is a Gaussian whose co-variance matrix is adapted following the framework described in Dureau et al (2013b).

The starting point of the MCMC chain is initialized using optimal values obtained from a simplex algorithm adapted for stochastic system (KSimplex), on a large number of parameter sets. Then, a pre-adaptation of the proposal co-variance matrix is performed with Kalman MCMC (KMCMC). The main idea behind these algorithmic improvements relies on less computationally costly algorithms in order to facilitate the exploration of parameter space. As our model is not Gaussian, we approximate the likelihood using the extended Kalman filter both in the simplex algorithm (KSimplex) (Dureau et al, 2013b) and in the MCMC (KMCMC) step (Dureau et al, 2013a). Then, the adaptive PMCMC is run on the output of the KMCMC step with

100,000 iterations and 2,000 particles for the estimates showed in figures included in the manuscript.

PMCMC algorithm as implemented in SSM (Dureau et al, 2013b)

In a model with n observations and J particles.
$q(.|\theta^{(i)})$ is the transition kernel

1. Initialize $\theta^{(0)}$
2. Using SMC algorithm, compute $\hat{p}(y_{1:n}|\theta^{(0)})$ and sample $x_{0:n}^{(0)}$ from $\hat{p}(x_{0:n}|y_{1:n},\theta^{(0)})$
3. **for** $(i = 0: N-1)$ **do**
4. Sample θ^* from $q(.|\theta^{(i)})$
5. Using SMC algorithm, compute $L(\theta^*) = \hat{p}(y_{1:n}|\theta^*)$ and sample $x_{0:n}^*$ from $\hat{p}(x_{0:n}|y_{1:n},\theta^*)$
6. Accept θ^* (and $x_{0:n}^*$) with probability $1 \wedge \frac{L(\theta^*)p(\theta^*)q(\theta^{(i)}|\theta^*)}{L(\theta^{(i)})p(\theta^{(i)})q(\theta^*|\theta^{(i)})}$
7. If accepted, $\theta^{(i+1)} = \theta^*$ and $x_{0:n}^{(i+1)} = x_{0:n}^*$. Otherwise, $\theta^{(i+1)} = \theta^{(i)}$ and $x_{0:n}^{(i+1)} = x_{0:n}^{(i)}$.
8. **end for**

References

- Andrieu C, Doucet A, Holenstein R. Particle markov chain monte carlo methods. Journal of the Royal Statistical Society: Series B. 2010;72:269-342.

- Bretó, C., He, D., Ionides, E.L., King, A.A. (2009). Time Series Analysis via Mechanistic Models. The Annals of Applied Statistics, 3, 319-348.

- Carrat, F., de Lamballerie, X., Rahib, D., Blanche, H., Lapidus, N., Artaud, F., ... & Lydie, N. Seroprevalence of SARS-CoV-2 among adults in three regions of France following the lockdown and associated risk factors: a multicohort study. MedRxiv. 2020;2020.09.16.20195693.

- Di Domenico L, Pullano G, Sabbatini CE, Boëlle PY, Colizza V. Impact of lockdown on COVID-19 epidemic in Île-de-France and possible exit strategies. BMC Medicine. 2020;18:240.

- Dureau J, Kalogeropoulos K, Baguelin M. Capturing the time-varying drivers of an epidemic using stochastic dynamical systems. Biostatistics. 2013a; 14(3): 541-555.

- Dureau, J., Ballesteros, S., Bogich, T. (2013b). SSM: Inference for time series analysis with State Space Models. https://github.com/JDureau/ssm/blob/master/doc/doc.pdf.

- Kloeden, P. & Platen, E. (1999). Numerical Solution to Stochastic Differential Equations. Springer.

- Le Vu, S., Jones, G., Anna, F., Rose, T., Richard, J. B., Bernard-Stoecklin, S., ... & Monnet, V. Prevalence of SARS-CoV-2 antibodies in France: results from nationwide serological surveillance. MedRxiv. 2020;2020.10.20.20213116.

- Liu Z, Magal P, Webb G. Predicting the number of reported and unreported cases for the COVID-19 epidemics in China, South Korea, Italy, France, Germany and United Kingdom. J Theor Biol. 2020;509:110501.

- Prague M, Wittkop L, Clairon Q, Dutartre D, Thiébaut R, Hejblum BP. Population modeling of early COVID-19 epidemic dynamics in French regions and estimation of the lockdown impact on infection rate. MedRxiv. 2020;2020.04.21.20073536.

- Prem K, Liu Y, Russell TW, Kucharski AJ, Eggo RM, Davies N et al. The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. The Lancet Public Health. 2020;5:e261-e270.

- Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, Shaman J. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science. 2020;368:489-493

- HSE. Preliminary report of the results of the Study to Investigate COVID-19 Infection in People Living in Ireland (SCOPI): A national seroprevalence study, June-July. 2020. Available from https://www.hpsc.ie/a-z/respiratory/coronavirus/novelcoronavirus/scopi/SCOPI%20report%20preliminary%20results%20final%20version.pdf (accessed 29th September 2020)

- Warszawski, J., Bajos, N., Meyer, L., de Lamballerie, X., Seng, R., Beaumont, A.-L., Slama, R., Hisbergues, M., Rahib, D., Lydié, N., Legendre, B., Barlet, M., Rey, S., Raynaud, P., Leduc, A., Costemalle, V., Beck, F. Legleye, S., Castell, L., Givord, P., Favre-Martinoz, C., Paliod, N., Silhol, J., Sillard, P., 2020. Ministère des Solidarités et de la Santé: enquête nationale EpiCov. https://drees.solidarites-sante.gouv.fr/IMG/pdf/er1167.pdf.

Table S1. Definition of the different parameters and their priors for Ile-de-France region, Ireland and four other French regions: Provence Alpes Côte d'Azur (PACA), Occitanie (OC), Nouvelle-Aquitaine (NA), Auvergne Rhône Alpes (ARA).
U stands for uniform distribution and tN for truncated normal distribution (tN[mean,std,limit inf,limit sup]).

Parameters	Definitions	Prior or constant value	Prior or constant value	Prior or constant value
		Ile de France*	**Ile de France****	**Ireland**
$I_1(0)$	Initial condition	U[10,1500]	U[10,1500]	U[5,100]
$S(0)$	Initial condition	N=12278000	N=12278000	N=5176000
$E_1(0)$, $E_2(0)$, $I_2(0)$, $A_1(0)$, $A_2(0)$	Initial conditions	Use of steady-state conditions ***	Use of steady-state conditions ***	Use of steady-state conditions ***
$H_1(0)$, $H_2(0)$, $ICU(0)$, $D(0)$, $G(0)$, $R(0)$	Initials conditions	0	0	0
$\beta(0)$	Initial condition of the transmission rate	0.85	0.85	0.70
ν	Volatility of the Brownian process	U[0.02,015]	U[0.02,015]	U[0.05,015]
$1/\sigma$	average duration of the incubation period	tN[4,0.1,3,5]	tN[4,0.1,3,5]	tN[4,0.1,3,5]
$1/\gamma$	average duration of the infectious period	tN[6,0.2,4.5,7.5]	tN[6,0.2,4.5,7.5]	tN[6,0.2,4.5,7.5]
$1/\kappa$	average hospitalizion period	U[10,20]	U[10,20]	U[8,20]
$1/\delta$	average time spent in ICU	U[10,20]	U[10,20]	U[8,20]
τ_A	fraction of asymptomatics	U[0.20,0.80]	U[0.20,0.80]	U[0.20,0.80]
τ_H	fraction of hospitalization	U[0.025,0.20]	U[0.025,0.20]	U[0.025,0.20]
τ_I	fraction of ICU admission	U[0.05,0.20]	U[0.05,0.20]	U[0.025,0.20]
τ_D	death rate	U[0.10,0.80]	U[0.10,0.80]	U[0.10,0.70]
q_1	reduction in transmissibility	$1.5*q_2$ but≤1	$1.5*q_2$ but≤1	$1.5*q_2$ but≤1
q_2	reduction in transmissibility	0.55 (Li et al, 2020)	0.55 (Li et al, 2020)	0.55 (Li et al, 2020)
q_I	reduction in ICU admission fraction	0.05	0.05	0.30
q_D	reduction in death rate	0.10	0.10	0.40
ρ_I	reporting rate for symptomatic infectious	U[0.01, 0.10]	U[0.01, 0.10]	U[0.05, 0.40]
ρ_H	reporting rate for hospitalized people	U[0.95,1]	U[0.95,1]	U[0.95,1]
ρ_{ICU}	reporting rate for ICU admission	0.96	0.96	0.96
ρ_G	reporting rate for hospital discharge	0.96	0.96	0.96
ρ_D	reporting rate for death	0.98	0.98	0.98

*using hospital discharge data and **not using hospital discharge data

***steady-state conditions are defined by: $\frac{dE_1}{dt} = \frac{dE_2}{dt} = \frac{dI_1}{dt} = \frac{dI_2}{dt} = \frac{dA_1}{dt} = \frac{dA_2}{dt} = 0$

Table S1. (continued)

Parameters	Prior or constant value	Prior or constant value	Prior or constant value	Prior or constant value
	PACA	**OC**	**NA**	**ARA**
$I_1(0)$	U[10,500]	U[10,300]	U[10,200]	U[10,500]
$S(0)$	N=5055000	N=5845000	N=5957000	N=5176000
$E_1(0)$, $E_2(0)$, $I_2(0)$, $A_1(0)$, $A_2(0)$,	Use of steady-state conditions ***	Use of steady-state conditions ***	Use of steady-state conditions ***	Use of steady-state conditions ***
$H_1(0)$, $H_2(0)$, $ICU(0)$, $D(0)$, $G(0)$, $R(0)$	0	0	0	0
$\beta(0)$	0.70	0.85	0.70	0.70
ν	U[0.02,015]	U[0.02,015]	U[0.02,015]	U[0.02,015]
$1/\sigma$	tN[4,0.1,3,5]	tN[4,0.1,3,5]	tN[4,0.1,3,5]	tN[4,0.1,3,5]
$1/\gamma$	tN[6,0.2,4.5,7.5]	tN[6,0.2,4.5,7.5]	tN[6,0.2,4.5,7.5]	tN[6,0.2,4.5,7.5]
$1/\kappa$	U[10,20]	U[8,18]	U[8,18]	U[10,20]
$1/\delta$	U[10,20]	U[10,20]	U[14,24]	U[10,20]
τ_A	U[0.20,0.80]	U[0.20,0.80]	U[0.20,0.80]	U[0.20,0.80]
τ_H	U[0.025,0.15]	U[0.02,0.20]	U[0.02,0.15]	U[0.02,0.20]
τ_I	U[0.05,0.15]	U[0.05,0.20]	U[0.05,0.15]	U[0.025,0.20]
τ_D	U[0.10,0.60]	U[0.10,0.60]	U[0.10,0.60]	U[0.10,0.60]
q_1	$1.5*q_2$ but≤1	$1.5*q_2$ but≤1	$1.5*q_2$ but≤1	$1.5*q_2$ but≤1
q_2	0.55 (Li et al, 2020)	0.55 (Li et al, 2020)	0.55 (Li et al, 2020)	0.55 (Li et al, 2020)
q_I	0.05	0.05	0.05	0.05
q_D	0.10	0.10	0.10	0.10
ρ_I	U[0.05, 0.15]	U[0.05, 0.20]	U[0.02, 0.15]	U[0.05, 0.40]
ρ_H	U[0.95,1]	U[0.95,1]	U[0.95,1]	U[0.95,1]
ρ_{ICU}	0.96	0.96	0.96	0.96
ρ_G	0.96	0.96	0.96	0.96
ρ_D	0.98	0.98	0.98	0.98

Table S2. Posteriors of the parameters for Ile-de-France region, Ireland, and four French regions: Provence Alpes Côte d'Azur (PACA), Occitanie (OC), Nouvelle-Aquitaine (NA), Auvergne Rhône Alpes (ARA).

Parameters	**Definitions**	**Posterior** Median, [95%CI]	**Posterior** Median, [95%CI]	**Posterior** Median, [95%CI]
		Ile de France*	**Ile de France****	**Ireland**
$I_1(0)$	Initial condition	709, [386-1253]	780, [386-1390]	15, [7-26]
ν	Volatility of the Brownian process	0.119, [0. 092-0.145]	0.113, [0.087-0.145]	0.133, [0.106-0.148]
$1/\sigma$	average duration of the incubation period	4.01, [3.84-4.17]	4.00, [3.85-4.17]	3.99, [3.85-4.16]
$1/\gamma$	average duration of the infectious period	6.01, [5.66-6.33]	6.01, [5.71-6.36]	6.00, [5.68-6.35]
$1/\kappa$	average hospitalizion period	14.56, [13.06-16.18]	11.95, [10.57-13.75]	11.88, [10.67-13.10]
$1/\delta$	average time spent in ICU	16.66, [14.52-18.86]	16.45, [14.35-18.71]	17.42, [15.17-19.45]
τ_A	fraction of asymptomatics	0.609, [0.258-0.795]	0.611, [0.266-0.789]	0.487, [0.226-0.771]
τ_H	fraction of hospitalization	0.028, [0.025-0.035]	0.028, [0.025-0.038]	0.100, [0.080-0.138]
τ_I	fraction of ICU admission	0.075, [0.067-0.087]	0.069, [0.060-0.084]	0.028, [0.025-0.036]
τ_D	death rate	0.521, [0.480-0.567]	0.4441, [0.446-0.533]	0.434, [0.397-0.472]
ρ_I	reporting rate for symptomatic infectious	0.027, [0.020-0.038]	0.025, [0.018-0.037]	0.297, [0.220-0.394]
ρ_H	reporting rate for hospitalized people	0.982, [0.954-0.998]	0.976, [0.953-0.998]	0.970, [0.952-0.995]

*using hospital discharge data and **not using hospital discharge data

Table S2. (continued)

Parameters	**Posterior** Median, [95%CI]	**Posterior** Median, [95%CI]	**Posterior** Median, [95%CI]	**Posterior** Median, [95%CI]
	PACA	**OC**	**NA**	**ARA**
$I_1(0)$	208, [116-419]	63, [35-138]	85, [35-148]	287, [162-527]
ν	0.118, [0. 093-0.144]	0.131, [0.107-0.148]	0.143, [0.125-0.148]	0.132, [0.105-0.148]
$1/\sigma$	4.03, [3.86-4.18]	3.99, [3.83-4.14]	3.97, [3.81-4.16]	3.99, [3.86-4.13]
$1/\gamma$	5.99, [5.67-6.28]	6.01, [5.61-6.31]	6.00, [5.72-6.32]	6.00, [5.72-6.34]
$1/\kappa$	12.47, [11.42-13.24]	11.28, [10.04-12.28]	13.75, [12.22-15.20]	15.88, [14.06-17.22]
$1/\delta$	14.18, [12.58-15.65]	12.17, [12.35-15.74]	18.73, [16.50-21.09]	15.44, [13.22-18.56]
τ_A	0.508, [0.249-0.758]	0.466, [0.236-0.764]	0.510, [0.241-0.779]	0.443, [0.220.745]
τ_H	0.030, [0.026-0.044]	0.031, [0.026-0.051]	0.029, [0.022-0.051]	0.032, [0.026-0.045]
τ_I	0.073, [0.063-0.091]	0.097, [0.080-0.129]	0.073, [0.056-0.100]	0.075, [0.065-0.093]
τ_D	0.399, [0.351-0.436]	0.354, [0.290-0.398]	0.440, [0.390-0.478]	0.485, [0.430-0.552]
ρ_I	0.051, [0.038-0.071]	0.036, [0.026-0.062]	0.033, [0.023-0.057]	0.029, [0.021-0.041]
ρ_H	0.976, [0.953-0.996]	0.991, [0.962-0.999]	0.983, [0.960-0.996]	0.976, [0.952-0.996]

Table S3. Seroprevalence results from different studies for Ile-de-France region, Ireland and four other French regions: Provence Alpes Côte d'Azur (PACA), Occitanie (OC), Nouvelle-Aquitaine (NA), Auvergne Rhône Alpes (ARA).

Regions or Country	**Seroprevalence 05-2020** Median, [95%CI] Le Vu et al, 2020	**Seroprevalence 05-2020 to 06-2020** Median, [95%CI] n = 14628 Carrat et al, 2020	**Seroprevalence 05-2020** Median, [95%CI] Warszawski et al, 2020	**Seroprevalence 06-2020** Median, [95%CI] HSE, SCOPI report, 2020
Ile-de-France	7.3%, [5.5%-9.4%] n = 536	10.0%, [9.1%-10.9%]	9.2%, [7.1%-11.2%] n = 2350	
Ireland				1.7%, [1.1%-2.4%] but 0.6%, [0.2%-1.4%] in the Sligo Provence
PACA	1.5%, [0.7%-2.6%] n=397		5.2%, [2.9%-7.5%] n=1643	
OC	1.2%, [0.4%-2.4%] n = 206		1.9%, [0.9%-2.9%] n = 581	
NA	1.4%, [0.7%-2.5%] n = 282	3.1%, [2.4%-3.7%]	2.0%, [0.8%-3.2%] n = 548	
ARA	2.8%, [1.5%-4.5%] n = 234		4.8%, [3.3%-6.2%] n = 732	

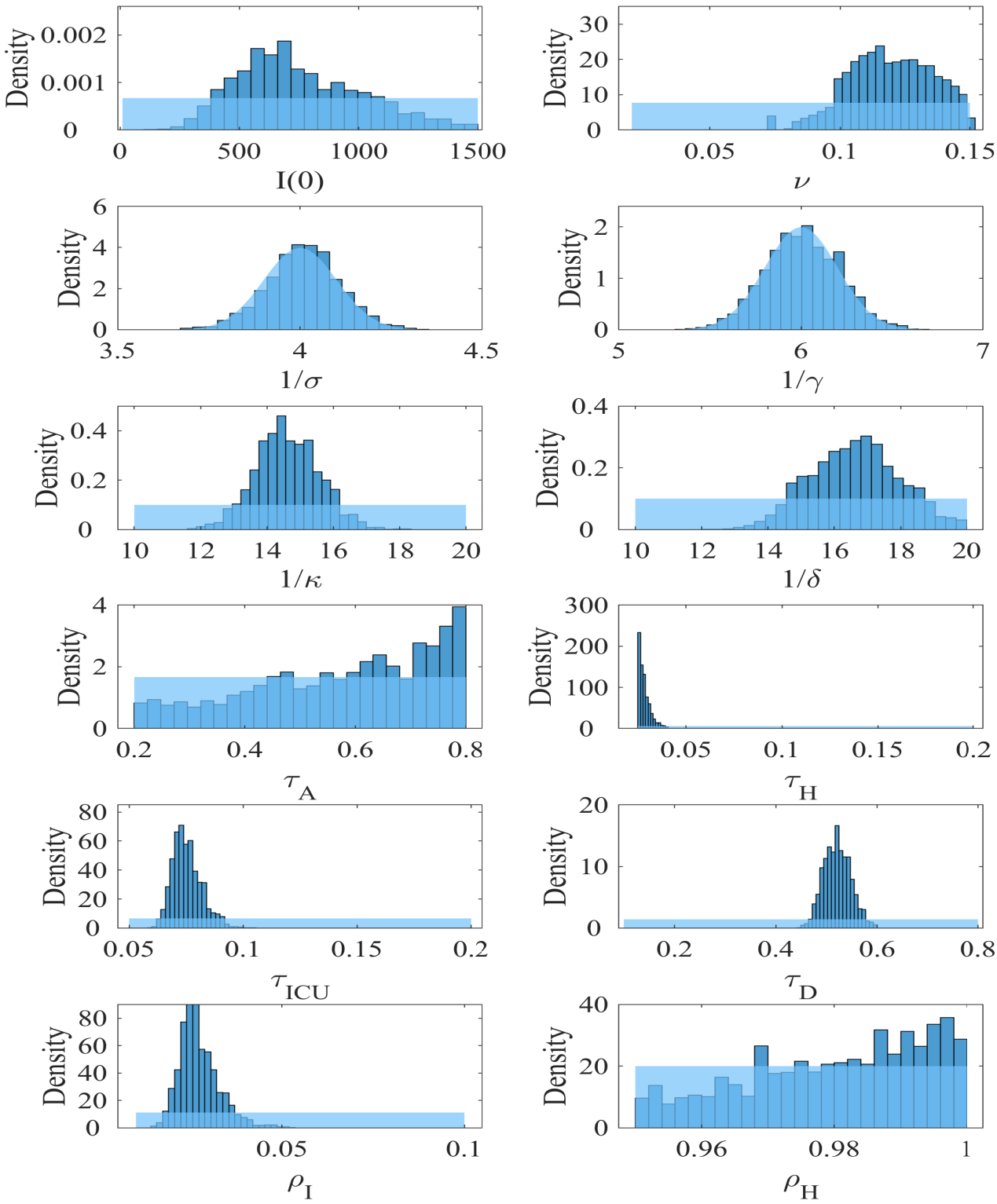

Fig. S1 – Prior and posterior distributions for the model inferences presented Fig. 2. $I_1(0)$ initial value of infectious, ν is the volatility of the Brownian process of $\beta(t)$, $1/\sigma$ the average duration of the incubation, $1/\gamma$ the average duration of infectious period, $1/\kappa$ the average hospitalized period, $1/\delta$ the average time spent in ICU, τ_A the fraction of asymptomatics, τ_H the fraction of infectious hospitalized, τ_I the fraction of ICU admission, τ_D the death rate, ρ_I the reporting rate for the infectious, ρ_H the reporting rate for the hospitalized people. The blue distributions are the priors and the discrete histograms are the posteriors.

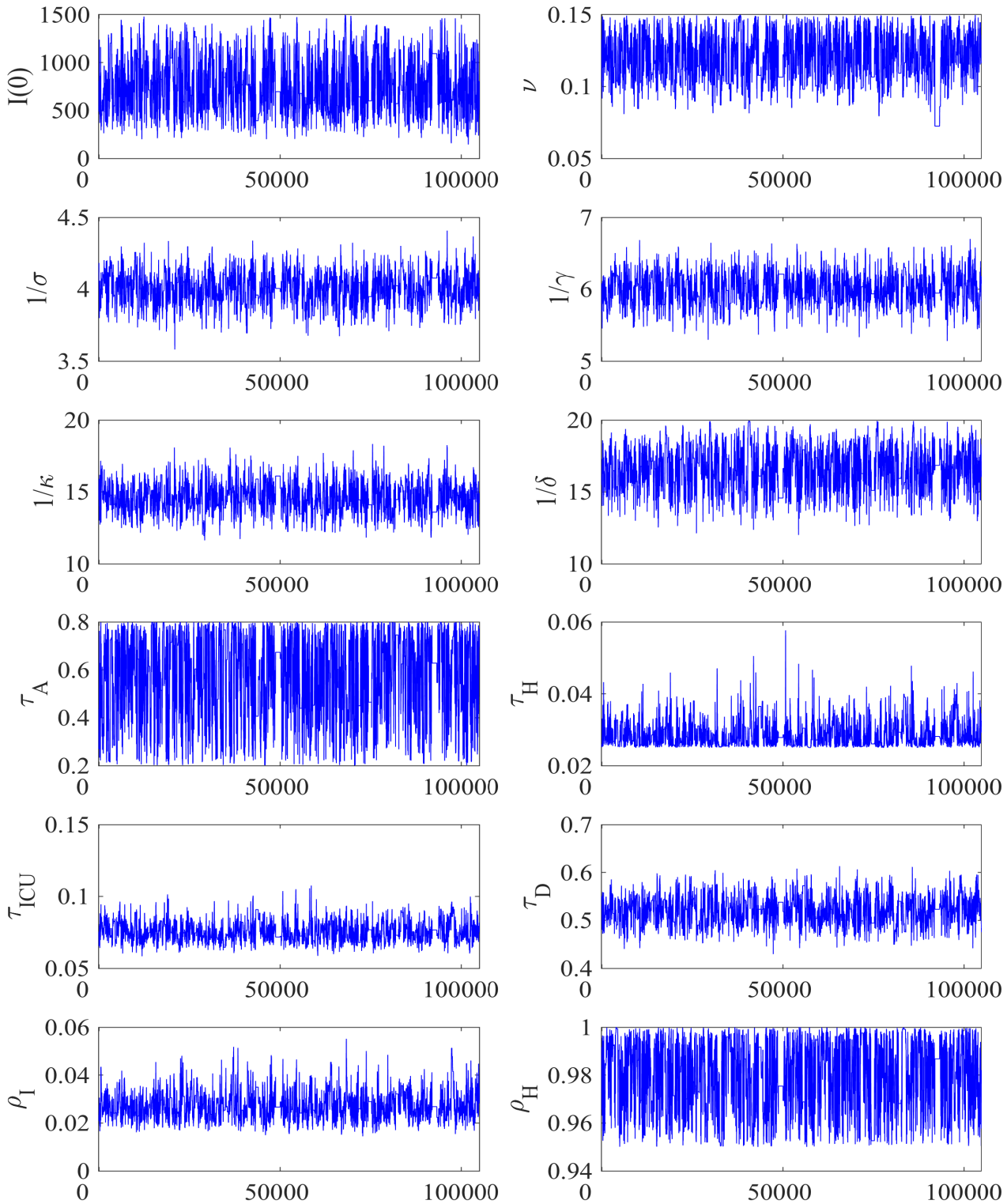

Fig. S2 – The traces of the MCMC chain for the model inferences of Fig 2.. $I_1(0)$ initial value of infectious, ν is the volatility of the Brownian process of $\beta(t)$, $1/\sigma$ the average duration of the incubation, $1/\gamma$ the average duration of infectious period, $1/\kappa$ the average hospitalized period, $1/\delta$ the average time spent in ICU, τ_A the fraction of asymptomatics, τ_H the fraction of infectious hospitalized, τ_I the fraction of ICU admission, τ_D the death rate, ρ_I the reporting rate for the infectious, ρ_H the reporting rate for the hospitalized people.

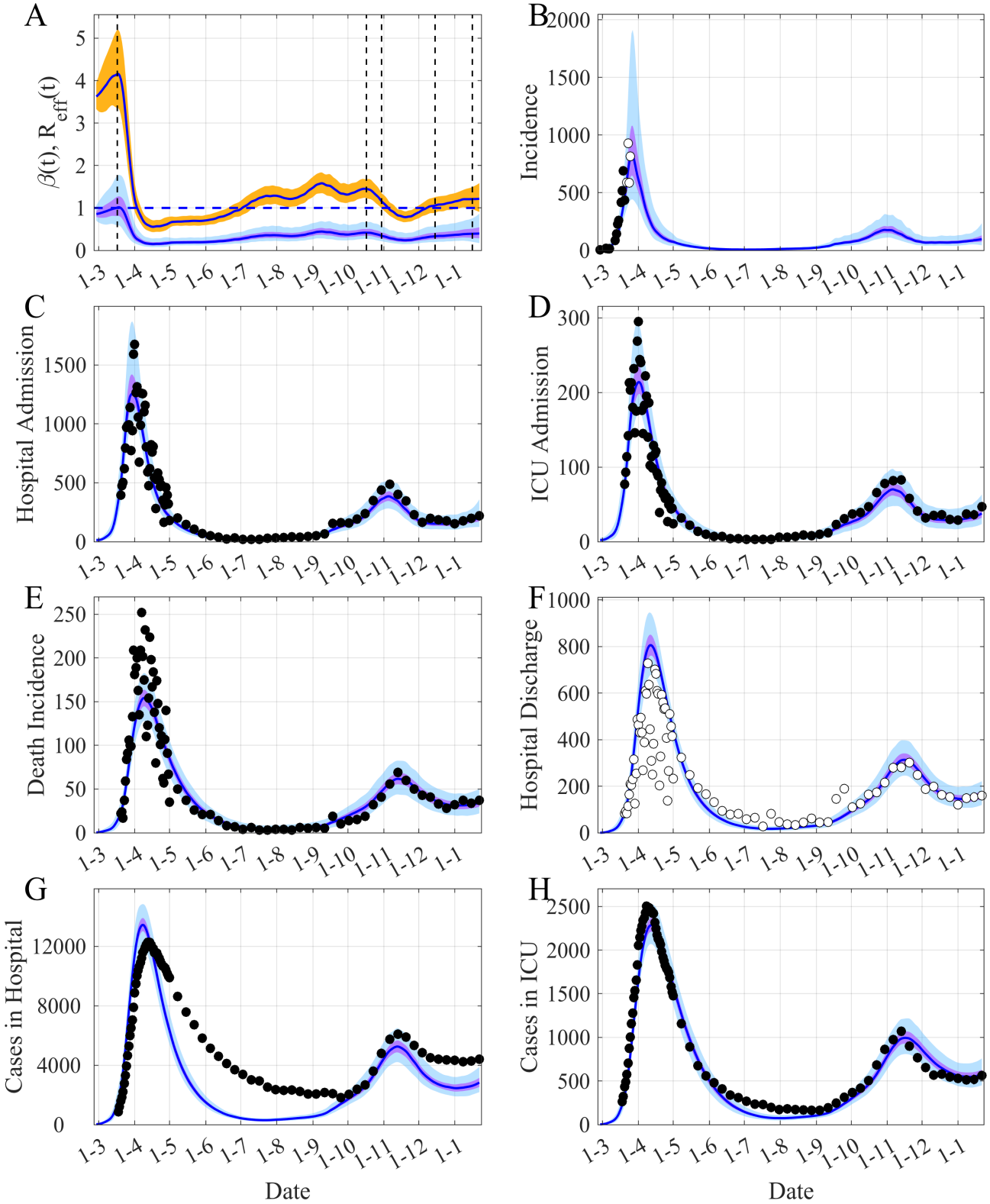

Fig. S3 – Reconstruction of the observed dynamics of COVID-19 in **Ile-de-France, the Paris region**. Caption as Fig. 2. The black points are observations used by the inference process, the white points are the observations not used.

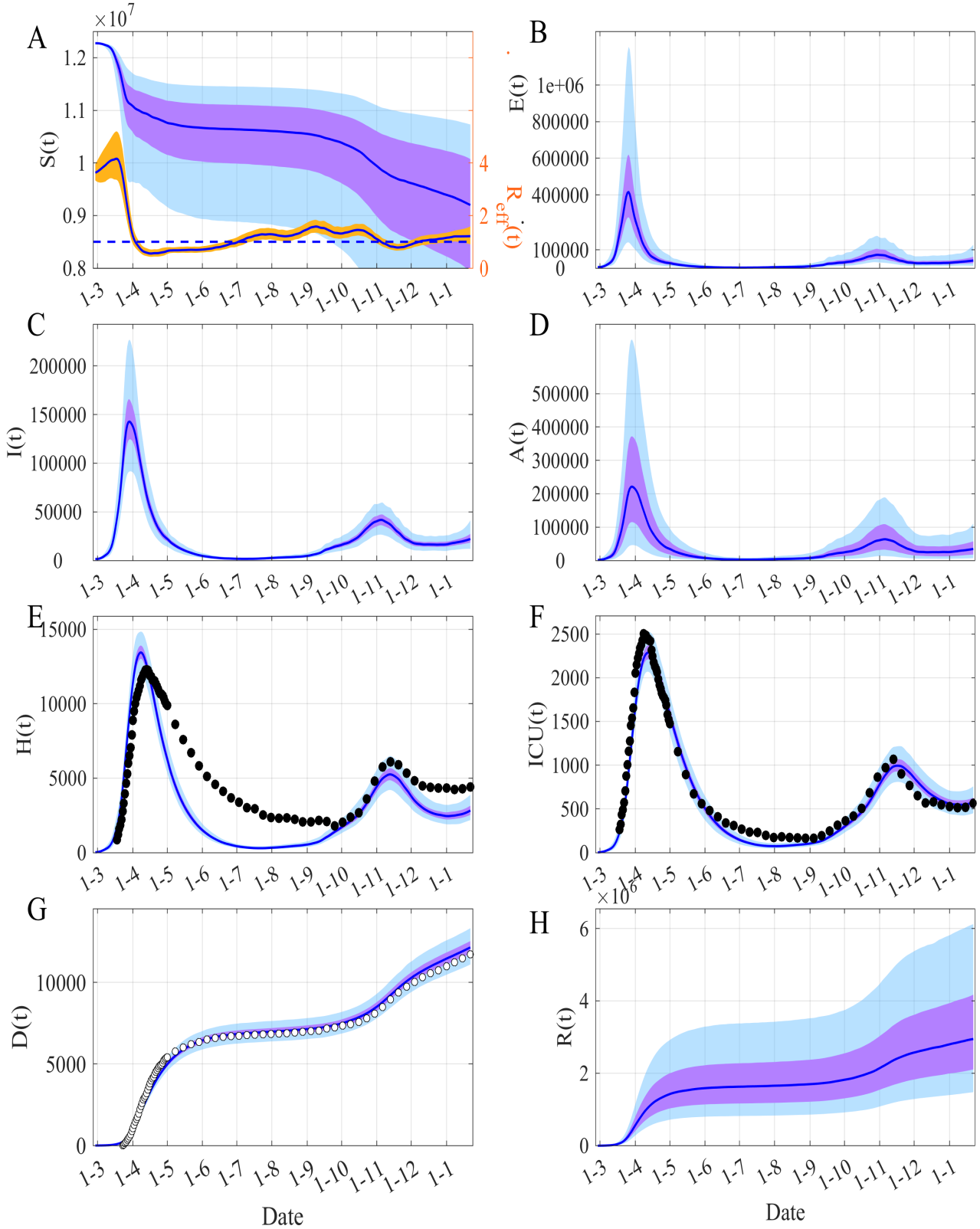

Fig. S4 – Dynamics of COVID-19 in **Ile-de-France, the Paris region**. Caption as Fig. 2. The black points are observations used by the inference process, the white points are the observations not used.

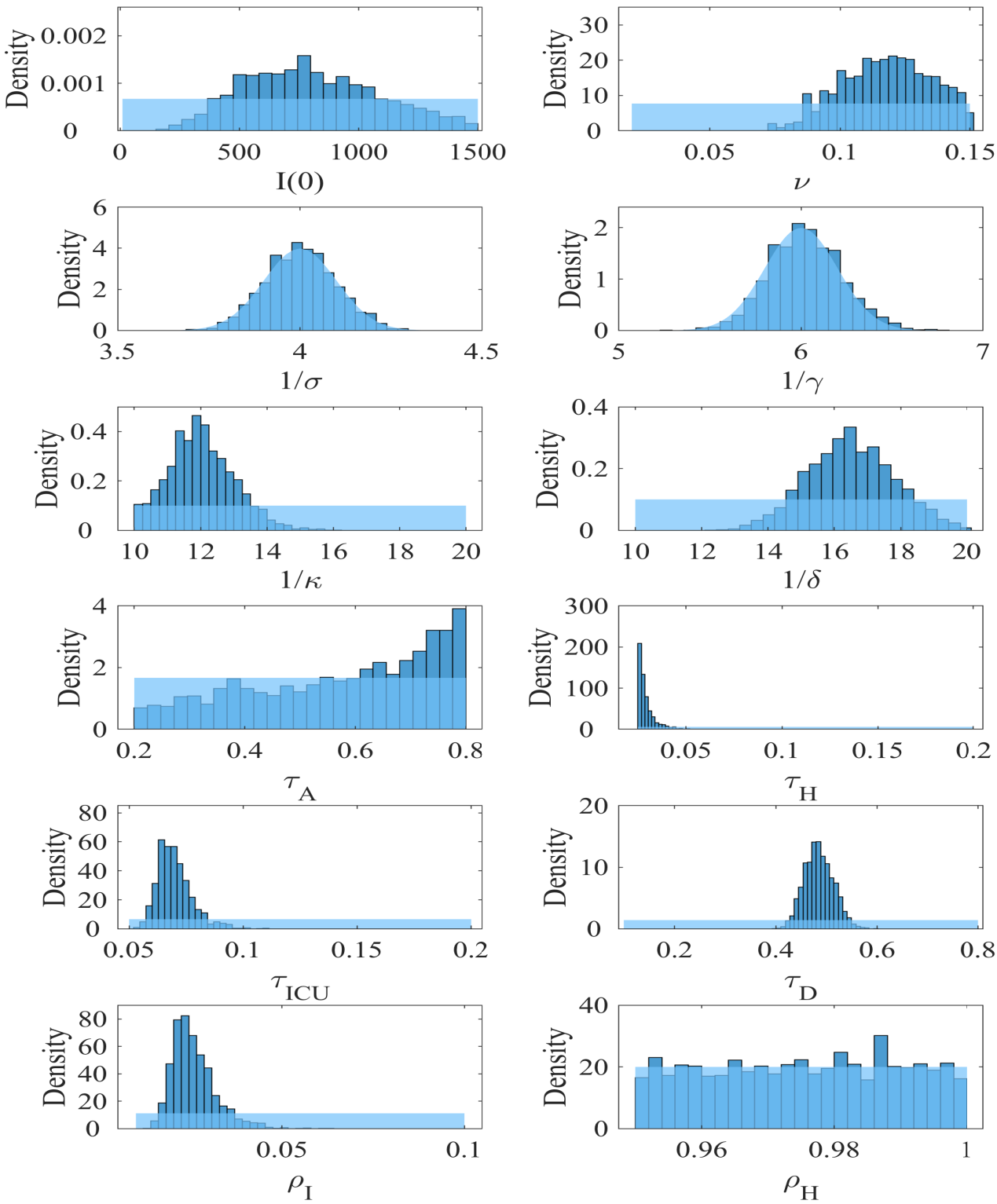

Fig. S5 – Prior and posterior distributions for the model inferences presented Fig. S3. Caption as Fig. S1.

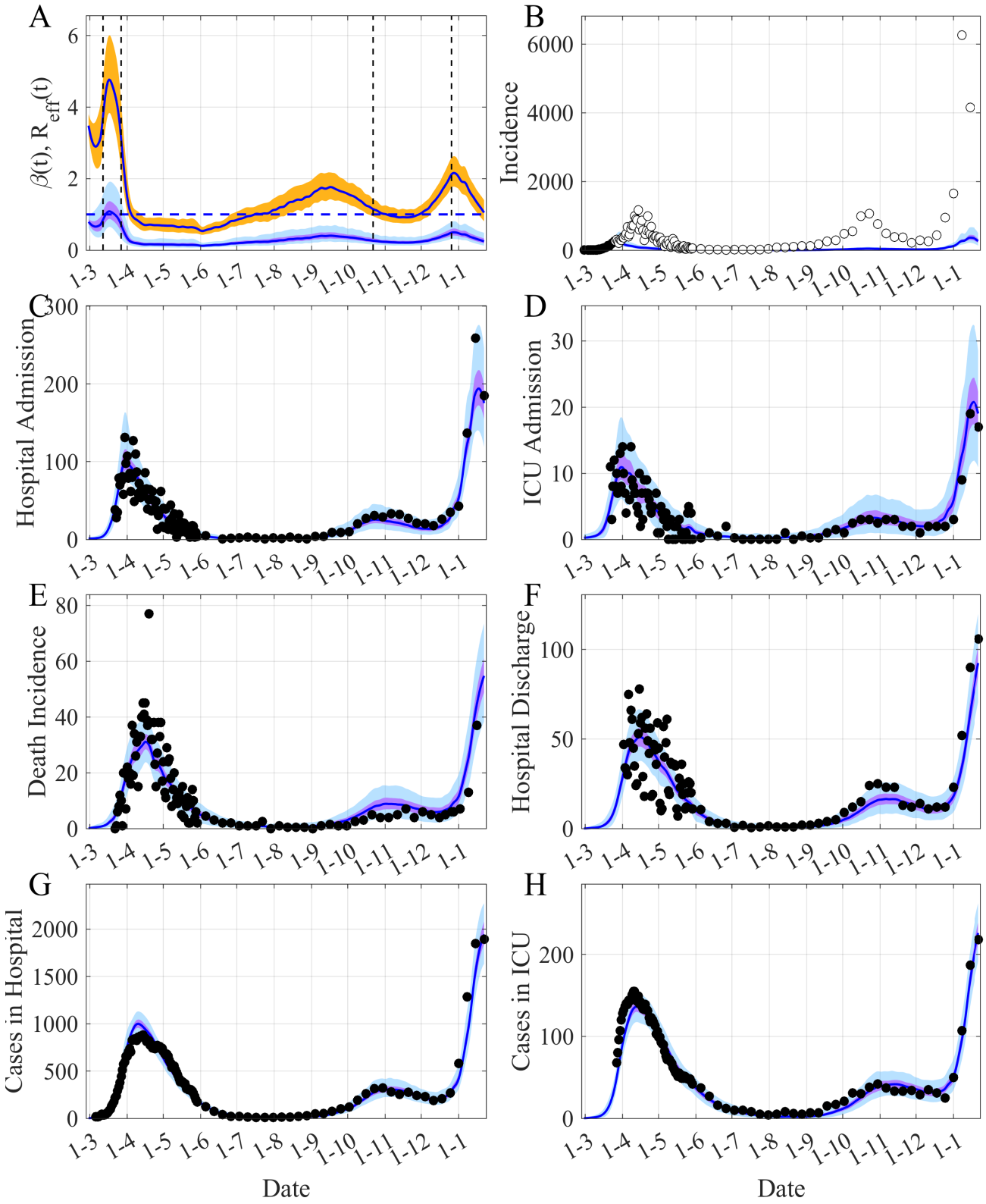

Fig. S6 – Reconstruction of the observed dynamics of COVID-19 in **Ireland**. Caption as Fig. 2 but average daily data of the current week is used after 01-06-2020. The black points are observations used by the inference process, the white points are the observations not used.

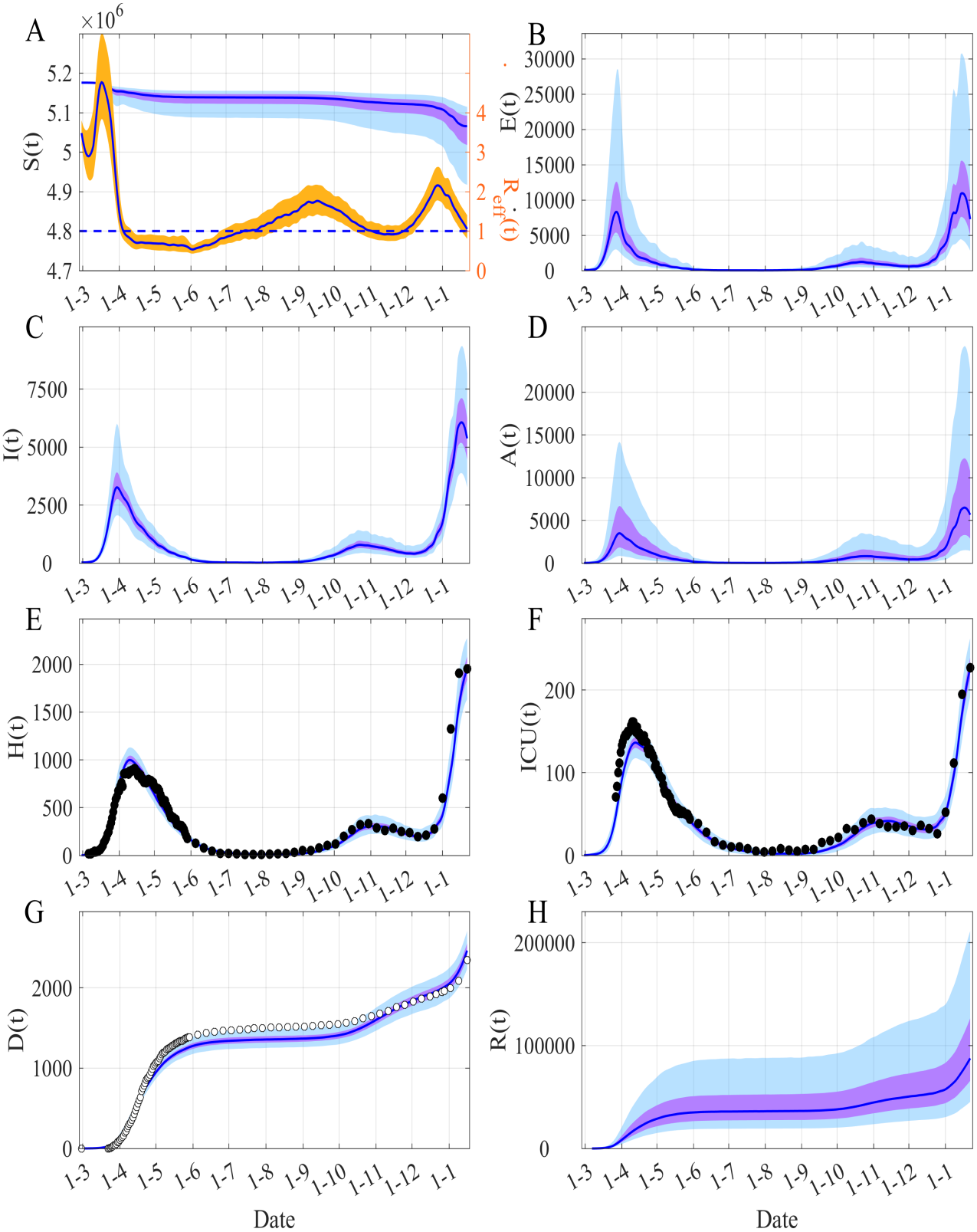

Fig. S7 – Dynamics of COVID-19 in **Ireland**. Caption as Fig. 2. The black points are observations used by the inference process, the white points are the observations not used.

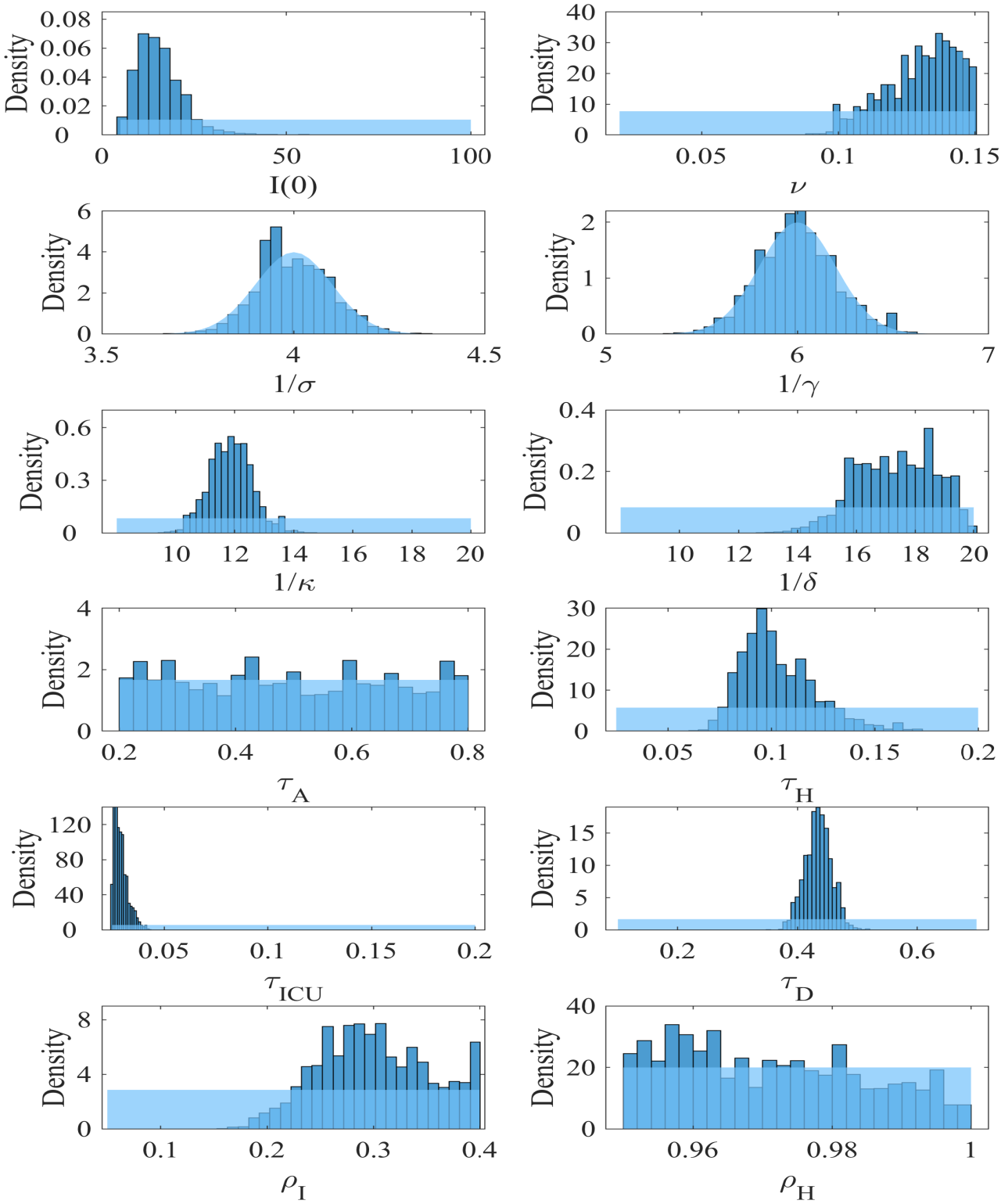

Fig. S8 – Prior and posterior distributions for the model inferences presented Fig. S6. Caption as Fig. S1.

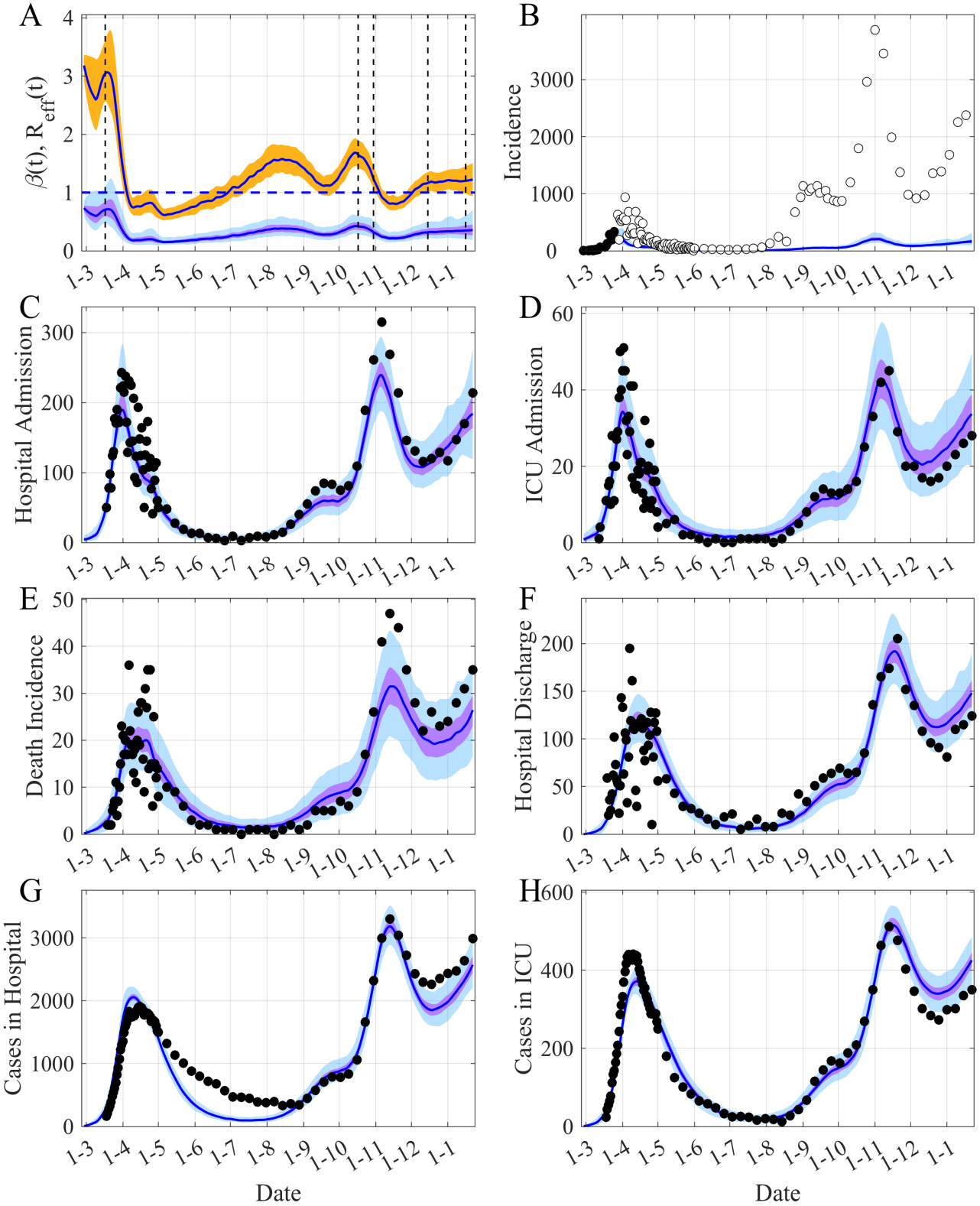

Fig. S9 – Reconstruction of the observed dynamics of COVID-19 in **Provence Alpes Côte d'Azur**. Caption as Fig. 2. The black points are observations used by the inference process, the white points are the observations not used.

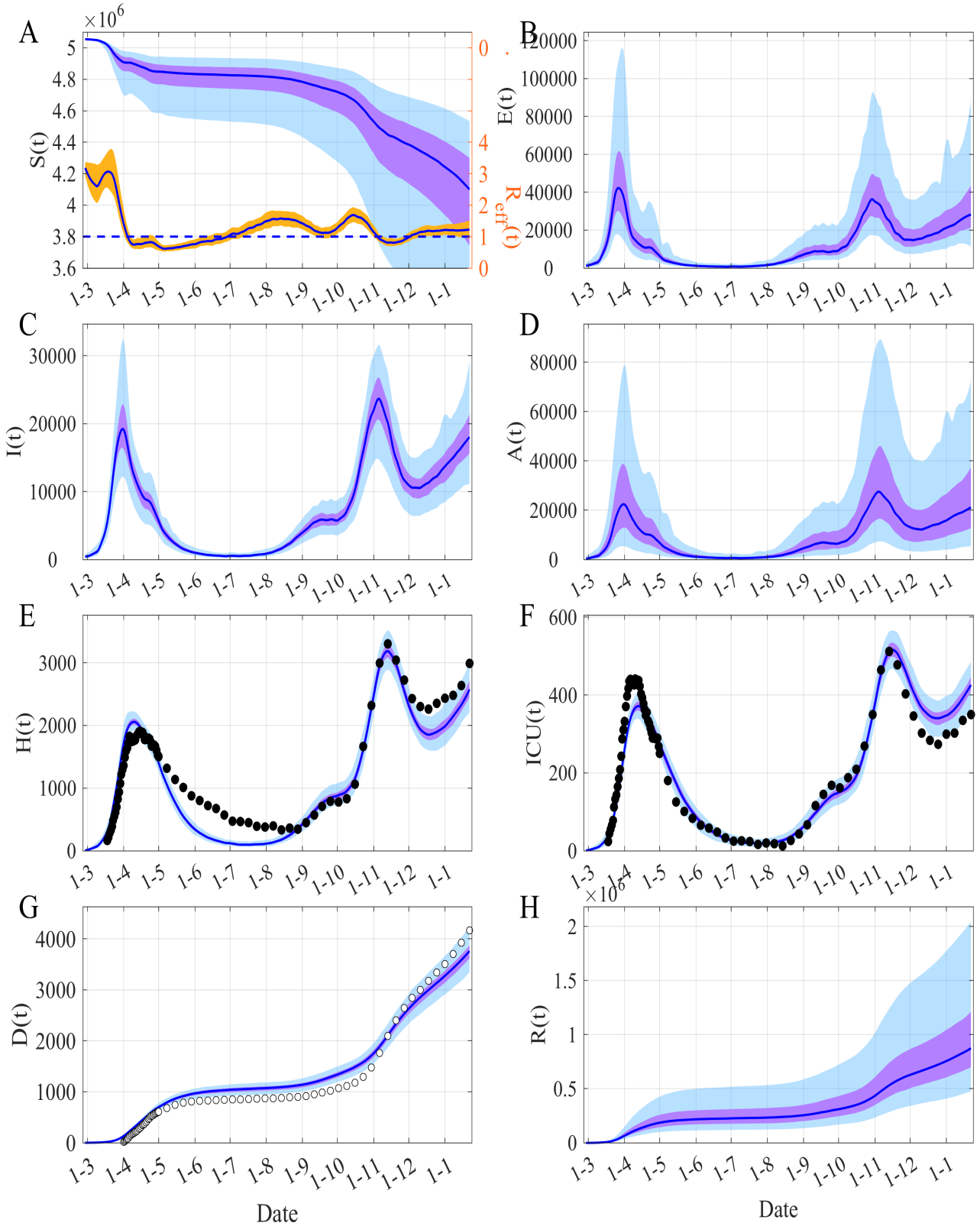

Fig. S10 – Dynamics of COVID-19 in **Provence Alpes Côte d'Azur**. Caption as Fig. 2. The black points are observations used by the inference process, the white points are the observations not used.

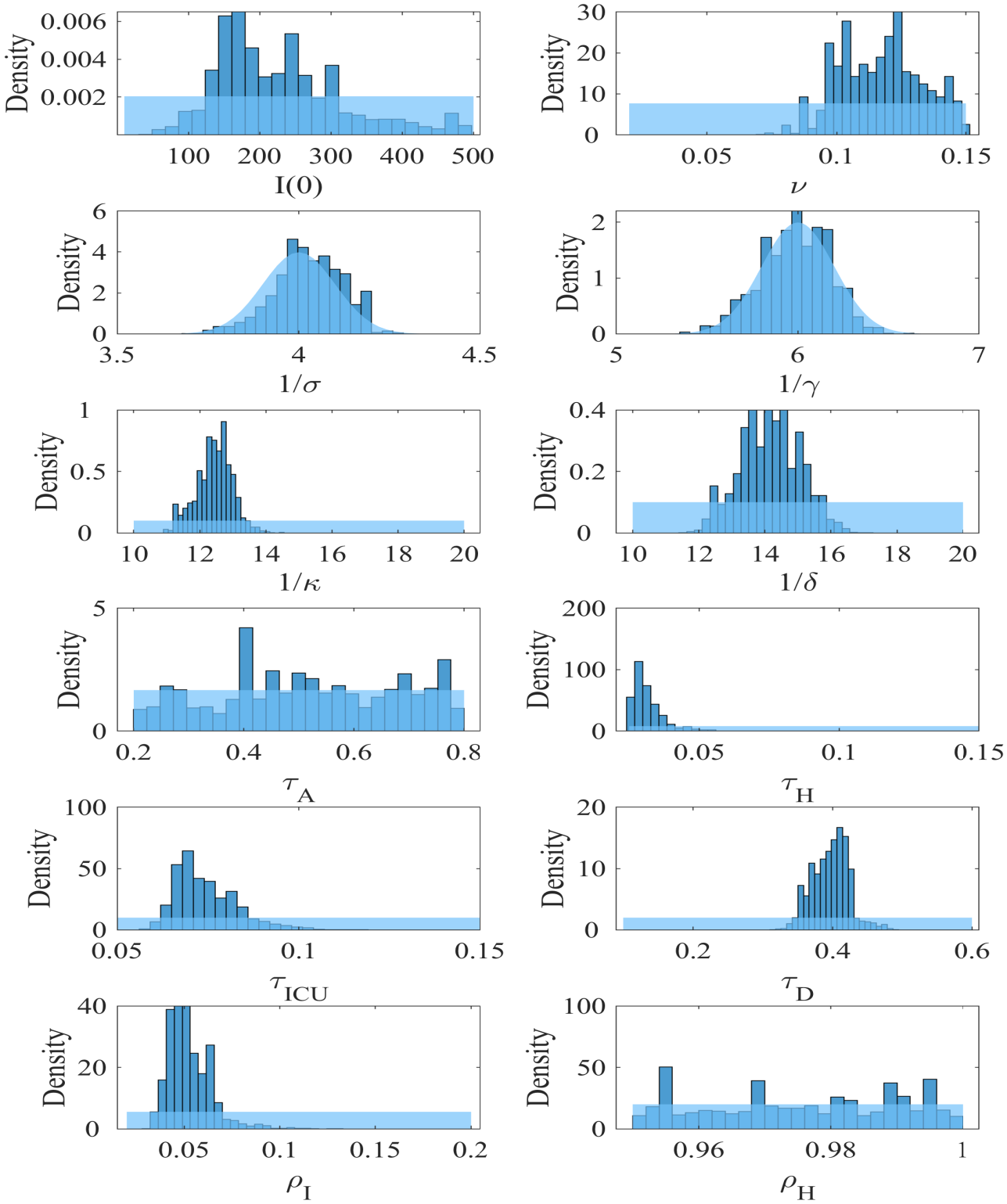

Fig. S11 – Prior and posterior distributions for the model inferences presented Fig. S9. Caption as Fig. S1.

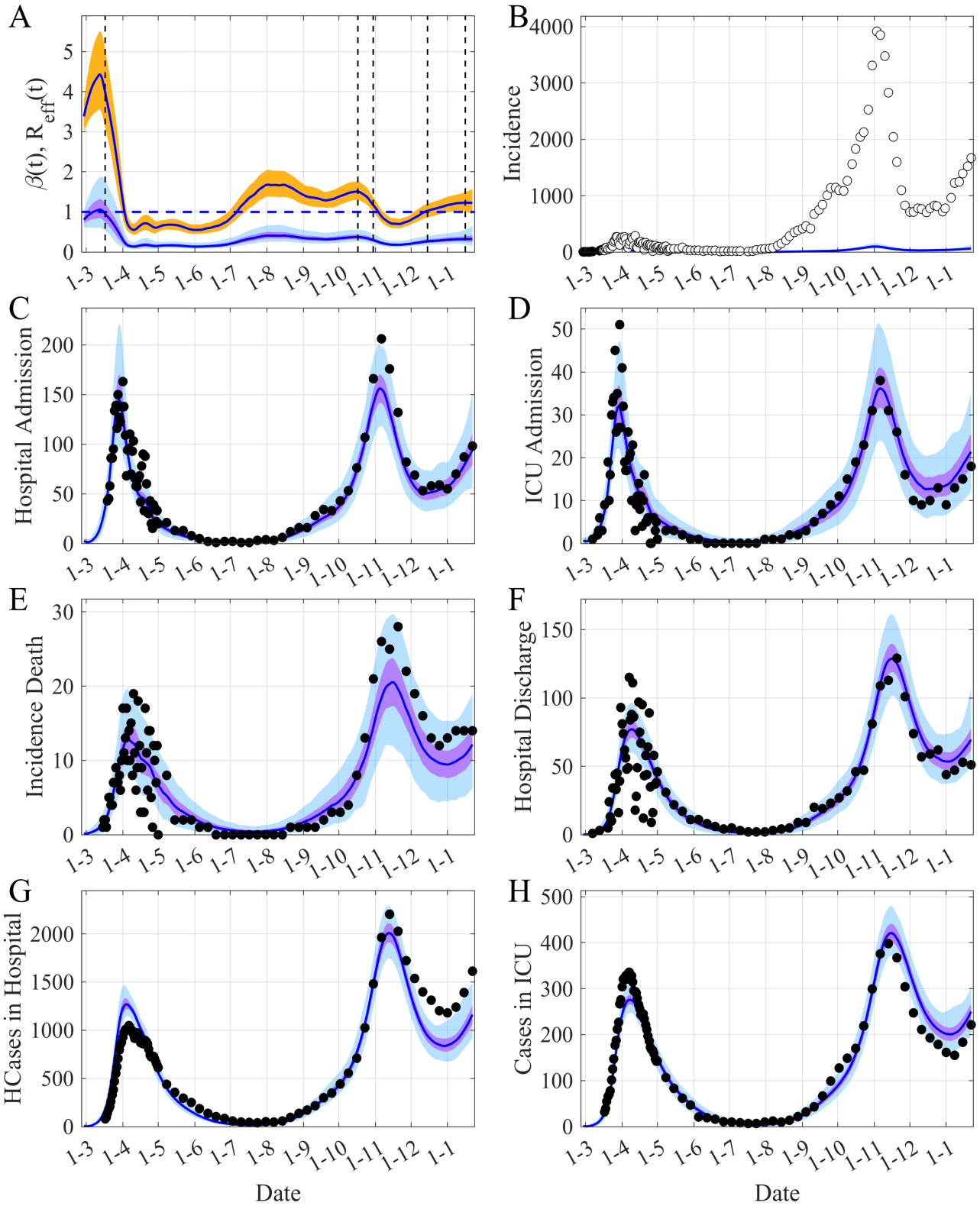

Fig. S12 – Reconstruction of the observed dynamics of COVID-19 in **Occitanie**. Caption as Fig. 2. The black points are observations used by the inference process, the white points are the observations not used.

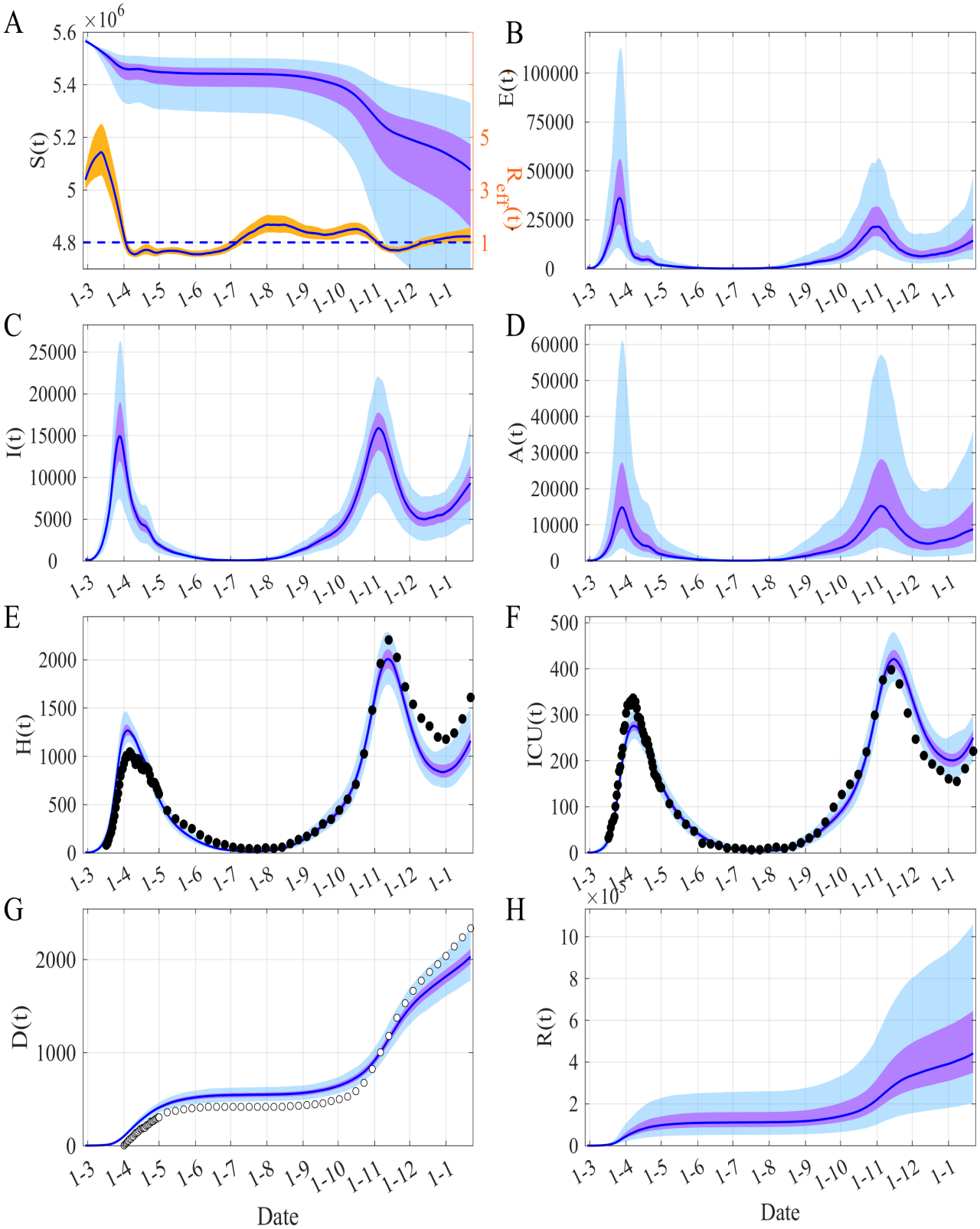

Fig. S13 – Dynamics of Covid-19 in **Occitanie**. Caption as Fig. 2. The black points are observations used by the inference process, the white points are the observations not used.

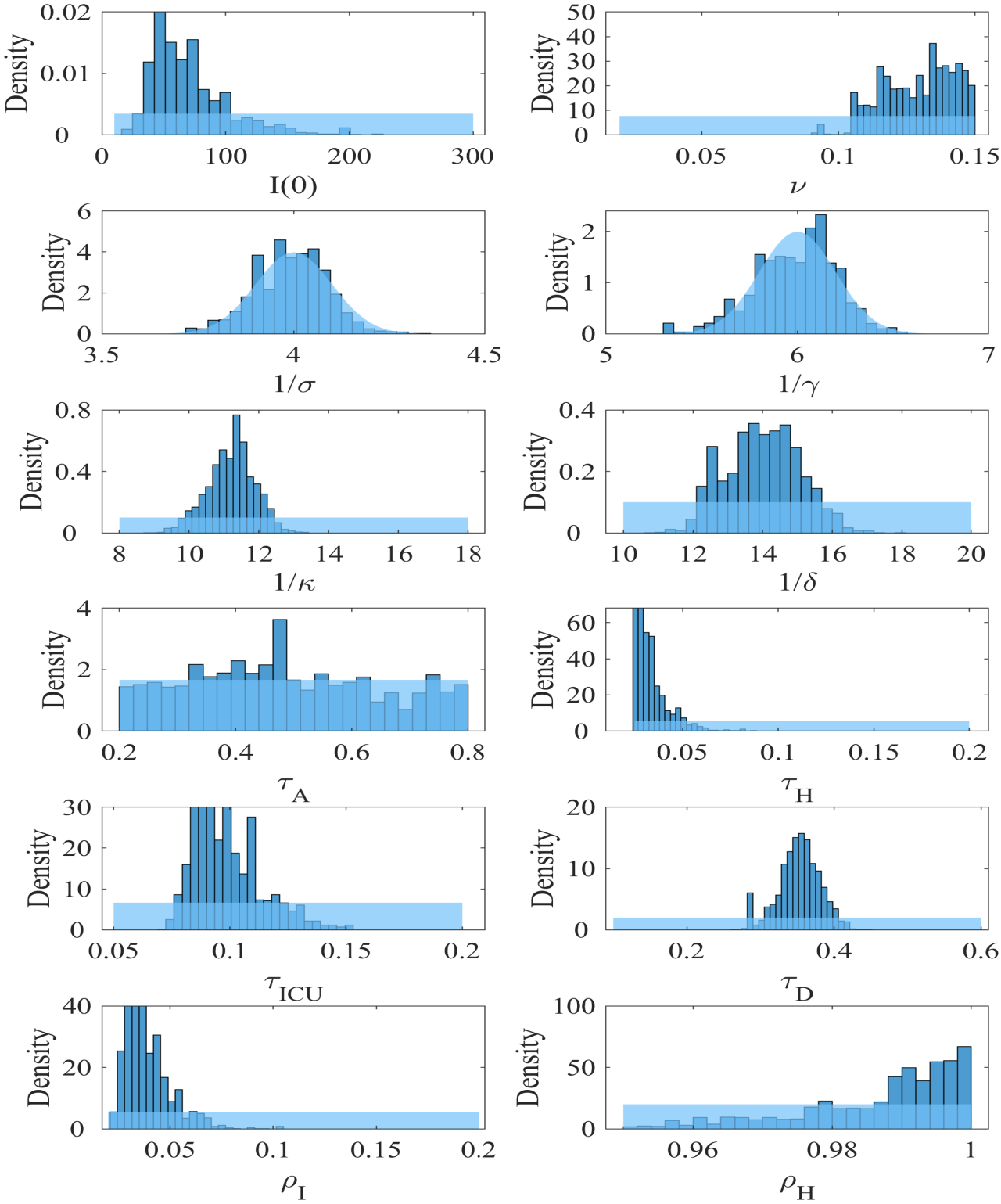

Fig. S14 – Prior and posterior distributions for the model inferences presented Fig. S12. Caption as Fig. S1.

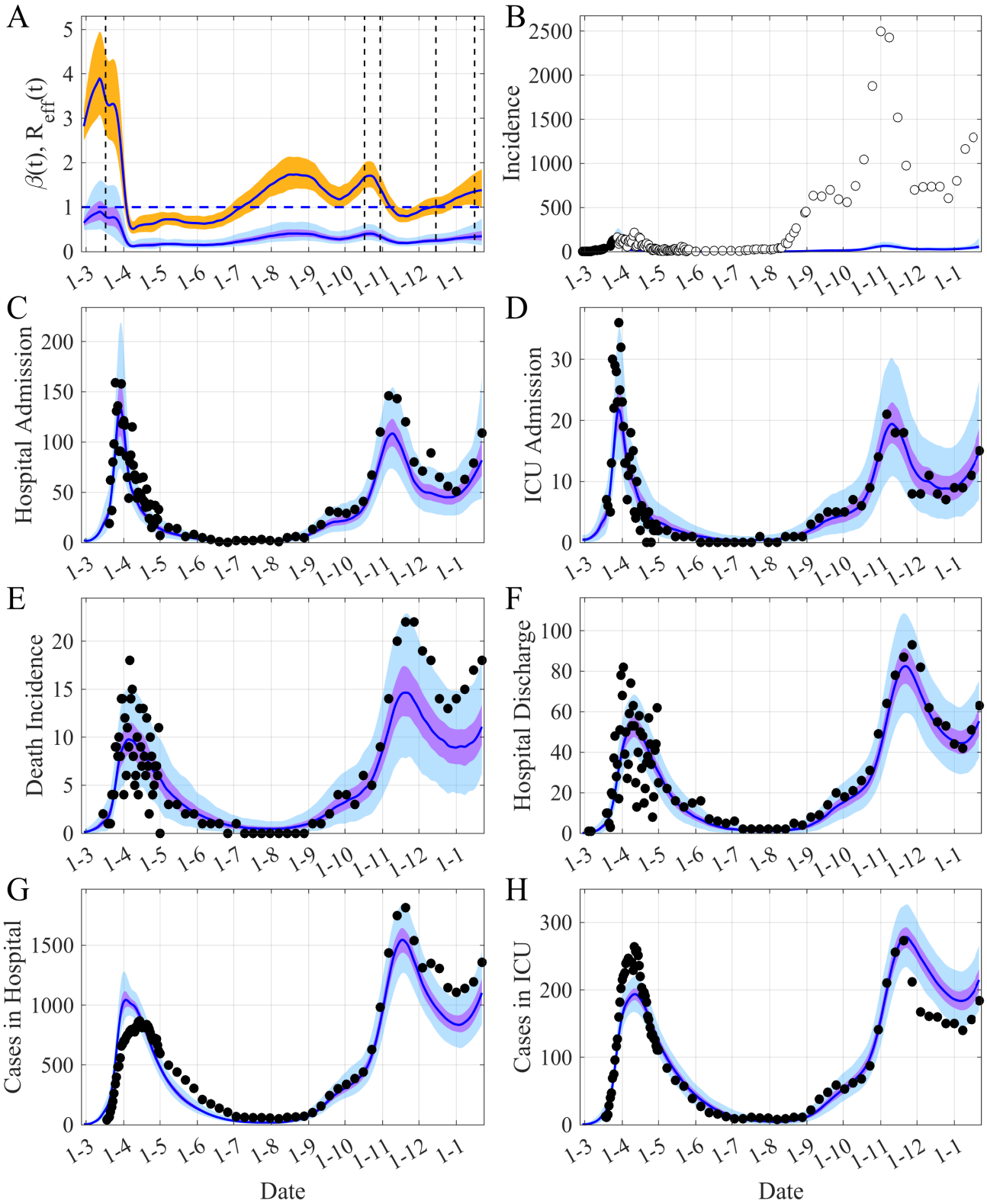

Fig. S15 – Reconstruction of the observed dynamics of COVID-19 in **Nouvelle Aquitaine**. Caption as Fig. 2. The black points are observations used by the inference prAQUIess, the white points are the observations not used.

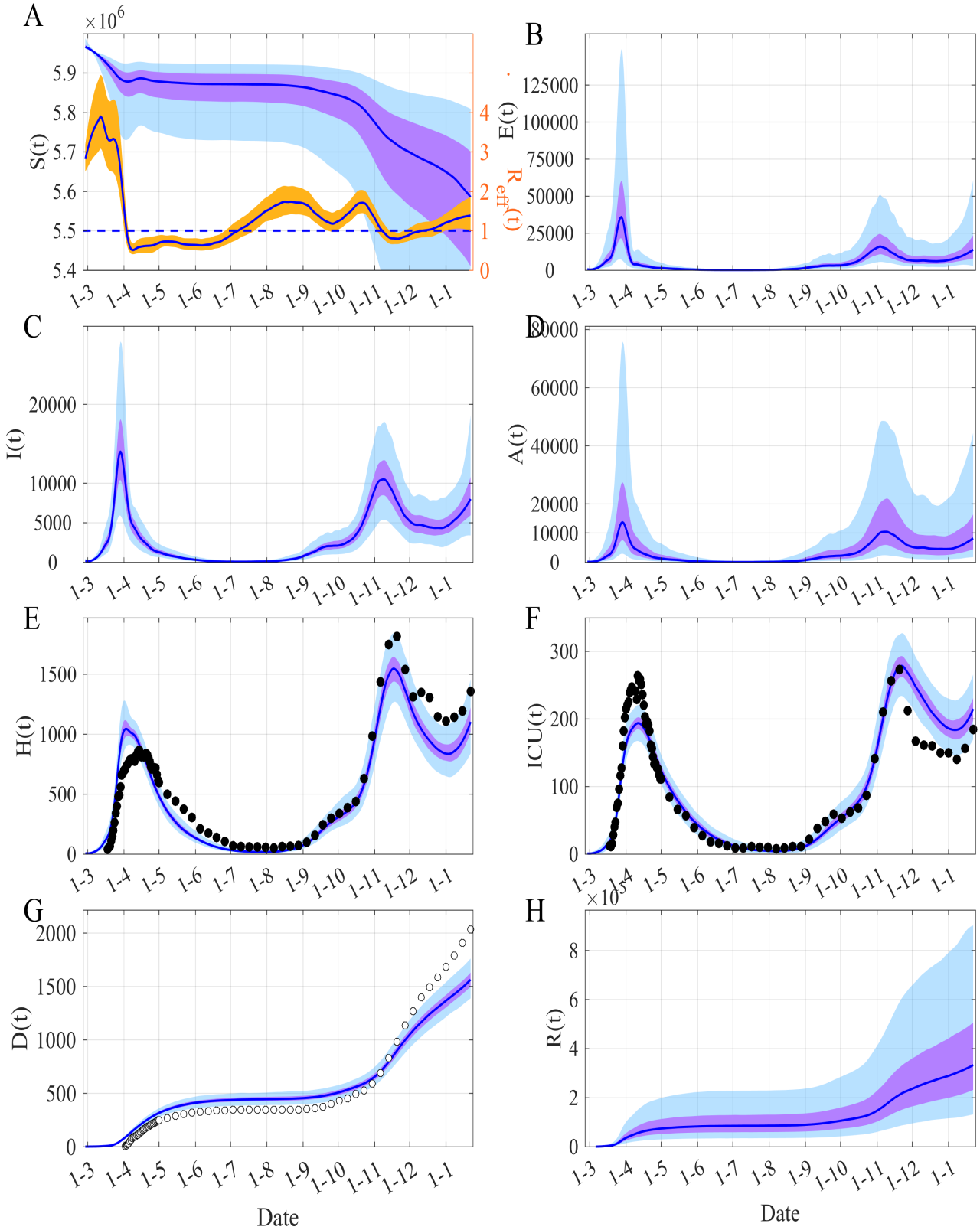

Fig. S16 – Dynamics of COVID-19 in **Nouvelle Aquitaine**. Caption as Fig. 2. The black points are observations used by the inference process, the white points are the observations not used.

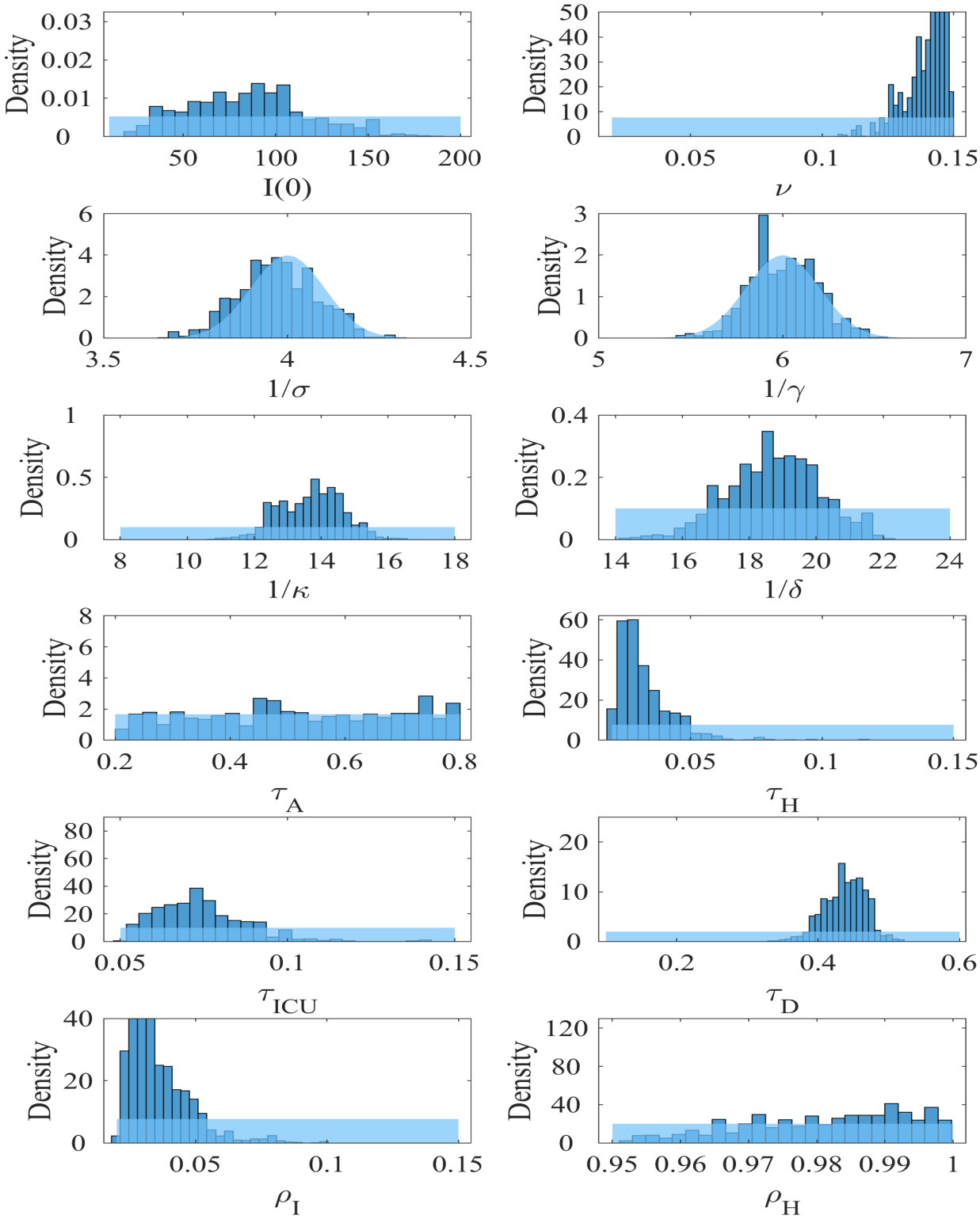

Fig. S17 – Prior and posterior distributions for the model inferences presented Fig. S15. Caption as Fig. S1.

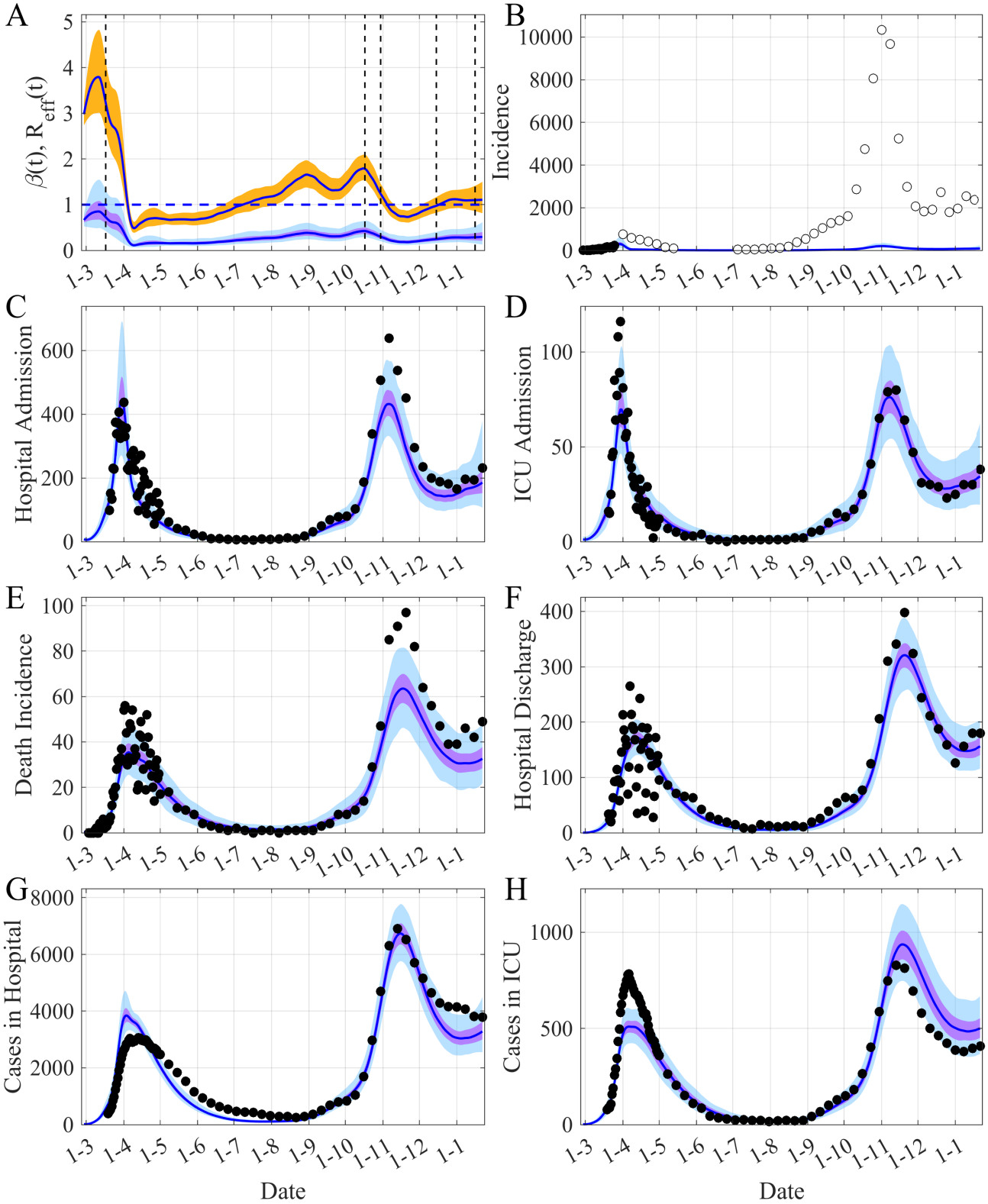

Fig. S18 – Reconstruction of the observed dynamics of COVID-19 in **Auvergne Rhône Alpes**. Caption as Fig. 2. The black points are observations used by the inference prARAess, the white points are the observations not used.

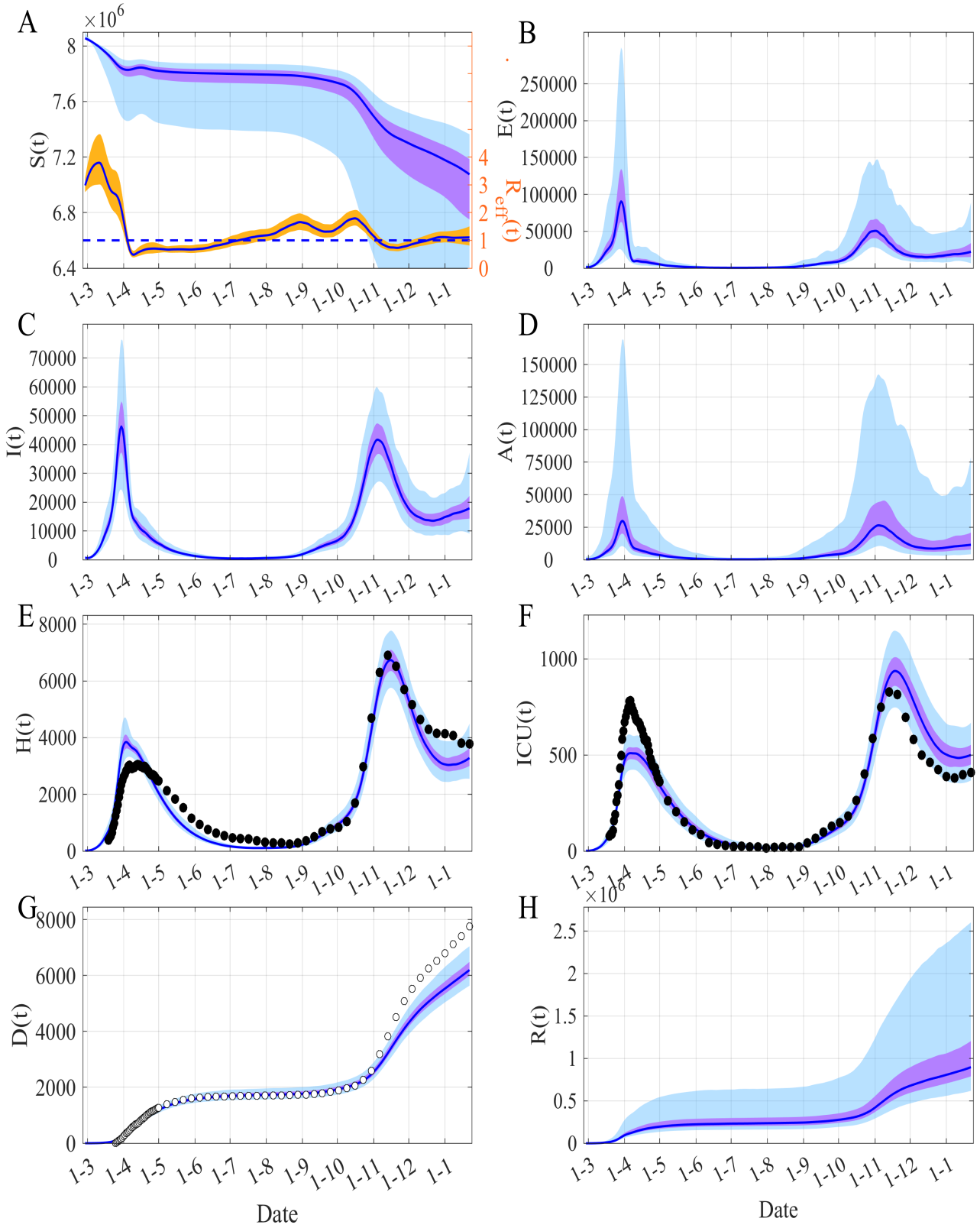

Fig. S19 – Dynamics of COVID-19 in **Auvergne Rhône Alpes**. Caption as Fig. 2. The black points are observations used by the inference prARAess, the white points are the observations not used.

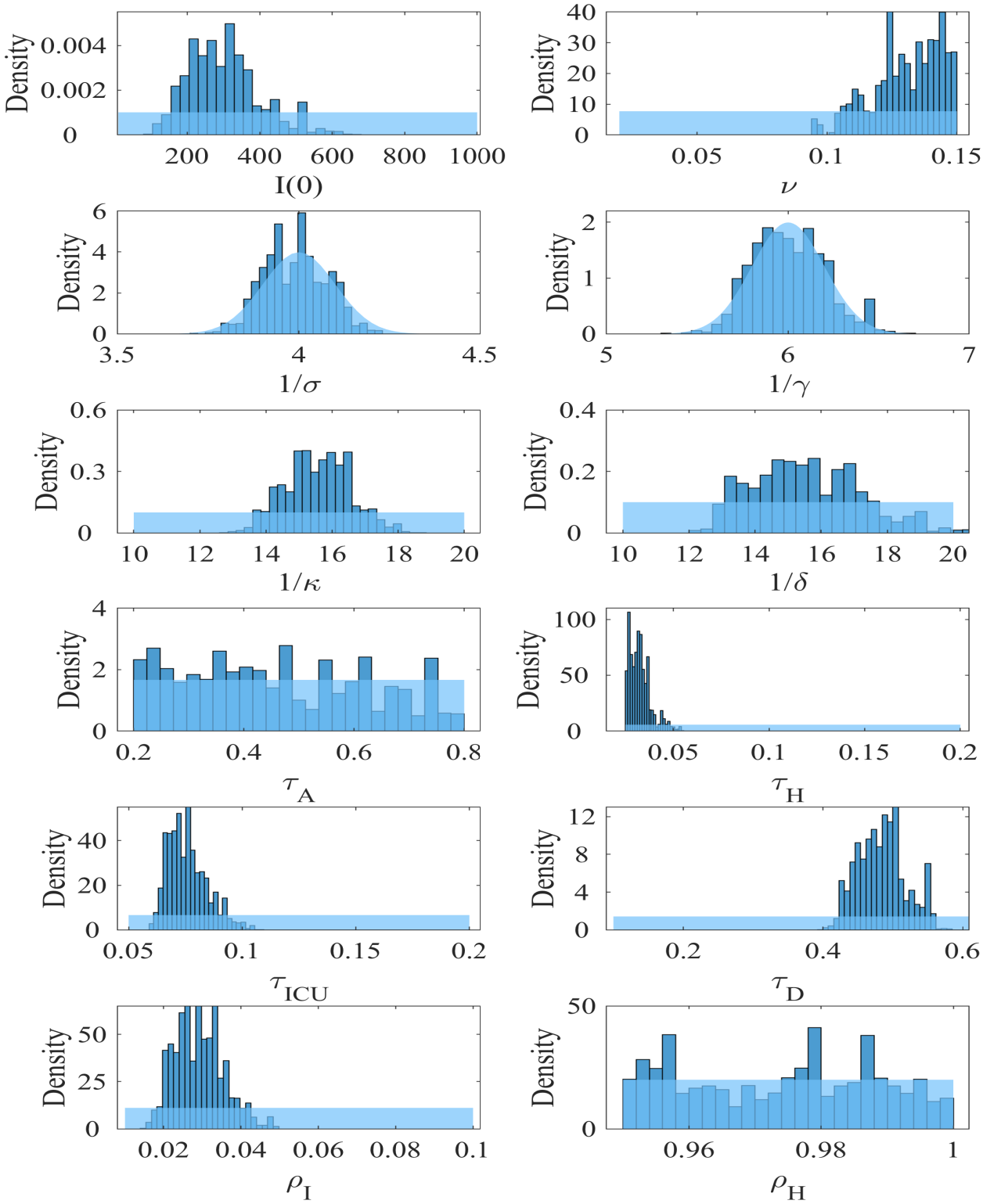

Fig. S20 – Prior and posterior distributions for the model inferences presented Fig. S18. Caption as Fig. S1.

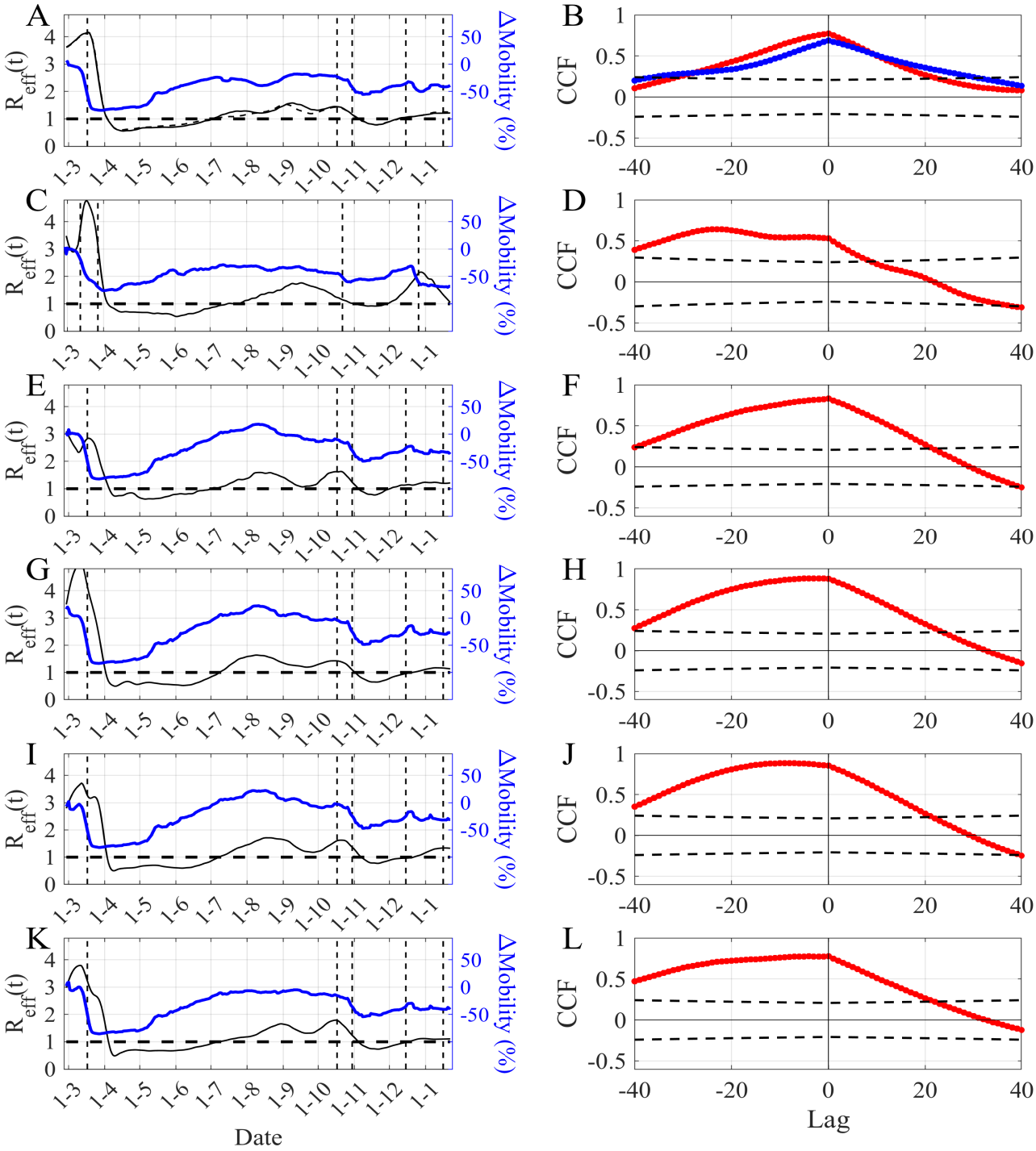

Fig. S21 – Correlations between the effective reproduction number and public transport mobility (https ://www.google.com/covid19/mobility/). Left column : Time evolution of the estimated $R_{eff}(t)$ (black line) and public transport mobility (blue line). Right column : Cross-correlation functions between the estimated $R_{eff}(t)$ and public transport mobility computed between 15-05-2020 and 15-10-2020. The dashed black lines delimit the significant region at 0.1%. (A-B) Ile-de-France region, (C-D) Ireland ; (E-F) Provence Alpes Côte d'Azur region, (F-G) Occitanie region, (H-I) Nouvelle-Aquitaine region, (J-K) Auvergne Rhône Alpes region.