Socio-hydrologic data assimilation: Analyzing human-flood interactions by model-data integration

Yohei Sawada[1], Risa Hanazaki[2]

[1] Institute of Engineering Innovation, School of Engineering, the University of Tokyo, Tokyo, Japan

[2] Institute of Industrial Science, the University of Tokyo, Tokyo, Japan

Corresponding author: Y. Sawada, Institute of Engineering Innovation, the University of Tokyo, Tokyo, Japan, 2-11-6, Yayoi, Bunkyo-ku, Tokyo, Japan, yohei.sawada@sogo.t.u-tokyo.ac.jp

Abstract

In socio-hydrology, human-water interactions are simulated by mathematical models. Although the integration of these socio-hydrologic models and observation data is necessary to improve the understanding of the human-water interactions, the methodological development of the model-data integration in socio-hydrology is in its infancy. Here we propose to apply sequential data assimilation, which has been widely used in geoscience, to a socio-hydrological model. We developed particle filtering for a widely adopted flood risk model and performed an idealized observation system simulation experiment and a real-data experiment to demonstrate the potential of the sequential data assimilation in socio-hydrology. In these experiments, the flood risk model's parameters, the input forcing data, and empirical social data were assumed to be somewhat imperfect. We tested if data assimilation can contribute to accurately reconstructing the historical human-flood interactions by integrating these imperfect models and imperfect and sparsely distributed data. Our results highlight that it is important to sequentially constrain both state variables and parameters when the input forcing is uncertain. Our proposed method can accurately estimate the model's unknown parameters even if the true model parameter temporally varies. The small amount of empirical data can significantly improve the simulation skill of the flood risk model.

Therefore, sequential data assimilation is useful to reconstruct historical socio-hydrological processes by the synergistic effect of models and data.

36

37 **1. Introduction**

38 Socio-hydrology is an emerging research field in which two-way feedbacks between
39 social and water systems are investigated (Sivapalan et al. 2012, 2014). Understanding
40 complex socio-hydrologic phenomena contributes to solving water crises around the
41 world. Socio-hydrology has been recognized as an important scientific grand challenge
42 to meet United Nations' Sustainable Development Goals (Di Baldassarre et al. 2019).

43

44 The most popular approach in socio-hydrology is to develop dynamic models which
45 compute non-linear interactions between human and water. For instance, Di Baldassarre
46 et al. (2013) developed a simplified model, which described human-flood interactions, to
47 understand the levee effect in which high levees generate a false sense of security and
48 induce social vulnerabilities to severe floods (see also Viglione et al. 2014; Ciullo et al.
49 2017). Van Emmerik et al. (2014) developed a stylized model, which described two-way
50 feedbacks between environment and economic activities, to understand the historical
51 competition for water between agricultural development and environment health in
52 Australia (see also Roobavannan et al. 2017). Pande and Savenije (2016) modeled
53 economic activities of smallholder farmers to analyze the agrarian crisis in Marathwada,
54 India. While socio-hydrologic models described above assumed the existence of a single

55 lumped decision maker, Yu et al. (2017) incorporated a collective action into their model
56 and analyzed the dynamics of community-managed flood protection systems in coastal
57 Bangladesh. Please refer to Di Baldassarre et al. (2019) for the comprehensive review of
58 socio-hydrologic modeling.
59
60 In addition to these modeling approaches, both qualitative and quantitative data related to
61 socio-hydrologic processes are important to understand human-water interactions. For
62 instance, Mostert (2018) revealed historical changes in river management from water
63 resources development to protection and restoration by analyzing qualitative data. Dang
64 and Konar (2018) applied econometric methods to analyze quantitative data in both
65 human and water domains and quantified the causal relationship between trade openness
66 and water use. Kreibich et al. (2017) performed the detailed case study analysis on paired
67 floods, consecutive flood events which occurred in the same region with the second flood
68 causing significantly lower damage. They found that the reduction of vulnerability played
69 a key role for successful adaptation to the second floods.
70
71 Although it is expected that the integration of model and data contributes to accurately
72 understanding the socio-hydrologic processes (Mount et al. 2016), the methodological

73 development of the model-data integration in socio-hydrology is in its infancy. Generally,
74 mathematical models can provide spatiotemporally continuous state variables and
75 quantitative scenarios for future socio-hydrologic developments. In addition,
76 mathematical models can quantitatively provide possible scenarios unrealized in the real-
77 world, which gives the insight to targeted processes (e.g., Viglione et al. 2014). The major
78 limitation of socio-hydrological models is that they are often inaccurate due to the
79 uncertainty in their input forcing, parameters, and descriptions of the processes. On the
80 other hand, hydrologic and social data are often more reliable than numerical models and
81 can provide more complete understanding of the socio-hydrological processes (e.g.,
82 Mostert 2018), although data also have uncertainties. However, in many cases, relevant
83 data in socio-hydrology are sparsely distributed so that it is difficult to completely
84 reconstruct the historical socio-hydrologic processes from data. The other limitation of
85 the data-driven approach is that the quantification of the causal relationship cannot be
86 easily done only by empirical data (e.g., Dang and Konar 2018). Considering this
87 advantages and disadvantages of model and data, previous studies used social statistics
88 to calibrate and validate their socio-hydrologic models (e.g., Barendrecht et al. 2019;
89 Roobavannan et al. 2017; Ciullo et al. 2017; van Emmerik et al. 2014; Gonzales and
90 Ajami 2017).

In geosciences, sequential data assimilation has been widely used for the model-data integration. Data assimilation sequentially adjusts the predicted state variables and parameters of dynamic models by integrating observation data into models based on Bayes' theorem. Data assimilation has been widely applied to numerical weather prediction (e.g., Miyoshi and Yamane 2007; Bauer et al. 2015; Poterjoy et al. 2019; Sawada et al. 2019), atmospheric reanalysis (e.g., Kobayashi et al. 2015; Hersbach et al. 2019), and hydrology and land surface modeling (e.g., Moradkhani et al. 2005; Sawada et al. 2015; Rasmussen et al. 2015; Lievens et al. 2017). Applicability of the data assimilation approach to the socio-hydrologic models has yet to be investigated.

In this study, we aim to develop the methodology of sequential data assimilation for the flood risk model proposed by Di Baldassarre et al. (2013). From a series of idealized experiments and a real-data experiment in the city of Rome, we demonstrate the potential of data assimilation to accurately reconstruct the historical human-flood interactions. We focus on the case in which the socio-hydrologic model's parameters, input forcing data, and social data are somewhat inaccurate.

2. Method

2.1. Model

In this study, we used a socio-hydrologic flood risk model proposed by Di Baldassarre et al. (2013). This model conceptualizes human-flood interactions by the set of simple equations which describe the states of flood, economy, technology, politics, and society. Based on this original model of Di Baldassarre et al. (2013), many similar flood risk models have been proposed, validated, and applied (e.g., Viglione et al. 2014; Ciullo et al. 2017; Barendrecht et al. 2019). Here we briefly describe this model. Please refer to Di Baldassarre et al. (2013) for the complete description of this model.

The governing equations of the flood risk model are shown below:

$$F = \begin{cases} 1 - \exp\left(-\frac{W+\xi_H H}{\alpha_H D}\right) & if\ W + \xi_H H > H \\ 0 & if\ W + \xi_H H \leq H \end{cases} \quad (1)$$

$$R = \begin{cases} \varepsilon_T (W + \xi_H H - H) & if\ (F > 0)\ and\ \left(FG > \gamma_E R\sqrt{G}\right)\ and\ (G - FG > \gamma_E R\sqrt{G}) \\ 0 & otherwise \end{cases}$$

(2)

$$S = \begin{cases} \alpha_S F & if\ (R > 0) \\ F & if\ (R = 0) \end{cases} \quad (3)$$

$$\frac{dG}{dt} = \rho_E \left(1 - \frac{D}{\lambda_E}\right) G - \Delta(\Upsilon(t))(FG + \gamma_E R\sqrt{G}) \quad (4)$$

126 $$\frac{dD}{dt} = (M - \frac{D}{\lambda_P})\frac{\varphi_P}{\sqrt{G}} \quad (5)$$

127 $$\frac{dH}{dt} = \Delta\big(\Upsilon(t)\big)R - \kappa_T H \quad (6)$$

128 $$\frac{dM}{dt} = \Delta\big(\Upsilon(t)\big)S - \mu_S M \quad (7)$$

129

130 This model has four state variables: G, D, H, and M. G(t) [L^2] is the size of the human
131 settlement; D(t) [L] is the distance of the center of mass of the human settlement from the
132 river; H(t) [L] is the flood protection level (or levee height); M(t) [.] is the social
133 awareness of the flood risk. The timestep was set to annual.

134

135 Equation (1) calculates the intensity of flooding events F(t) [.] from the high water level
136 W(t) [L], the height of the levee H(t) [L], and the distance of the human settlement from
137 the river D(t) [L]. Equation (2) calculates R(t) [L], the amount by which the levees are
138 raised responding to the flood event. There are three required conditions under which
139 people decide to raise the levee. First, the flood event occurs. Second, the damage of flood
140 (FG) should be larger than the cost of raising levee. Third, the cost of raising levee should
141 be lower than the wealth remaining after the flooding. Equation (3) shows the magnitude
142 of the psychological shock by the flood event S(t) [.]. If the levee is raised, the
143 psychological shock is assumed to be mitigated. Equation (4) explains the dynamics of

144 G(t), the size of the human settlement or the wealth of the community. Following the
145 notation of Di Baldassarre et al. (2013), $\Delta(\Upsilon(t)) = 1$ with integral only when time t
146 passes the time of the flooding event (F>0), otherwise, $\Delta(\Upsilon(t)) = 0$. The term $FG +$
147 $\gamma_E R\sqrt{G}$ (total cost of flood damage and construction of levees) appears only if flood
148 occurs. Equation (5) shows the dynamics of the distance of the center of mass of the
149 human settlement from the river D(t). When the social awareness of the flood risk is high,
150 people tend to live far from the river. Equation (6) computes the dynamics of the flood
151 protection level H(t) and equation (7) shows the dynamics of the social awareness of the
152 flood risk M(t). The explanation of parameters can be found in Table 1.
153
154
155 **2.2. Data Assimilation**
156 In this study, we used Sampling Importance Resampling Particle Filtering (SIRPF) as the
157 method of data assimilation. SIRPF has been widely used in hydrologic data assimilation
158 (e.g., Moradkhani et al. 2005; Qin et al. 2009; Sawada et al. 2015). Compared with the
159 other data assimilation algorithms such as ensemble Kalman filter, SIRPF is robust
160 against model nonlinearity and associated non-Gaussian error distribution. The
161 disadvantage of SIRPF is that the infeasible computational resources are required if the

numerical model is computationally expensive, which is not the case in the flood risk model.

The flood risk model can be formulated as a discrete state-space dynamic system:

$$\boldsymbol{x}(t+1) = f\big(\boldsymbol{x}(t), \boldsymbol{\theta}, \boldsymbol{u}(t)\big) + \boldsymbol{q}(t) \quad (8)$$

where $\boldsymbol{x}(t)$ is the state variables (i.e. G, D, H, and M), $\boldsymbol{\theta}$ is the model parameters, $\boldsymbol{u}(t)$ is the external forcing (i.e., the high water level), and $\boldsymbol{q}(t)$ is the noise process which represents the model error. In data assimilation, it is useful to formulate an observation process as follows:

$$\boldsymbol{y}^f(t) = h\big(\boldsymbol{x}(t)\big) + \boldsymbol{r}(t) \quad (9)$$

where $\boldsymbol{y}^f(t)$ is the simulated observation, h is the observation operator which maps the model's state variables into the observable variables, and $\boldsymbol{r}(t)$ is the noise process which represents the observation error.

The SIRPF is a Monte Carlo approximation of Bayesian update of the state variables and parameters:

$$p(\boldsymbol{x}(t), \boldsymbol{\theta} | \boldsymbol{y}^o(1:t)) \propto p(\boldsymbol{y}^o(t) | \boldsymbol{x}(t), \boldsymbol{\theta}) p(\boldsymbol{x}(t), \boldsymbol{\theta} | \boldsymbol{y}^o(1:t-1)) \quad (10)$$

where $p(\boldsymbol{x}(t), \boldsymbol{\theta}|\boldsymbol{y}^o(1:t))$ is the posterior probability of the state variables $\boldsymbol{x}(t)$ and parameters $\boldsymbol{\theta}$ given all observations up to time t $\boldsymbol{y}^o(1:t)$. The prior knowledge, $p(\boldsymbol{x}(t), \boldsymbol{\theta}|\boldsymbol{y}^o(1:t-1))$, based on the model integration is updated using the likelihood which includes the new observation at time t $p(\boldsymbol{y}^o(t)|\boldsymbol{x}(t), \boldsymbol{\theta})$. In this study, we assumed that our observation error follows Gaussian distribution so that the likelihood can be formulated as follows:

$$p(\boldsymbol{y}^o(t)|\boldsymbol{x}(t), \boldsymbol{\theta}) \equiv L(\boldsymbol{y}^o(t), \boldsymbol{x}(t), \boldsymbol{\theta}) =$$

$$\frac{1}{\sqrt{\det(2\pi \boldsymbol{R})}} \exp\left[-\frac{1}{2}\left(\boldsymbol{y}^o(t) - \boldsymbol{y}^f(t)\right)^T \boldsymbol{R}^{-1}\left(\boldsymbol{y}^o(t) - \boldsymbol{y}^f(t)\right)\right] \quad (11)$$

where **R** is the covariance matrix of the observation error process $\boldsymbol{r}(t)$. The prior knowledge of the state variables is approximated by the ensemble simulation:

$$p(\boldsymbol{x}(t)|\boldsymbol{y}^o(1:t-1)) \approx \frac{1}{N}\sum_{i=1}^{N} \delta\left[\boldsymbol{x}(t) - f\left(\boldsymbol{x}^i(t-1), \boldsymbol{\theta}^i, \boldsymbol{u}^i(t-1)\right)\right] \quad (12)$$

where N is the ensemble size, $\boldsymbol{x}^i, \boldsymbol{\theta}^i, \boldsymbol{u}^i$ are the realizations of the ensemble member i, and $\delta[.]$ is the Direc delta function.

The posterior probability of the state variables and parameters can be approximated as follows:

$$p(\boldsymbol{x}(t)|\boldsymbol{y}^o(1:t)) \approx \sum_{i=1}^{N} w(i)\delta(\boldsymbol{x}(t) - \boldsymbol{x}^i(t)) \quad (13)$$

$$p(\boldsymbol{\theta}|\boldsymbol{y}^o(1:t)) \approx \sum_{i=1}^{N} w(i)\delta(\boldsymbol{\theta} - \boldsymbol{\theta}^i) \quad (14)$$

where $w(i)$ is the normalized weight for the realization of the ensemble member i and is calculated using the likelihood (see also equation (11)).

$$w(i) = \frac{L(\boldsymbol{y}^o(t),\boldsymbol{x}^i(t),\theta^i)}{\sum_{k=1}^{N} L(\boldsymbol{y}^o(t),\boldsymbol{x}^k(t),\theta^k)} \quad (15)$$

Note that equations (13) and (14) update all state variables and parameters of the model although the weight is calculated using only observable variables. Therefore, it is not necessary to observe all state variables in order to update all system variables.

The implementation of SIRPF is the following:

1. Model state variables are updated from time t-1 to t using ensemble simulation (equations (8) and (12)).
2. Simulated observations are calculated for all ensembles (equation (9)).
3. The likelihood for each ensemble member is calculated (equation (11))
4. The weights are obtained for all ensembles (equation (15))
5. We applied a resampling procedure according to the normalized weights. The normalized weights of ensemble i, $w(i)$, can be recognized as the probability that the ensemble i is selected after resampling. Resampled state variables and parameters are defined as $\boldsymbol{x}^i_{resamp}$ and $\boldsymbol{\theta}^i_{resamp}$, respectively.

6. Since there are no mechanisms to increase the variance of parameters of ensemble members, Moradkhani et al. (2005) proposed to perturb the ensembles of parameters:

$$\boldsymbol{\theta}^i \leftarrow \boldsymbol{\theta}^i_{resamp} + \varepsilon^i \quad (16)$$

$$\varepsilon^i \sim N(0, \max(\boldsymbol{\omega}, s \times Var^{\theta}) \quad (17)$$

where $N(.)$ is the Gaussian distribution, Var^{θ} is the variance of $\boldsymbol{\theta}^i$, $\boldsymbol{\omega}$ is the fixed hyperparameter (see Table 1 for its variable) which guarantees that the ensembles of parameters do not converge into a single value. s is an adaptively changed factor according to the effective ensemble size, N_{eff}.

$$s = s_0\left(1 - \left(\frac{N_{eff}}{N}\right)^2\right) \quad (18)$$

$$N_{eff} = \frac{1}{\sum_{i=1}^{N} w(i)} \quad (19)$$

where $s_0 = 0.05$. The effective ensemble size is the measure of the diversity of ensembles. If the effective ensemble size becomes small, ensembles should be strongly perturbed in order to maintain the diversity of ensembles. Similar strategy has been used in many SIRPF systems (e.g., Moradkhani et al. 2005; Poterjoy et al. 2019).

3. Experiment design

3.1. Observation System Simulation Experiment

In this study, we performed three observation system simulation experiments (OSSEs). In the OSSE, we generated the synthetic truth of the state and flux variables by driving the flood risk model with the specified parameters and input. Then, we generated synthetic observations by adding the noise to this synthetic truth. Those synthetic observations were assimilated into the model by SIRPF. The performance of SIRPF was evaluated by comparing the estimated state variables by SIRPF with the synthetic truth. Model parameters used to generate the synthetic truth can be found in Table 1. They are identical to Di Baldassarre et al. (2013). The OSSE has been recognized as an important preliminary step to verify the newly developed data assimilation systems (e.g., Moradkhani et al. 2005; Vrugt et al. 2013; Penny and Miyoshi 2016; Sawada et al. 2018).

The high water level for the synthetic truth was generated by the following:

$$W = \min(v - 10, 0) \tag{20}$$

v follows the Gumbel distribution:

$$p(v) = \frac{\exp\left(-\frac{v-\mu}{\beta}\right)}{\beta} \exp\left(-\exp(-(v-\mu)\beta)\right) \tag{21}$$

where $\mu = 9, \beta = 2.5$. Although our high water level is not identical to Di Baldassarre et al. (2013), the estimated trajectory of the state variables is similar to Di Baldassarre et al. (2013).

Synthetic observations were generated by adding the Gaussian white noise to the F, G, D, H, and M (see section 2.1) of the synthetic truth. The mean of the Gaussian white noise was 0. The observation error, the standard deviation of the Gaussian white noise, was firstly set to 10% of the synthetic true variables. Although this observation error is generally larger than that used in meteorology and hydrology, we further increased the observation error and tested the sensitivity of the observation error to the SIRPF's performance. We firstly assumed that all of the F, G, D, H, and M can be observed every 10 years or every 10 model integration steps. Then, we evaluated the sensitivity of the observation network (i.e. the observable variables and the observation intervals) to the SIRPF's performance. Although it is not straightforward to observe social memory M, several previous studies obtained the proxy of the social memory by interview data (Barendrecht et al. 2019) and the number of Google searches (Gonzales and Ajami 2017).

We used the ensemble mean of root-mean square errors (mRMSE) as an evaluation metrics:

$$RMSE^i = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(x^i(t) - z(t))} \quad (22)$$

$$mRMSE = \frac{1}{N}\sum_{i=1}^{N} RMSE^i \quad (23)$$

where $RMSE^i$ is root-mean-square-error for i th ensemble, T is the computational period, $x^i(t)$ is the simulated state variables of ensemble i at time t, $z(t)$ is the synthetic truth at time t.

3.1.1. Experiment 1: Perfect model with uncertain high water levels

In the first OSSE, we assumed that there is no uncertainty in model parameters. We used the same parameter variables as the synthetic truth run and we did not perform the estimation of parameters. Our SIRPF updated only state variables. Although the model had no uncertainty, it was assumed that the input data, the timeseries of the high water level, were uncertain. Lognormal multiplicative noise was added to the synthetic true high water level so that different ensemble members have different high water levels in the data assimilation experiment. The two parameters of the lognormal distribution, commonly called μ and σ, were set to 0 and 0.15, respectively.

284

285

286 **3.1.2. Experiment 2: Unknown model parameters and uncertain high water levels**

287 In the second OSSE, we assumed that some of the synthetic true parameter values were
288 unknown. The unknown parameters in the experiment 2 were the cost of levee raising γ_E,
289 the rate by which new properties can be built φ_P, the rate of decay of levees κ_T, and
290 memory loss rate μ_S (see Table 1). We selected these unknown parameters one by one
291 from four equations of economy, politics, technology, and social to discuss how each state
292 variable's observation affects the estimation of parameters across these four equations
293 (see section 2.1). We have no unknown parameters related to F (equation (1)) since it is
294 unlikely that the parameters in equation (1) are much more inaccurate than the other
295 parameters. The parameters related to flood are mainly determined by the topography of
296 the flood plain so that the process described in equation (1) can be replaced by more
297 accurate hydrodynamic models in the real-world case study. The initial parameter
298 variables were assumed to be distributed in the bounded uniform distributions whose
299 ranges were found in Table 1. The uncertainty of the simulation induced by these
300 parameters' uncertainty is large enough to demonstrate the potential of data assimilation
301 to minimize the simulation's uncertainty (see Results). Our SIRPF sequentially

assimilated observations and estimated both state variables and parameters in the experiment 2. The high water level data were uncertain as the experiment 1.

3.1.3. Experiment 3: Unknown and time-variant model parameters and uncertain high water levels

To further demonstrate the potential of sequential data assimilation in socio-hydrology, we assumed that the description of the model was biased in the experiment 3. Here we assumed that two of the model parameters were temporally varied by the unknown dynamics. Specifically, the rate by which new properties can be built, φ_P, and the memory loss rate, μ_S, were temporally varied in the experiment 3:

$$\varphi_P(t) = \begin{cases} 5000\ (t < 250) \\ 5000 + (t - 250) \times \frac{40000-5000}{500}\ (250 \le t < 750) \\ 40000\ (750 \le t) \end{cases} \quad (24)$$

$$\mu_S(t) = \begin{cases} 0.01\ (t < 250) \\ 0.01 + (t - 250) \times \frac{0.10-0.01}{500}\ (250 \le t < 750) \\ 0.10\ (750 \le t) \end{cases} \quad (25)$$

In the data assimilation experiment, we assumed that the dynamics of φ_P and μ_S was unknown, and we integrated the flood risk model with time-invariant φ_P and μ_S. We evaluated if SIRPF could track this time-variant parameter and reveal the bias of the model's description. The cost of levee raising γ_E, and the rate of decay of levees κ_T

were assumed to be time-invariant unknown parameters as they were in the experiment 2. The cost of levee raising γ_E affects the state variables of the flood risk model mainly in the initial early years and the gradual change of the rate of decay of levees κ_T has few impacts on the state variables. Therefore, we found that it is difficult to track the temporal change of these two parameters. The input forcing data, high water level, were uncertain as described in the experiment 1.

3.2. Real-data experiment

In addition to the OSSEs, we performed the real-world experiment in the city of Rome, Italy. Ciullo et al. (2017) collected real-world data and calibrated their flood risk model. Using the data collected by Ciullo et al. (2017), we performed the data assimilation experiment. It should be noted that the flood risk model of Ciullo et al. (2017) is different from our model (i.e. Di Baldassarre et al. 2013), although they are conceptually similar.

All the data were collected from Figure 1 of Ciullo et al. (2017) by WebPlotDigitizer (https://automeris.io/WebPlotDigitizer/). The observed high water level of Tiber River was used as input forcing data (W). The levee height (H) and population (G) were used

337 as the observation data to be assimilated into the flood risk model. In Ciullo et al. (2017),
338 population values within the Tiber's floodplain were normalized by the theoretical
339 maximum Tiber's floodplain population which is estimated to the range between 10^6
340 and 2×10^6. Since our flood risk model needs the population values (not normalized
341 values), we multiplied 1.5×10^6 and the normalized values shown in Figure 1 of Ciullo
342 et al. (2017) to obtain population in the floodplain.
343
344 We added lognormal multiplicative noise to the observed high water level as we did in
345 the OSSEs. The observation errors of levee height and population were set to 10% and
346 25% of the observed values, respectively. Since Ciullo et al. (2017) showed the large
347 uncertainty in the estimation of the theoretical maximum population (see above), it is
348 reasonable to assume that the estimation of population values also has relatively large
349 uncertainty.
350
351 As the second and third OSSEs, we have 4 unknown parameters in this real-world
352 experiment. We used the same settings of parameters as the OSSEs, which are shown in
353 Table 1, except for ξ_H, proportion of additional high water level due to levee heightening.
354 In this real-world experiment, we set $\xi_H = 0$ because the observed high water level

355 includes the effects of levee heightening. This treatment is consistent to Ciullo et al.
356 (2017) (see their Table 2).
357
358 The initial conditions of H and M were set to 0. The initial conditions of D were obtained
359 from the uniform distribution between 1000 and 5000. The initial conditions of G were
360 obtained from the uniform distribution between 1500 and 50000.
361
362
363 **4. Results**
364 **4.1. Observation System Simulation Experiment**
365 **4.1.1. Experiment 1: Perfect model with uncertain high water levels**
366 Figure 1 shows the timeseries of the model variables calculated by 5000 ensembles with
367 no data assimilation. Although the ensemble mean of the state variables is close to the
368 synthetic truth, the ensembles have the large spread especially for G. The uncertainty in
369 the input forcing brings the uncertainty in the estimation of the historical socio-hydrologic
370 condition.
371

372 Figure 2 indicates that this uncertainty is mitigated by assimilating the observations of F,
373 G, D, H, and M into the model every 10 years with 5000 ensembles. Table 2 shows that
374 RMSE is reduced for all state variables by data assimilation.
375
376 While we can observe all of F, G, D, H, and M in Figure 2 and Table 2, Figure 3 shows
377 the performance of our SIRPF in which only one of them can be observed. Our SIRPF
378 updates all state variables although only one of them is assimilated. Figure 3 reveals that
379 we can accurately propagate the observation information into the model state space. In
380 other words, our SIRPF can positively impact the estimation of not only observed state
381 variables but unobserved state variables. For instance, even if we can observe only G, the
382 simulation of all G, D, H, and M is improved. This finding is promising since all of the
383 state variables cannot be observed in the real-world applications. Figure 3 also shows that
384 observing F is not effective compared with the other variables. This is because F is a flux
385 and F can be observed only when floods occur so that the number of effective
386 observations is small. In addition, observing F, D, and M negatively impacts the
387 estimation of H and observing H does not significantly improve the simulation of D and
388 M. Although the dynamics of F, D, and M strongly affects the decision making of whether
389 the levees are raised or not, the amount by which the levees are raised, R, is fully

390 determined by the high water level, W, once the community determines to raise the levees
391 (see equation (2)). Therefore, the uncertainty of H is largely induced by the uncertainty
392 of the high water level, W, whose uncertainty is not directly mitigated by our SIRPF. This
393 is why observing F, D, and M is not helpful to mitigate the uncertainty of H.
394
395 While we can observe every 10 years in Figure 2 and Table 2, Figure 4 shows the
396 sensitivity of the observation intervals to the performance of our SIRPF. Our SIRPF
397 improves the estimation of the state variables when we can obtain observation once in
398 50-year or 100-year (see also Figure S1 for timeseries of the model's variables), which is
399 promising since we cannot expect the frequent observations in the real-world applications.
400
401 We set the observation error to 10% of the synthetic truth thus far. The improvement of
402 the simulation skill can be found with larger observation errors (Figure S2). Although the
403 SIRPF's performance gradually declines as the observation error increases, our SIRPF
404 can significantly improve the simulation skill with 25% observation error.
405
406 Although we demonstrate the potential of our SIRPF with 5000 ensembles thus far, the
407 improvement of the simulation skill can be found in much smaller ensemble sizes. The

408 performance of our SIRPF with 20 ensembles is similar to that with 5000 ensembles
409 (Figure S3).
410
411
412 **4.1.2. Experiment 2: Unknown model parameters and uncertain high water levels**
413 Figure 5 reveals that the flood risk model completely loses its skill to estimate the human-
414 flood interactions if there are uncertainties in model parameters and high water levels
415 prescribed in Section 3. In contrast to the experiment 1, the ensemble mean cannot
416 accurately reproduce the synthetic truth.
417
418 Figure 6 indicates that our SIRPF can accurately estimate the model state variables by
419 assimilating the observations of F, G, D, H, and M into the model every 10 years with
420 5000 ensembles. Figure 7 indicates that four unknown parameters can also be accurately
421 estimated. We find that it is relatively difficult to estimate the rate of levee's decay, κ_T,
422 compared with the other parameters. This is because κ_T strongly affects the dynamics
423 of H and the uncertainty in H is largely determined by the uncertainty in high water levels,
424 which is not directly mitigated by our SIRPF system. Table 3 shows that RMSE is reduced
425 for both state variables and parameters by data assimilation.

426

427 We analyzed the impacts of the individual observation types on the simulation skill as we
428 did in the experiment 1. Figure 8a shows that the effects of the individual observation
429 types are similar to what we found in the experiment 1: (1) our SIRPF can improve the
430 skill to simulate unobservable state variables; (2) observing F is not effective compared
431 with the other observations; (3) observing H does not significantly improve the simulation
432 of D and M. Figure 8b reveals that the parameters can be efficiently estimated by
433 assimilating the observation of the state variables which are tightly related to the targeted
434 parameters. For instance, observing D can greatly improve the rate by which new
435 properties can be built, φ_P, in equation (5) which governs the dynamics of D. However,
436 assimilating a single observation type can contribute to accurately estimating all four
437 parameters in many cases, which is the promising result considering the sparsity of the
438 observation in the real-world applications.

439

440 The good performance of our SIRPF can be found with the longer observation intervals
441 as we found in the experiment 1. Figure 9 indicates that our SIRPF can improve the
442 estimation of the state variables and parameters when we can obtain observation once in
443 50-year or 100-year (see also Figures S4 and S5 for timeseries of the model's variables).

As we found in the experiment 1, the SIRPF's performance declines with the increased observation error (Figure S6). However, it is promising that our SIRPF can improve the simulation skill with larger observation errors up to 25% of the synthetic truth considering that the observations in the socio-hydrologic domain are often inaccurate.

In contrast to the experiment 1, the larger ensemble size is required to stably estimate both state variables and parameters (Figure S7). The increased degree of freedom and the nonlinear relationship between parameters and observations increase the necessary ensemble size.

4.1.3. Experiment 3: Unknown and time-variant model parameters and uncertain high water levels

In addition to the experiment 2, two of the unknown parameters (φ_P and μ_S) temporally vary in the synthetic truth of the experiment 3. We found that a larger spread of φ_P is required to stably track the time-variant synthetic true φ_P so that we increased s_0 in equation (18) from 0.05 to 0.5 only for φ_P in this experiment 3. Figure 10 and Table 4

462 indicate that despite the error in the model's description, our SIRPF can greatly improve
463 the simulation of the flood risk model. Please note that the synthetic truth shown in Figure
464 10 is different from that of the previous experiments especially for D and M. Figures 11b
465 and 11d indicate that we can accurately estimate the time-variant parameters (φ_P and
466 μ_S) as well as the other time-invariant parameters (Figures 11a and 11c). This result is
467 promising since we cannot expect the perfect description of the socio-hydrologic model
468 in the real-world applications. We also performed the sensitivity test on observation types,
469 observation intervals, and ensemble sizes, which results in the same conclusions as the
470 experiment 2 (not shown).
471
472
473 **4.2. Real-data experiment**
474 Figure 12 shows the timeseries of the model variables calculated by 5000 ensembles with
475 no data assimilation. The 5000-ensemble simulation reveals the two bifurcated social
476 systems. One builds a high levee and maintains a course of stable economic growth. The
477 other one has no levee and its economy is damaged by severe floods many times
478 (ensemble mean shown in Figure 12b implies that there are many ensemble members with
479 zero levee height).

480

481 In reality, the city of Rome constructed the levee responding to the severe flood occurred
482 on 28 December 1870. After the construction of this levee, no major flood losses occurred,
483 allowing the steady and undisturbed growth. Figure 13 indicates that our SIRPF
484 successfully constrains the trajectory of the ensemble simulation to the real-world (i.e.
485 high levee and stable economic growth) by assimilating the real data of H and G. Figure
486 S8 shows the SIRPF-estimated unknown parameters. Our SIRPF suggests lower γ_E than
487 the initial ensemble mean to promote the levee construction with lower costs. Lower κ_T
488 is also obtained because the assimilated real data show no decay of levee from 1874 to
489 2009. Compared with the OSSE experiment 2, the large uncertainty in estimated
490 parameters remains at the final timestep due to the limited number of assimilated
491 observations. In contrast to the OSSEs, our observation network has the uneven temporal
492 distribution. Figure 13 clearly indicates that our SIRPF is robust to these intermittent
493 observations whose intervals temporally change.

494

495 We analyzed the impacts of the individual observation types (i.e. H and G) on the
496 simulation skill as we did in the OSSEs. Figure 14 indicates that our SIRPF realistically
497 simulates the socio-hydrologic dynamics in the city of Rome and provides the similar

498 estimated state variables shown in Figure 13 by assimilating only population data. As we
499 found in the OSSEs, observations of the size of the human settlement G are informative
500 to effectively constrain the flood risk model. The dynamics of the parameter estimation
501 is similar to the case in which data of both G and H are assimilated (Figure S9).
502
503 On the other hand, assimilating only levee height data cannot provide the similar results
504 to those shown above. Figure 15 shows the timeseries of the model variables by the data
505 assimilation experiment in which we assimilated the observation data of H only.
506 Observations of the levee height cannot effectively constrain D, G, and M compared with
507 the observations of G. This finding is consistent to the OSSEs. The uncertainty in
508 estimated parameters becomes larger when we omit to assimilate observations of G
509 (Figure S10). Although the impact of levee height data is limited compared with
510 population data, it is promising that we can estimate the socio-hydrologic dynamics to
511 some extent only from the levee height data whose distribution is temporally sparse.
512
513
514 **5. Discussion and Conclusions**

515 In this study, we developed the sequential data assimilation system for the widely adopted
516 socio-hydrological model, the flood risk model by Di Baldassarre et al. (2013). We
517 demonstrated that our SIRPF for the flood risk model is useful to reconstruct the historical
518 human-flood interactions, which can be called "socio-hydrologic reanalysis", by
519 integrating sparsely distributed observations and imperfect numerical simulation. In the
520 atmospheric science, atmospheric reanalysis has been intensively analyzed to understand
521 complex feedback in the atmosphere, which cannot be done by analyzing only
522 observation data due to their sparsity. Socio-hydrologic reanalysis can work as a reliable
523 and spatio-temporally homogeneous dataset and may be helpful to deepen the
524 understanding of human and water. In addition, socio-hydrologic reanalysis can be used
525 as initial condition to predict the future change of socio-hydrologic processes as
526 atmospheric scientists predict the future weather/climate using atmospheric reanalysis.
527 Since it is impossible to directly observe all state variables and parameters as initial
528 conditions, socio-hydrologic reanalysis is crucially important for accurate prediction.
529 Socio-hydrologic data assimilation has a high potential to improve the understanding of
530 the complex feedback between social and flood systems and predict their future. Our
531 idealized OSSE and real-data experiment reveal several important findings.
532

First, the sequential data assimilation can mitigate the negative impact of the uncertainty in the input forcing on the simulation of socio-hydrologic state variables. We found that the small perturbation of high water levels greatly affects the long-term trajectory of the socio-hydrologic state variables as Viglione et al. (2014) found. It is necessary to sequentially constrain the state variables and parameters by sequential data assimilation if the input forcing is uncertain although previous studies on the model-data integration in socio-hydrology mainly focused on parameter calibration assuming no uncertainty in the input forcing (e.g., Barendrecht et al. 2019; Roobavannan et al. 2017; Ciullo et al. 2017; van Emmerik et al. 2014; Gonzales and Ajami 2017). To deeply understand the socio-hydrologic processes, the long-term historical analysis should be performed. Although there are many studies on the accurate reconstruction of the historical weather condition (e.g., Toride et al. 2017), it may be necessary to tackle with the uncertainty in hydrometeorological datasets used for the input forcing of the socio-hydrologic models.

Second, our SIRPF can efficiently improve the simulation of the socio-hydrologic state variables using the sparsely distributed data. All model variables should not necessarily be observed to constrain the model's state variables and parameters. In some cases, observations of a single state variable are enough to reconstruct the accurate socio-

551 hydrologic state. In addition, observation intervals can be longer than 10-year. Since it is
552 difficult to obtain the large volume of data in socio-hydrology, this finding is promising.
553 We also give some insights about the informative observation types in the flood risk
554 model. With uncertain high water levels, observations of the intensity of flooding events
555 F and the height of levee H are not informative (i.e. the assimilation of these observations
556 cannot greatly improve the simulation skill) although the empirical data which can be
557 related to F and H may be easily found. On the other hand, observations of the size of the
558 human settlement G are informative to constrain the flood risk model. Model parameters
559 can be efficiently estimated by assimilating the state variables which is tightly related to
560 the targeted parameters, which is consistent to the findings of the idealized experiment by
561 Barendrecht et al. (2019).
562
563 Third, our SIRPF is robust to the imperfectness of the socio-hydrologic model. The
564 unknown parameters can be efficiently estimated by the sequential data assimilation.
565 While previous studies evaluated the trajectory in the whole study period to calibrate the
566 socio-hydrologic models by iteratively performing the long-term model integration (e.g.,
567 Barendrecht et al. 2019; Roobavannan et al. 2017; Ciullo et al. 2017; van Emmerik et al.
568 2014; Gonzales and Ajami 2017), we sequentially optimize parameters based on the

569 relatively short-term timeseries allowing parameters to temporally vary in the study
570 period. The advantage of this strategy is that we can deal with time-variant parameters as
571 previously demonstrated in the applications to hydrologic models (e.g., Pathiraja et al.
572 2018). In the model development, parameters are formulated as time-invariant values so
573 that the existence of time-variant parameters indicates the imperfect description of
574 dynamic models. Sequential data assimilation can mitigate the negative impact of this
575 imperfect model description. Vrugt et al. (2013) pointed out that the parameter
576 optimization by the sequential filters is unstable if parameter sensitivity temporally
577 changes (e.g., parameters affects the model's dynamics differently in the different
578 seasons), which may be the potential limitation of our strategy compared with Bayesian
579 inference based on the long-term trajectory such as Barendrecht et al. (2019).
580
581 The major limitation of this study is that we assume the modeled state variables can
582 directly be observed although it is difficult to directly observe state variables of the socio-
583 hydrologic models. For example, it is impossible to directly observe social awareness of
584 flood risk in the flood risk model and several previous studies obtained the proxy of the
585 social memory by interview data (Barendrecht et al. 2019) and the number of Google
586 searches (Gonzales and Ajami 2017). When these indirect observations are assimilated

into a model, the (non-linear) observation operator (see equation (9)), the assignment of the observation error, and assimilation methods should be carefully designed as previously discussed in the context of numerical weather prediction (e.g., Sawada et al. 2019; Okamoto et al. 2019; Minamide and Zhang 2017). Future work will focus on the methodological development to efficiently assimilate observations in the social domain with complicated structure of observation operators and errors.

Acknowledgements

We thank Di Baldassarre for sharing the original source code of the flood risk model. We thank two anonymous referees for their constructive comments. Data Integration and Analysis System (DIAS) provided us the computational resources.

Code/Data availability

Code and data are available upon the request to the corresponding author.

Author Contribution

YS designed the study. RH and YS jointly developed the data assimilation system for the flood risk model and performed the numerical experiments. YS and RH contributed to interpreting the results. YS wrote the first draft of the paper and RH contributed to editing the paper.

Competing interests

The authors declare that they have no conflict of interest.

References

Barendrecht, M. H., Viglione, A., Kreibich, H., Merz, B., Vorogushyn, S., and Blöschl, G.: The Value of Empirical Data for Estimating the Parameters of a Sociohydrological Flood Risk Model. *Water Resources Research*. https://doi.org/10.1029/2018WR024128, 2019

Bauer, P., Thorpe, A., and Brunet, G.: The quiet revolution of numerical weather prediction. *Nature*, *525*(7567), 47–55. https://doi.org/10.1038/nature14956, 2015

Ciullo, A., Viglione, A., Castellarin, A., Crisci, M., and Di Baldassarre, G.: Socio-hydrological modelling of flood-risk dynamics: comparing the resilience of green

621 and technological systems. *Hydrological Sciences Journal*, *62*(6), 880–891.
622 https://doi.org/10.1080/02626667.2016.1273527, 2017
623 Dang, Q., and Konar, M.: Trade Openness and Domestic Water Use. *Water Resources*
624 *Research*, *54*(1), 4–18. https://doi.org/10.1002/2017WR021102, 2018
625 Di Baldassarre, G., Viglione, A., Carr, G., Kuil, L., Salinas, J. L., and Blöschl, G.: Socio-
626 hydrology: Conceptualising human-flood interactions. *Hydrology and Earth*
627 *System Sciences*, *17*(8), 3295–3303. https://doi.org/10.5194/hess-17-3295-2013,
628 2013
629 Di Baldassarre, G., et al.: Socio-hydrology: Scientific Challenges in Addressing a Societal
630 Grand Challenge. *Water Resources Research*, 1–29.
631 https://doi.org/10.1029/2018wr023901, 2019
632 Gonzales, P., and Ajami, N.: Social and Structural Patterns of Drought-Related Water
633 Conservation and Rebound. *Water Resources Research*, *53*(12), 10619–10634.
634 https://doi.org/10.1002/2017WR021852, 2017
635 Hersbach, H. et al.: Global reanalysis: goodbye ERA-Interim, hello ERA5, *ECMWF*
636 *Newsletter, 159,* 17-24, doi: 10.21957/vf291hehd7, 2019

637 Kobayashi, S., et al.: The JRA-55 Reanalysis: General Specifications and Basic
638 Characteristics. *Journal of the Meteorological Society of Japan, 93,* 5-48.
639 https://doi.org/10.2151/jmsj.2015-001, 2015
640 Kreibich, H., et al.: Adaptation to flood risk: Results of international paired flood event
641 studies. *Earth's Future*. https://doi.org/10.1002/eft2.232, 2017
642 Lievens, H., et al.: Joint Sentinel-1 and SMAP data assimilation to improve soil moisture
643 estimates. *Geophysical Research Letters*, *44*(12), 6145–6153.
644 https://doi.org/10.1002/2017GL073904, 2017
645 Minamide, M., and Zhang, F: Adaptive Observation Error Inflation for Assimilating All-
646 Sky Satellite Radiance., *Monthly Weather. Review*, *145,* 1063–1081,
647 https://doi.org/10.1175/MWR-D-16-0257.1, 2017
648 Miyoshi, T., and Yamane, S.: Local Ensemble Transform Kalman Filtering with an
649 AGCM at a T159/L48 Resolution. *Monthly Weather Review*, *135*(2002), 3841–
650 3861. https://doi.org/10.1175/2007MWR1873.1, 2007
651 Moradkhani, H., Hsu, K. L., Gupta, H., and Sorooshian, S.: Uncertainty assessment of
652 hydrologic model states and parameters: Sequential data assimilation using the
653 particle filter. *Water Resources Research*, *41*(5), 1–17.
654 https://doi.org/10.1029/2004WR003604, 2005

Mostert, E.: An alternative approach for socio-hydrology: Case study research. *Hydrology and Earth System Sciences*, *22*(1), 317–329. https://doi.org/10.5194/hess-22-317-2018, 2018

Mount, N., J., et al.: Data-driven modelling approaches for sociohydrology: opportunities and challenges within the Panta Rhei Science Plan. *Hydrological Sciences Journal, 61(7),* 1192-1208. https://doi.org/10.1080/02626667.2016.1159683, 2016

Okamoto, K, Sawada, Y, Kunii, M. Comparison of assimilating all-sky and clear-sky infrared radiances from Himawari-8 in a mesoscale system. *Q J R Meteorol Soc.*, *145*, 745-766. https://doi.org/10.1002/qj.3463, 2019

Pande, S., and Savenije, H. H. G.: A sociohydrological model for smallholder farmers in Maharashtra, India. *Water Resources Research*, *52*(3), 1923–1947. https://doi.org/10.1002/2015WR017841, 2016

Pathiraja, S., Anghileri, D., Burlando, P., Sharma, A., Marshall, L., and Moradkhani, H.: Time-varying parameter models for catchments with land use change: the importance of model structure, Hydrol. Earth Syst. Sci., *22*, 2903–2919, https://doi.org/10.5194/hess-22-2903-2018, 2018.

Penny, S. G., and Miyoshi, T.: *A local particle filter for high-dimensional geophysical systems*. 391–405. https://doi.org/10.5194/npg-23-391-2016, 2016

673 Poterjoy, J., Wicker, L., and Buehner, M.: Progress toward the application of a localized
674 particle filter for numerical weather prediction. *Monthly Weather Review*, *147*(4),
675 1107–1126. https://doi.org/10.1175/MWR-D-17-0344.1, 2019
676 Qin, J., Liang, S., Yang, K., Kaihotsu, I., Liu, R., and Koike, T.: Simultaneous estimation
677 of both soil moisture and model parameters using particle filtering method through
678 the assimilation of microwave signal. *Journal of Geophysical Research*, *114*(D15),
679 1–13. https://doi.org/10.1029/2008JD011358, 2009
680 Rasmussen, J., Madsen, H., Jensen, K. H., and Refsgaard, J. C.: Data assimilation in
681 integrated hydrological modeling using ensemble Kalman filtering: evaluating the
682 effect of ensemble size and localization on filter performance. *Hydrology and Earth*
683 *System Sciences*, *19*(7), 2999–3013. https://doi.org/10.5194/hess-19-2999-2015,
684 2015
685 Roobavannan, M., Kandasamy, J., Pande, S., Vigneswaran, S., and Sivapalan, M.: Role
686 of Sectoral Transformation in the Evolution of Water Management Norms in
687 Agricultural Catchments: A Sociohydrologic Modeling Analysis. *Water Resources*
688 *Research*, *53*(10), 8344–8365. https://doi.org/10.1002/2017WR020671, 2017

689 Sawada, Y., Koike, T., and Walker, J. P.: A land data assimilation system for simultaneous
690 simulation of soil moisture and vegetation dynamics. *J. Geophys. Res. Atmos.*, *120*,
691 5910– 5930. doi: 10.1002/2014JD022895, 2015
692 Sawada, Y., Nakaegawa, T. and Miyoshi, T.: Hydrometeorology as an inversion problem:
693 Can river discharge observations improve the atmosphere by ensemble data
694 assimilation? *Journal of Geophysical Research: Atmospheres*, 123, 848– 860.
695 https://doi.org/10.1002/2017JD027531, 2018
696 Sawada, Y., Okamoto, K., Kunii, M., and Miyoshi, T.: Assimilating every-10-minute
697 Himawari-8 infrared radiances to improve convective predictability. *Journal of*
698 *Geophysical Research: Atmospheres*, 124, 2546–2561.
699 https://doi.org/10.1029/2018JD029643, 2019
700 Sivapalan, M., Savenije, H.H.G. and Blöschl, G.: Socio-hydrology: A new science of
701 people and water. Hydrol. Process., 26: 1270-1276. doi:10.1002/hyp.8426, 2012
702 Sivapalan, M., Konar, M., Srinivasan, V., Chhatre, A., Wutich, A., Scott, C. A., and
703 Wescoat, J. L.: Socio-hydrology: Use-inspired water sustainability science for the
704 Anthropocene, *Earth's Future*, *2*, 225–230. https://doi.org/10.1002/2013EF000164,
705 2014.

706 Toride, K., Neluwala, P., Kim, H. and Yoshimura, K.: Feasibility Study of the
707 Reconstruction of Historical Weather with Data Assimilation. *Mon. Wea. Rev.,* **145**,
708 3563–3580, https://doi.org/10.1175/MWR-D-16-0288.1, 2017
709 Van Emmerik, T. H. M., et al.: Socio-hydrologic modeling to understand and mediate the
710 competition for water between agriculture development and environmental health:
711 Murrumbidgee River basin, Australia. Hydrology and Earth System Sciences,
712 18(10), 4239–4259. https://doi.org/10.5194/hess-18-4239-2014, 2014
713 Viglione, A., et al.: Insights from socio-hydrology modelling on dealing with flood risk -
714 Roles of collective memory, risk-taking attitude and trust. *Journal of Hydrology*,
715 *518*(PA), 71–82. https://doi.org/10.1016/j.jhydrol.2014.01.018, 2014
716 Vrugt, J. A., ter Braak, C. J. F., Diks, C. G. H., and Schoups, G.: Hydrologic data
717 assimilation using particle Markov chain Monte Carlo simulation: Theory, concepts
718 and applications. *Advances in Water Resources*, *51*, 457–478.
719 https://doi.org/10.1016/j.advwatres.2012.04.002, 2013
720 Yu, D. J., Sangwan, N., Sung, K., Chen, X., and Merwade, V.: Incorporating institutions
721 and collective action into a sociohydrological model of flood resilience. *Water*
722 *Resources Research*, *53*(2), 1336–1353. https://doi.org/10.1002/2016WR019746,
723 2017

724

725

726

727

728 **Table 1.** Parameters of the flood risk model

729

	description	Values	Ranges in data assimilation	ω in equation (17)
ξ_H	proportion of additional high water level due to levee heightening	0.5	-	-
α_H	parameter related to the slope of the floodplain and the resilience of the human settlement	0.01	-	-
ρ_E	maximum relative growth rate	0.02	-	-
λ_E	critical distance from the river beyond which the settlement can no longer grow	5000	-	-
γ_E	Cost of levee raising	0.5	0.2-5.0	0.01
λ_P	distance at which people would accept to live when they remember past floods whose total consequences were perceived as a total destruction of the settlement	12000	-	
φ_P	rate by which new properties can be built	10000	1000-50000	100
ε_T	safety factor for levees rising	1.1	-	-
κ_T	rate of decay of levees	0.001	0-0.0015	0.0000025
α_S	proportion of shock after flooding if levees are risen	0.5	-	-
μ_S	memory loss rate	0.05	0-0.4	0.0025

730

731

732 **Table 2.** RMSE of the no data assimilation experiment (NoDA) and the data
733 assimilation experiment (DA) in which all observations are assimilated every 10 years
734 with 5000 ensembles in the experiment 1 (see section 3.1).
735

	NoDA	DA
G	1.06×10^{6}	1.64×10^{4}
D	3.60×10^{2}	3.92×10^{1}
H	2.65	1.41
M	1.08×10^{-1}	8.32×10^{-2}

736
737

738 **Table 3.** RMSE of the no data assimilation experiment (NoDA) and the data
739 assimilation experiment (DA) in which all observations are assimilated every 10 years
740 with 5000 ensembles in the experiment 2 (see section 3.2).
741

	NoDA	DA
G	2.97×10^{6}	1.64×10^{4}
D	1.86×10^{3}	1.01×10^{2}
H	9.35	1.63
M	2.24×10^{-1}	8.99×10^{-2}
γ_E	2.08	4.27×10^{-1}
φ_P	1.72×10^{4}	3.81×10^{3}
κ_T	4.12×10^{-4}	2.36×10^{-4}
μ_S	1.55×10^{-1}	2.43×10^{-2}

742
743

744 **Table 4.** RMSE of the no data assimilation experiment (NoDA) and the data
745 assimilation experiment (DA) in which all observations are assimilated every 10 years
746 with 5000 ensembles in the experiment 3 (see section 3.3).
747

	NoDA	DA
G	2.91×10^{6}	6.20×10^{3}
D	2.20×10^{3}	2.02×10^{2}
H	9.21	1.65
M	2.48×10^{-1}	1.05×10^{-1}
γ_E	2.08	5.20×10^{-1}
φ_P	1.98×10^{4}	7.68×10^{3}
κ_T	4.12×10^{-4}	2.54×10^{-4}
μ_S	1.60×10^{-1}	3.03×10^{-2}

748
749

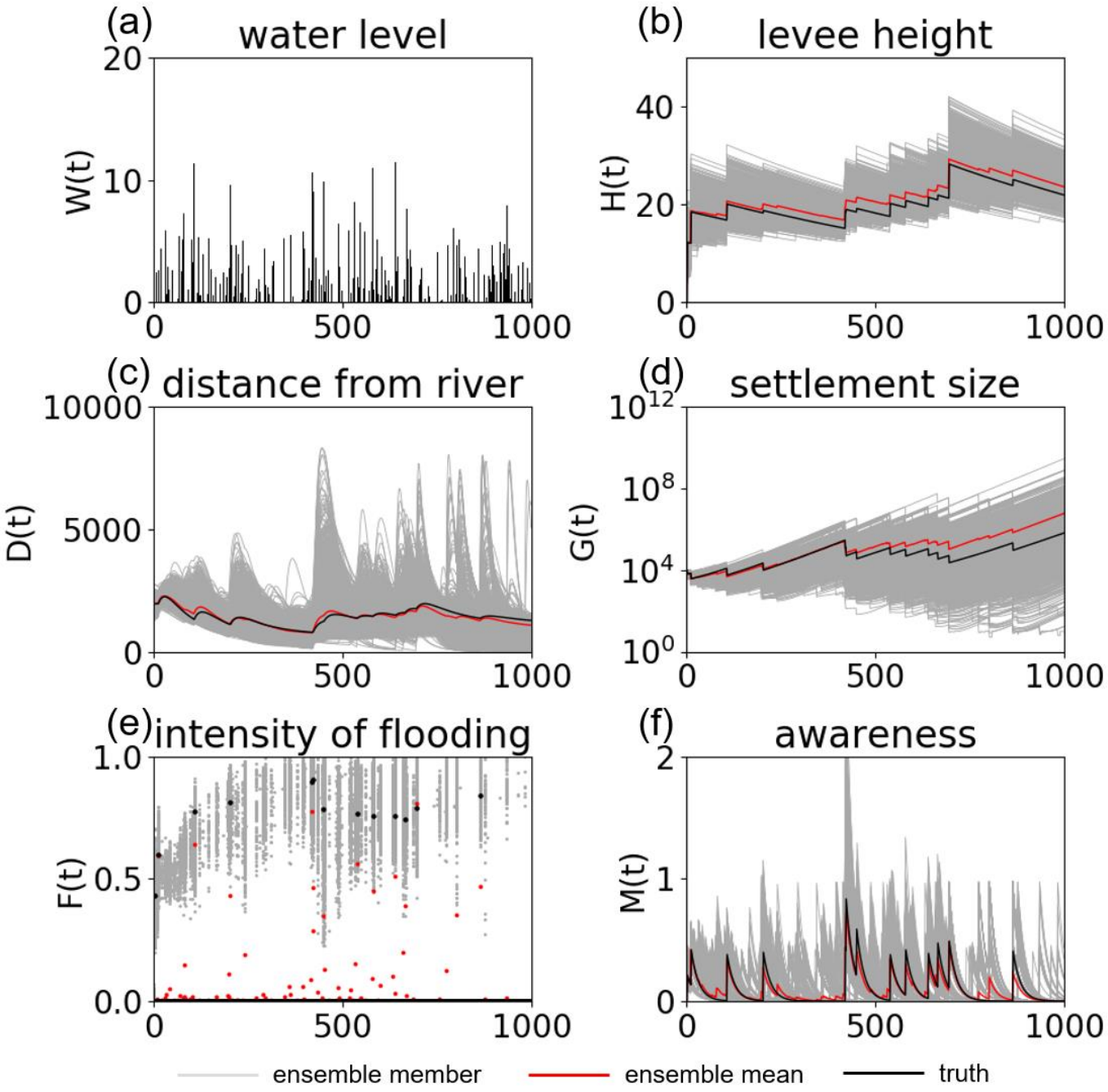


Figure 1. Timeseries of (a) high water level W(t), (b) the flood protection level (or levee height) H(t), (c) the distance of the center of mass of the human settlement from the river D(t), (d) the size of the human settlement G(t), (e) the intensity of flooding events F(t), and (f) the social awareness of the flood risk M(t) simulated by 5000 ensembles with uncertain high water levels and no data assimilation in the experiment 1 (see section

3.1.1). The time step is annual. Grey, red, and black lines are the ensemble members, their mean, and the synthetic truth, respectively.

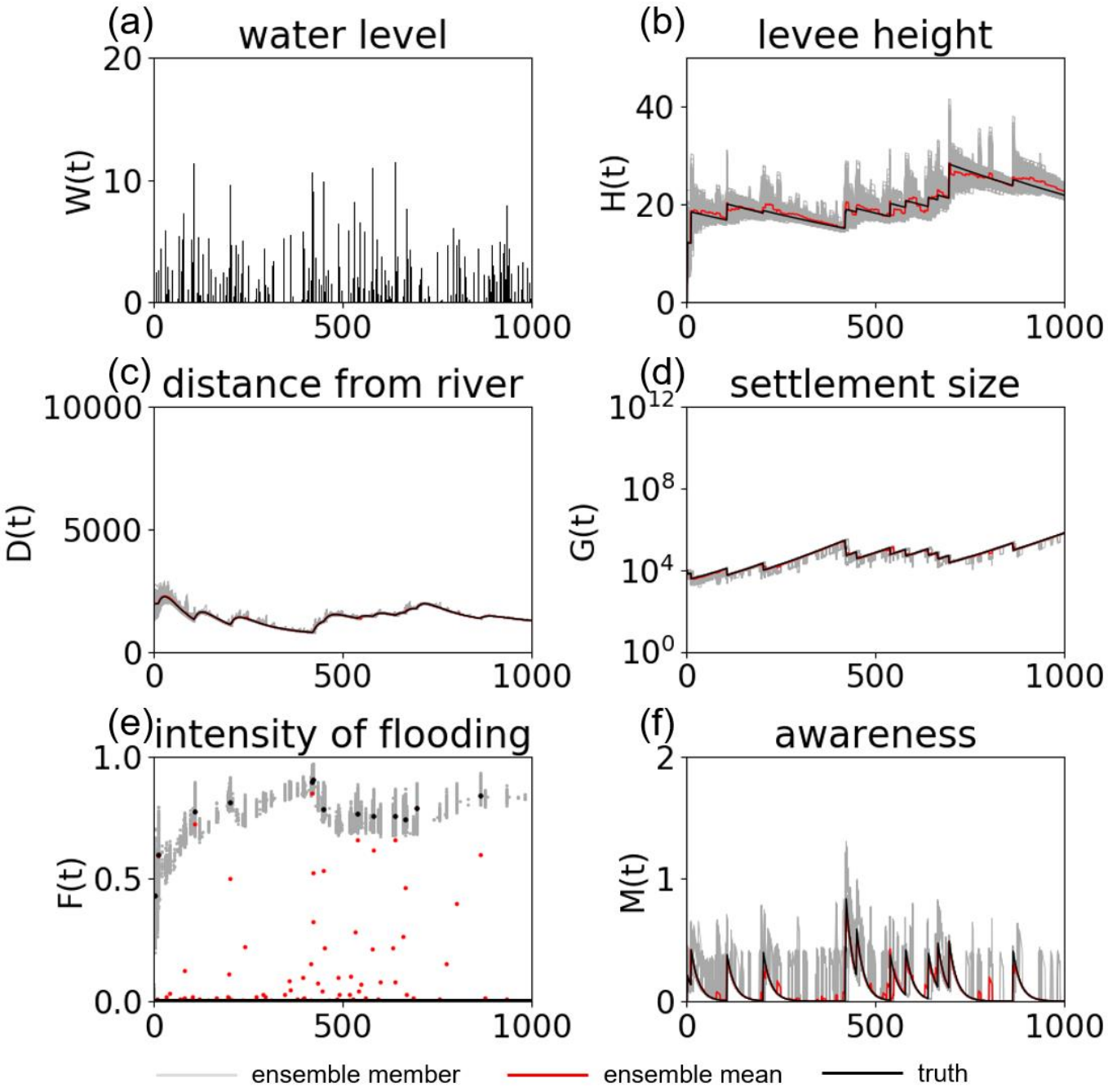
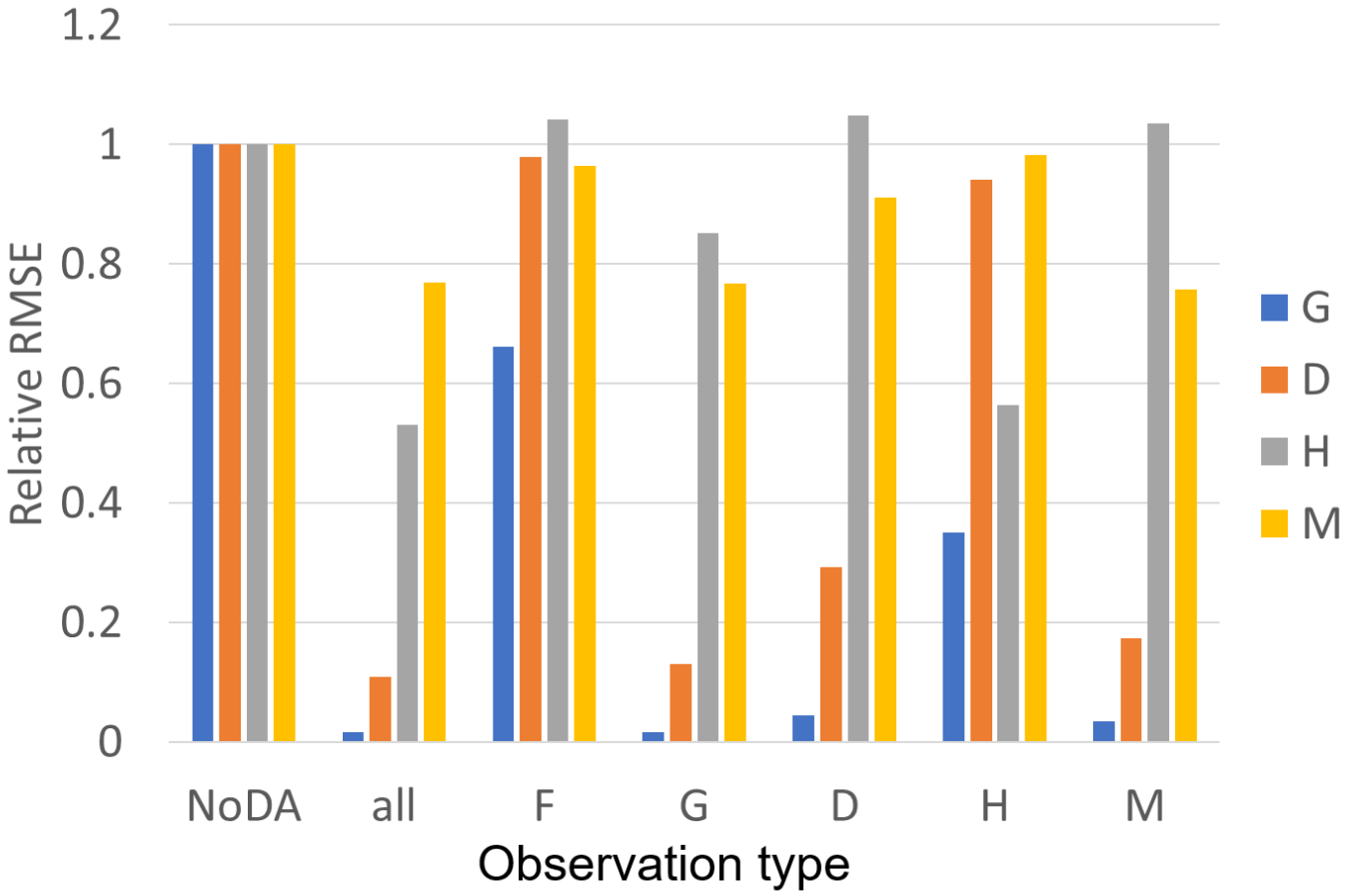


Figure 2. Timeseries of (a) high water level W(t), (b) the flood protection level (or levee height) H(t), (c) the distance of the center of mass of the human settlement from the river D(t), (d) the size of the human settlement G(t), (e) the intensity of flooding events F(t), and (f) the social awareness of the flood risk M(t) simulated by the data assimilation experiment in which the observations of F, G, D, H, and M are assimilated into the model

763 every 10 years with 5000 ensembles in the experiment 1 (see section 3.1.1). The time step is annual. Grey, red,

764 and black lines are the ensemble members, their mean, and the synthetic truth, respectively.

765



766

767 **Figure 3.** The ratio of RMSEs of the no data assimilation experiment (NoDA) to those of the data assimilation

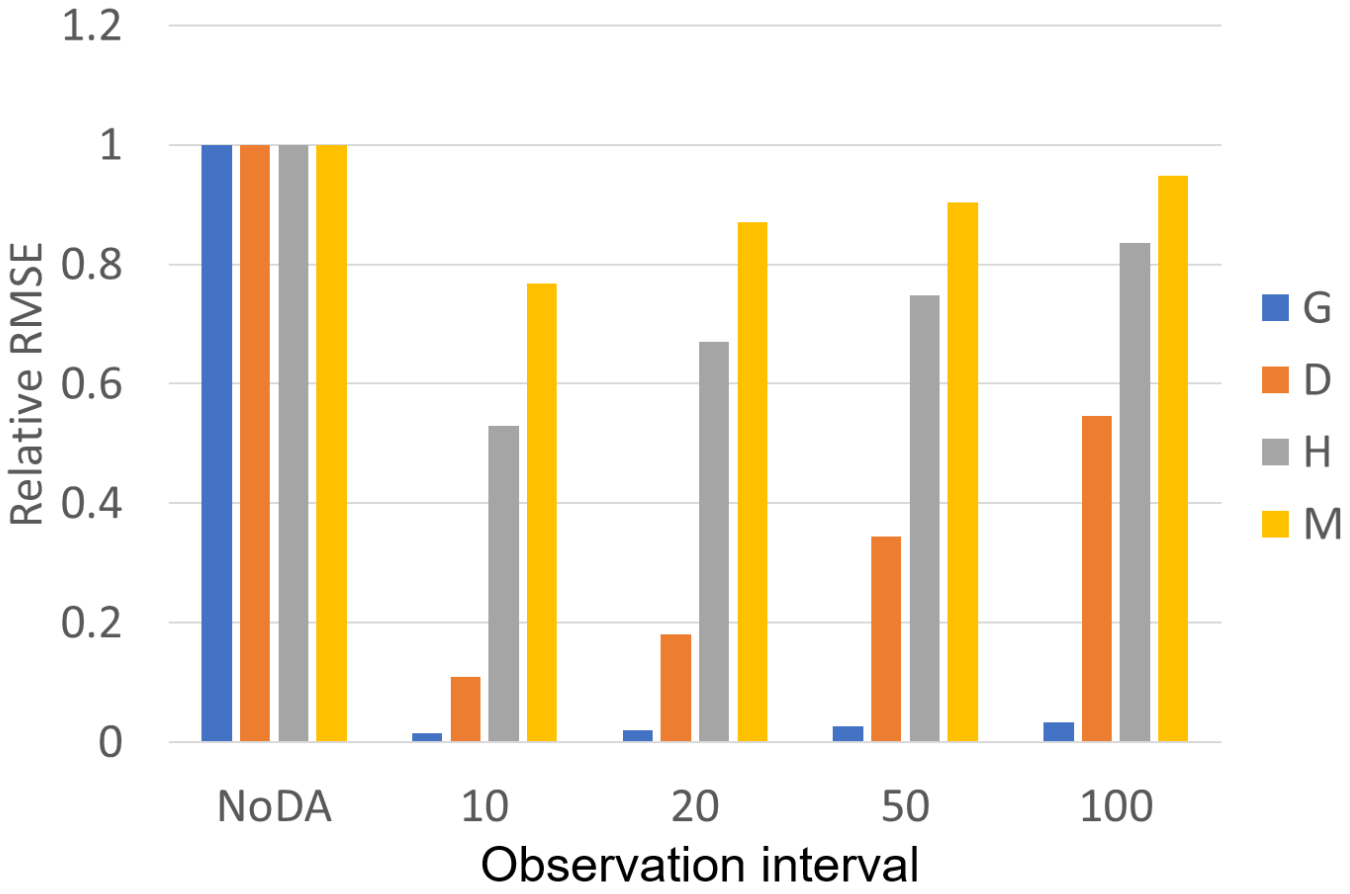
768 experiments in which all of observations (F, G, D, H, and M) are assimilated (all) and each one of them is

769 assimilated in the experiment 1 (see section 3.1.1). Blue, orange, gray, and yellow bars are RMSEs of the size

770 of the human settlement G(t), the center of mass of the human settlement from the river D(t), the flood

771 protection level (or levee height) H(t), and the social awareness of the flood risk M(t).

772



773

774 **Figure 4.** The ratio of RMSEs of the no data assimilation experiment (NoDA) to those of the data assimilation

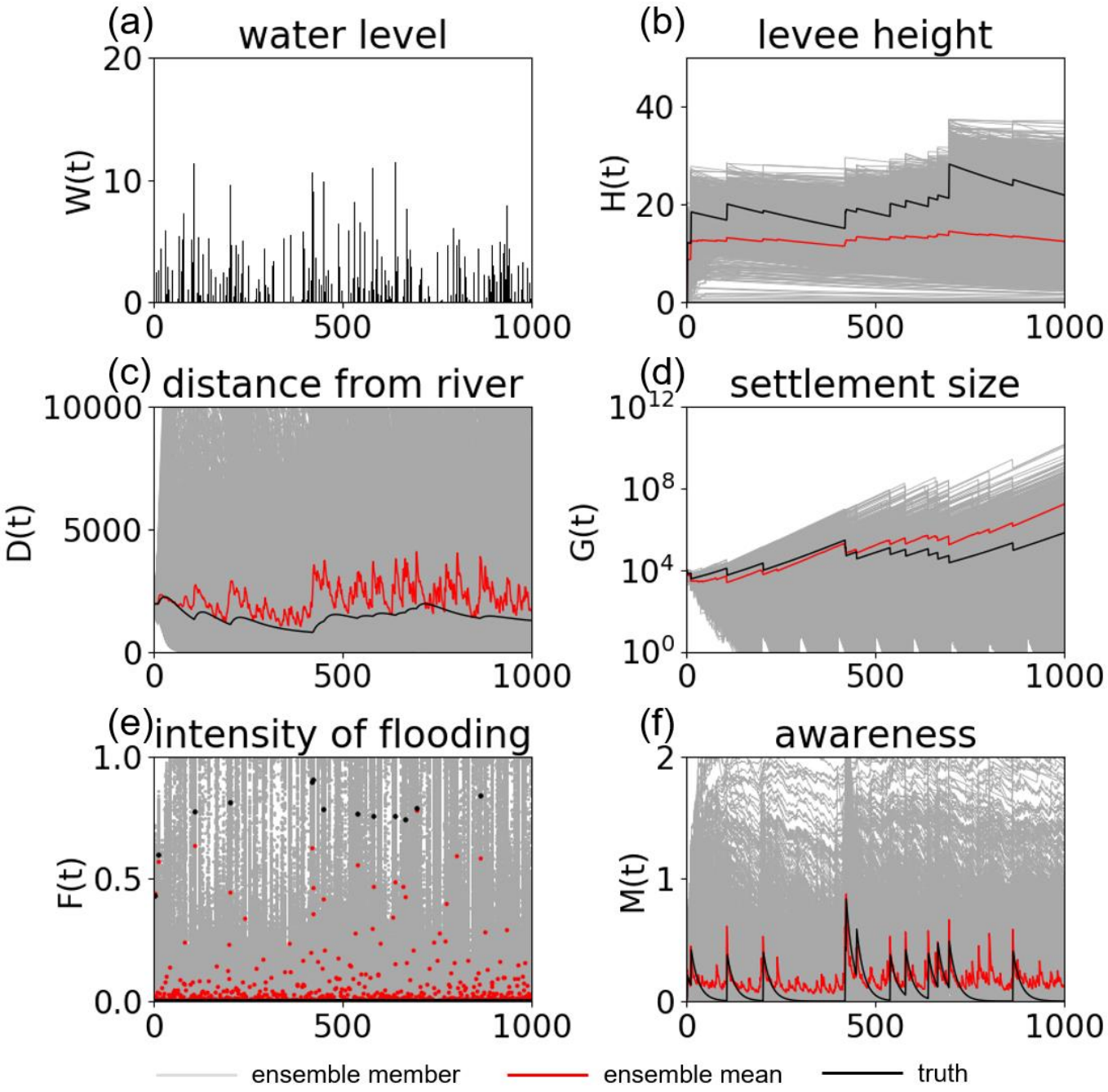
775 experiments in which all of observations (F, G, D, H, and M) are assimilated every 10, 20, 50, and 100 years

776 in the experiment 1 (see section 3.1.1). Blue, orange, gray, and yellow bars are RMSEs of the size of the

777 human settlement G(t), the center of mass of the human settlement from the river D(t), the flood protection

778 level (or levee height) H(t), and the social awareness of the flood risk M(t).

779



780

781 **Figure 5.** Timeseries of (a) high water level W(t), (b) the flood protection level (or levee height) H(t), (c) the

782 distance of the center of mass of the human settlement from the river D(t), (d) the size of the human settlement

783 G(t), (e) the intensity of flooding events F(t), and (f) the social awareness of the flood risk M(t) simulated by

784 5000 ensembles with uncertain high water levels and no data assimilation in the experiment 2 (see section

785 3.1.2). The time step is annual. Grey, red, and black lines are the ensemble members, their mean, and the

786 synthetic truth, respectively.

787

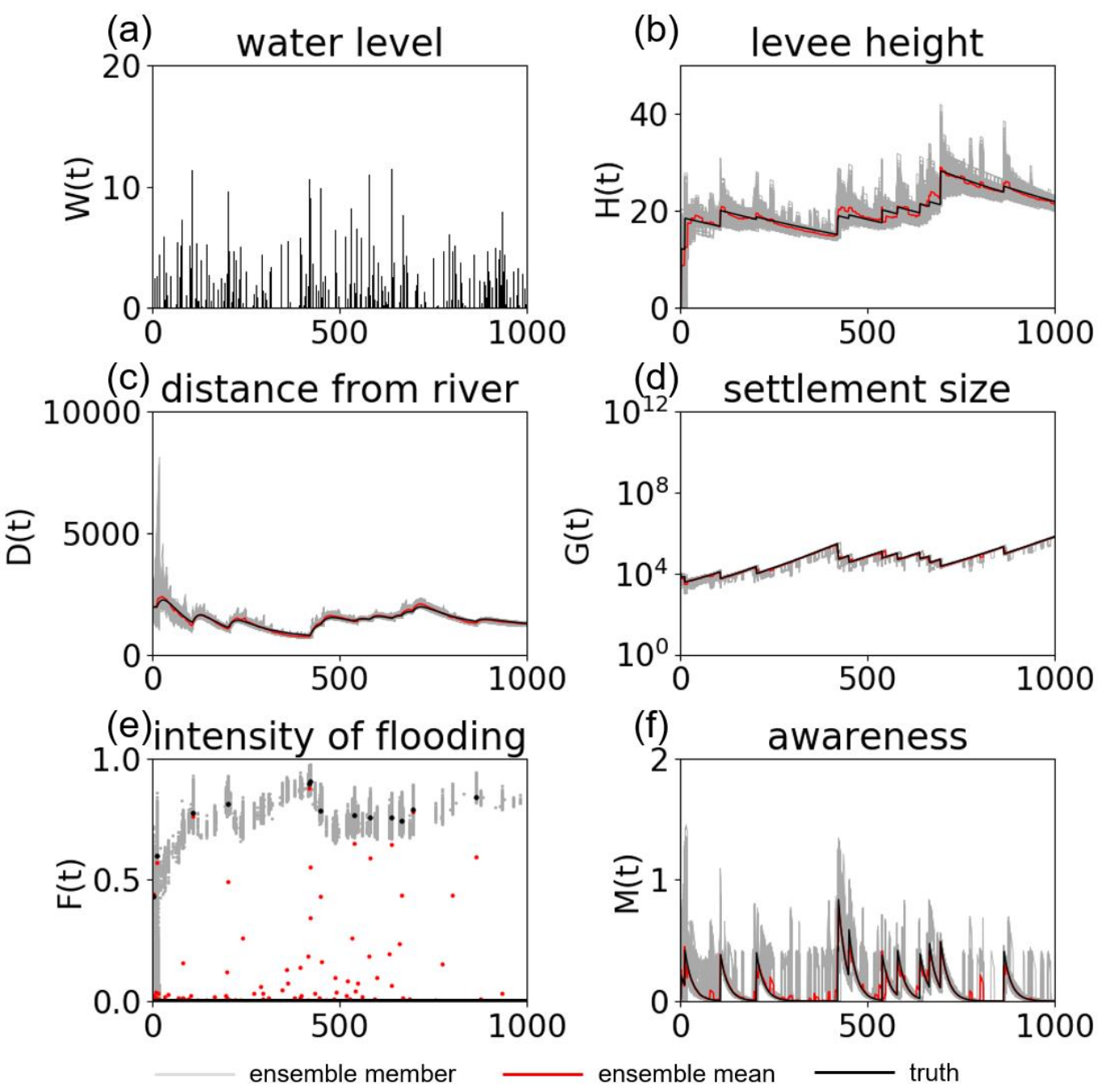


Figure 6. Timeseries of (a) high water level W(t), (b) the flood protection level (or levee height) H(t), (c) the distance of the center of mass of the human settlement from the river D(t), (d) the size of the human settlement G(t), (e) the intensity of flooding events F(t), and (f) the social awareness of the flood risk M(t) simulated by the data assimilation experiment in which the observations of F, G, D, H, and M are assimilated into the model

every 10 years with 5000 ensembles in the experiment 2 (see section 3.1.2). The time step is annual. Grey, red, and black lines are the ensemble members, their mean, and the synthetic truth, respectively.

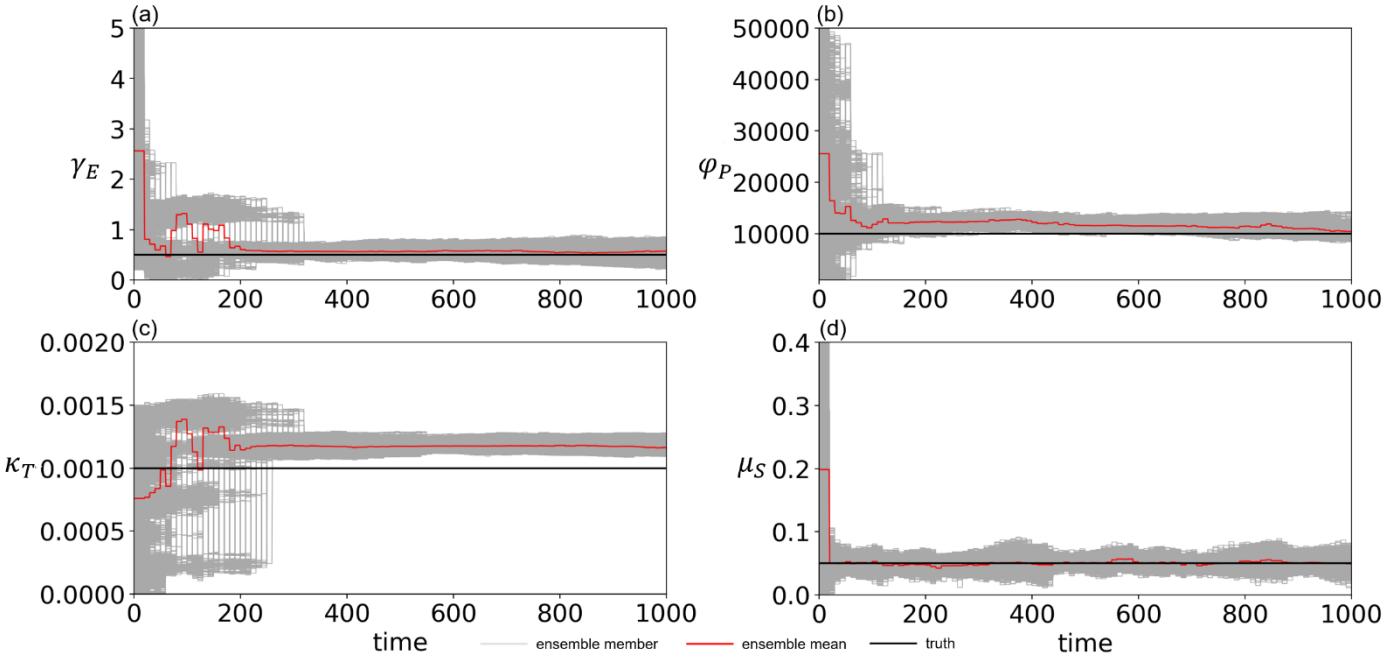


Figure 7. Timeseries of (a) the cost of levee raising γ_E, (b) the rate by which new properties can be built φ_P, (c) the rate of decay of levees κ_T, (d) memory loss rate μ_S estimated by the data assimilation of all observations (F, G, D, H, and M) with 5000 ensembles every 10 years in the experiment 2 (see section 3.1.2). The time step is annual. Grey, red, and black lines are the ensemble members, their mean, and the synthetic truth, respectively.

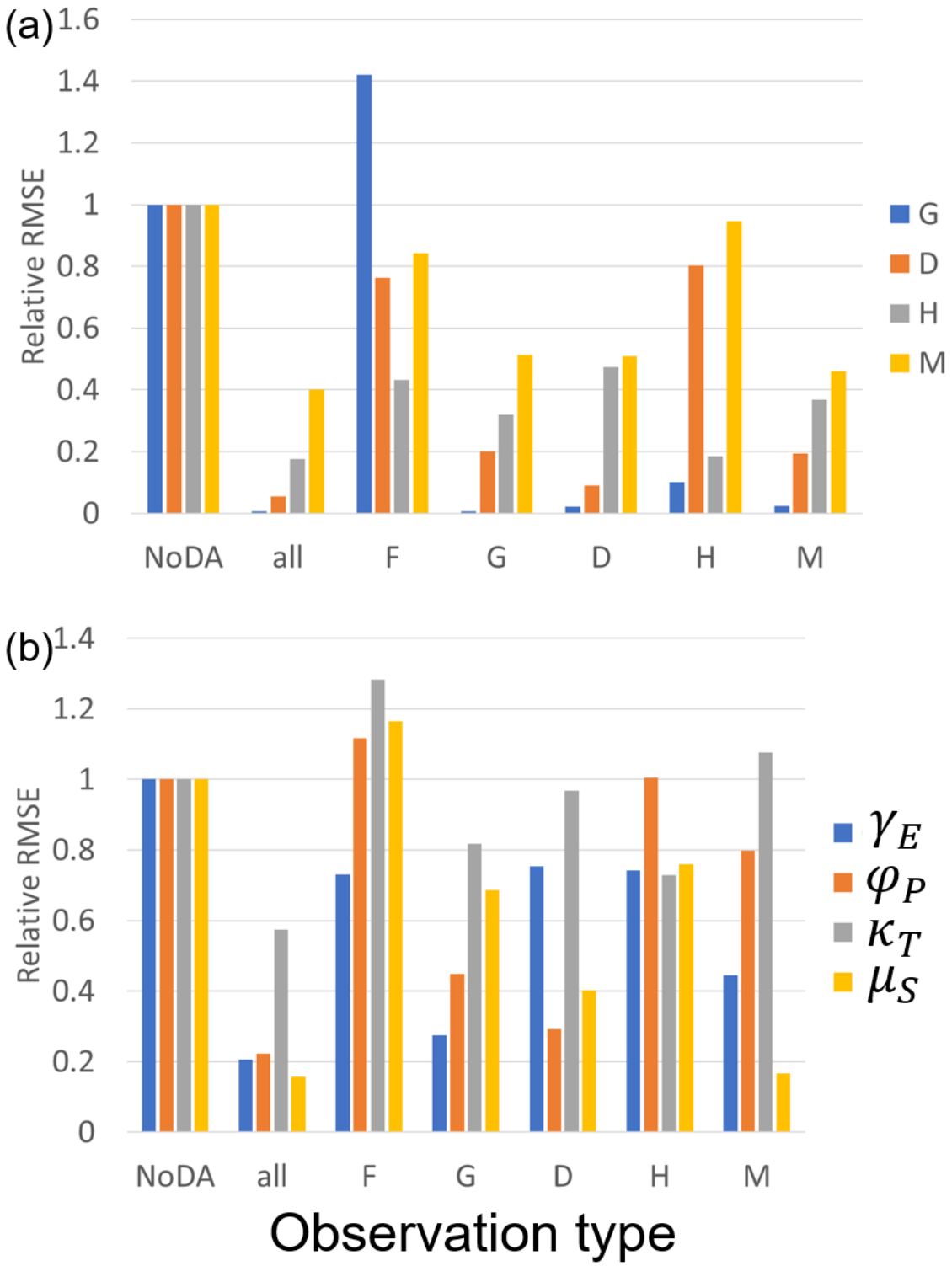
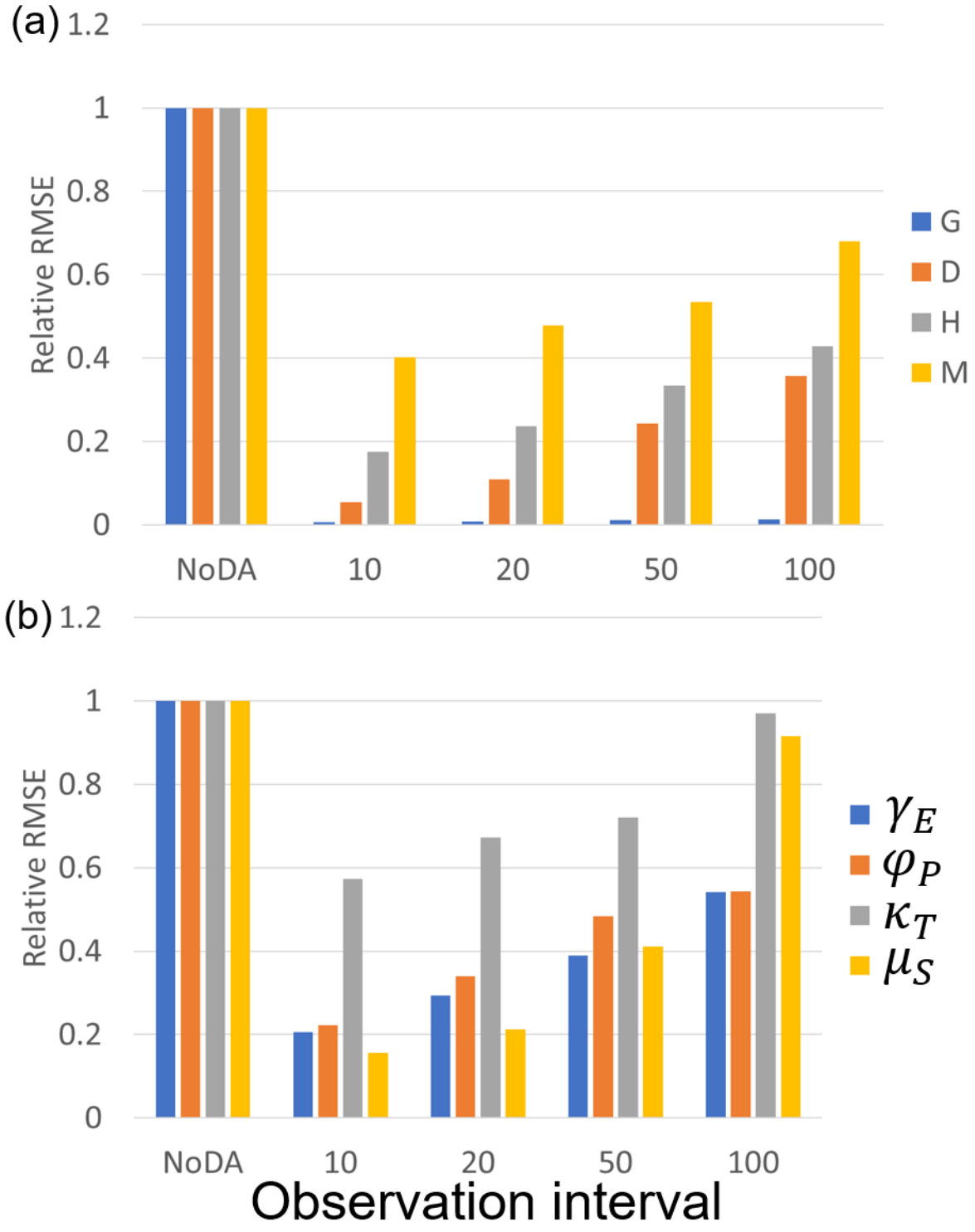


Figure 8. The ratio of RMSEs of the no data assimilation experiment (NoDA) to those of the data assimilation experiments in which all of observations (F, G, D, H, and M) are assimilated (all) and each one of them is assimilated in the experiment 2 (see section 3.1.2). (a) Blue, orange, gray, and yellow bars are RMSEs of the

size of the human settlement G(t), the center of mass of the human settlement from the river D(t), the flood protection level (or levee height) H(t), and the social awareness of the flood risk M(t). (b) Blue, orange, gray, and yellow bars are RMSEs of the cost of levee raising γ_E, the rate by which new properties can be built φ_P, the rate of decay of levees κ_T, memory loss rate μ_S.



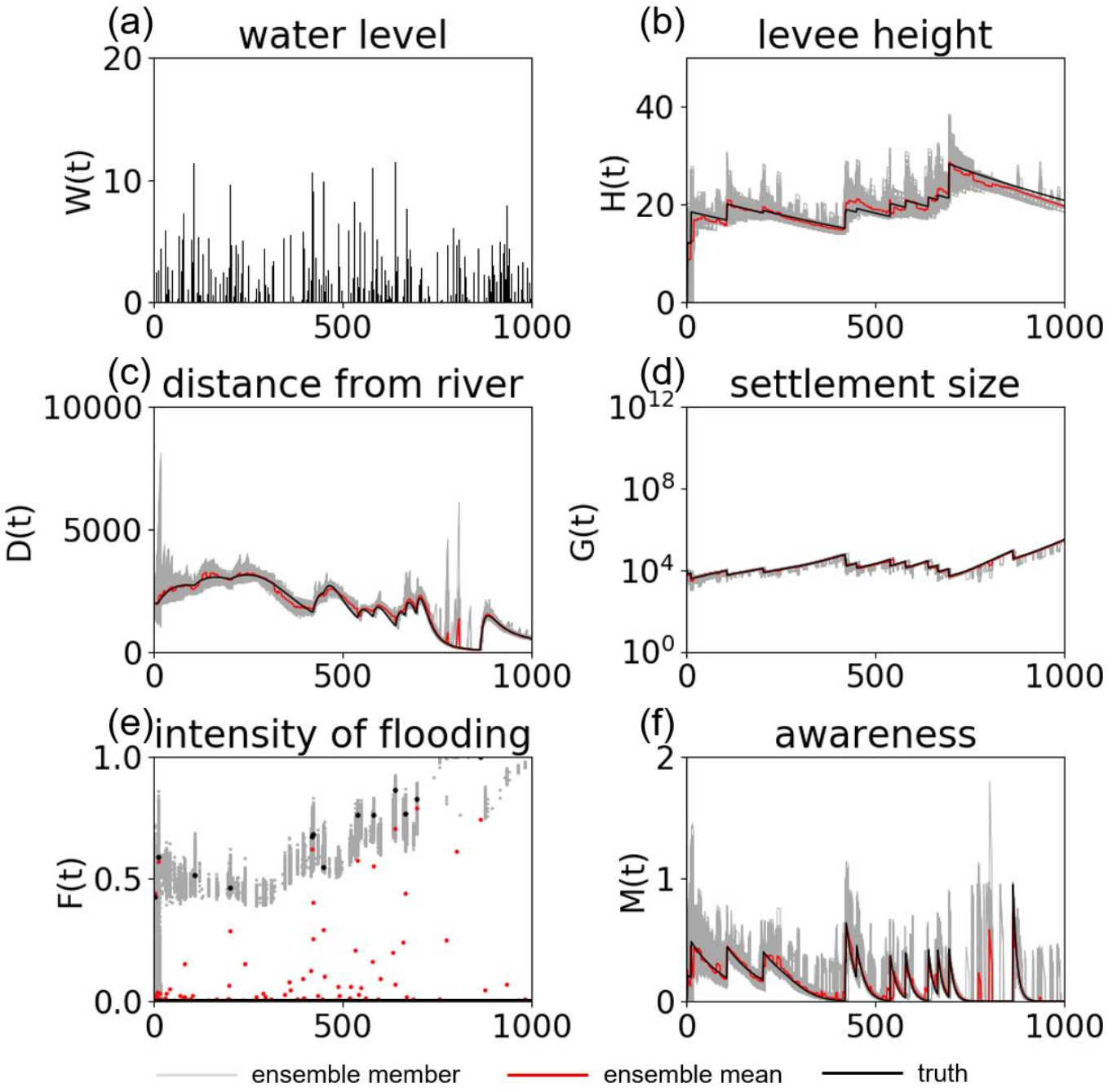
814

815 **Figure 9.** The ratio of RMSEs of the no data assimilation experiment (NoDA) to those of the data assimilation

816 experiments in which all of observations (F, G, D, H, and M) are assimilated every 10, 20, 50, and 100 years

817 in the experiment 2 (see section 3.1.2). (a) Blue, orange, gray, and yellow bars are RMSEs of the size of the

818 human settlement G(t), the center of mass of the human settlement from the river D(t), the flood protection
819 level (or levee height) H(t), and the social awareness of the flood risk M(t). (b) Blue, orange, gray, and yellow
820 bars are RMSEs of the cost of levee raising γ_E, the rate by which new properties can be built φ_P, the rate of
821 decay of levees κ_T, memory loss rate μ_S.
822
823



824

825 **Figure 10.** Timeseries of (a) high water level W(t), (b) the flood protection level (or levee height) H(t), (c) the

826 distance of the center of mass of the human settlement from the river D(t), (d) the size of the human settlement

827 G(t), (e) the intensity of flooding events F(t), and (f) the social awareness of the flood risk M(t) simulated by

828 the data assimilation experiment in which the observations of F, G, D, H, and M are assimilated into the model

every 10 years with 5000 ensembles in the experiment 3 (see section 3.1.3). The time step is annual. Grey, red, and black lines are the ensemble members, their mean, and the synthetic truth, respectively.

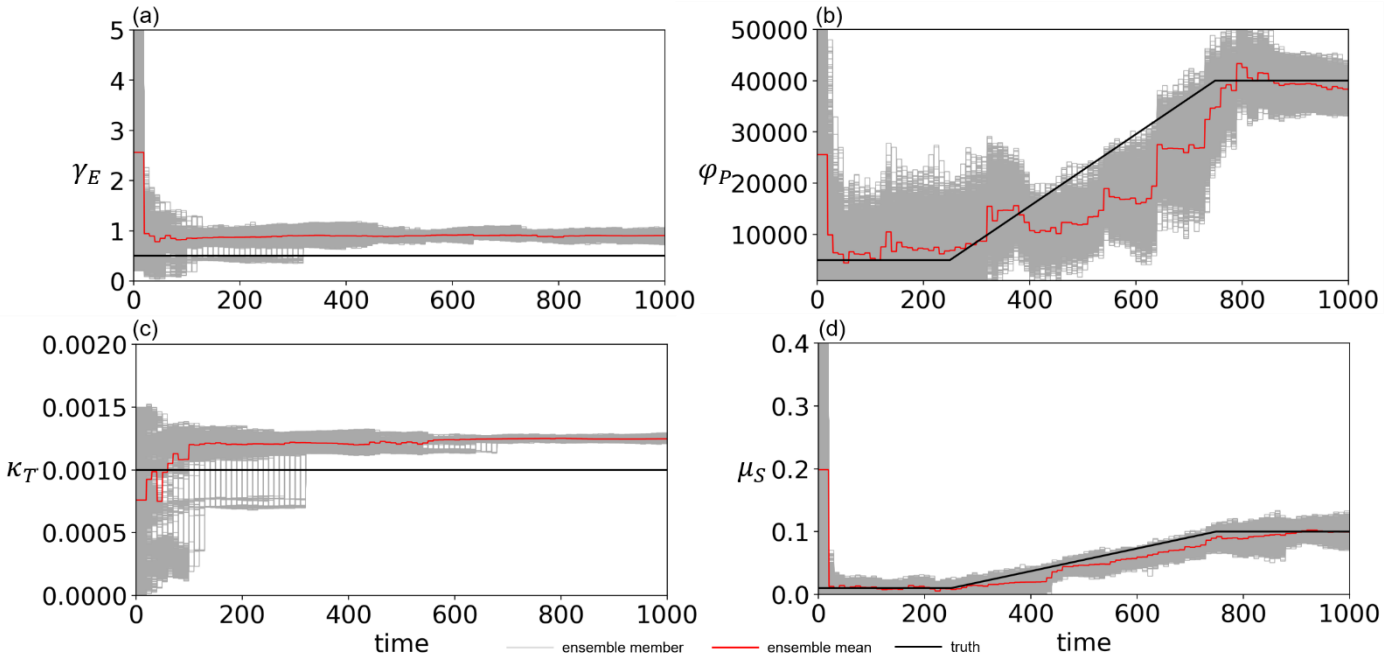


Figure 11. Timeseries of (a) the cost of levee raising γ_E, (b) the rate by which new properties can be built φ_P, (c) the rate of decay of levees κ_T, (d) memory loss rate μ_S estimated by the data assimilation of all observations (F, G, D, H, and M) with 5000 ensembles every 10 years in the experiment 3 (see section 3.1.3). The time step is annual. Grey, red, and black lines are the ensemble members, their mean, and the synthetic truth, respectively.

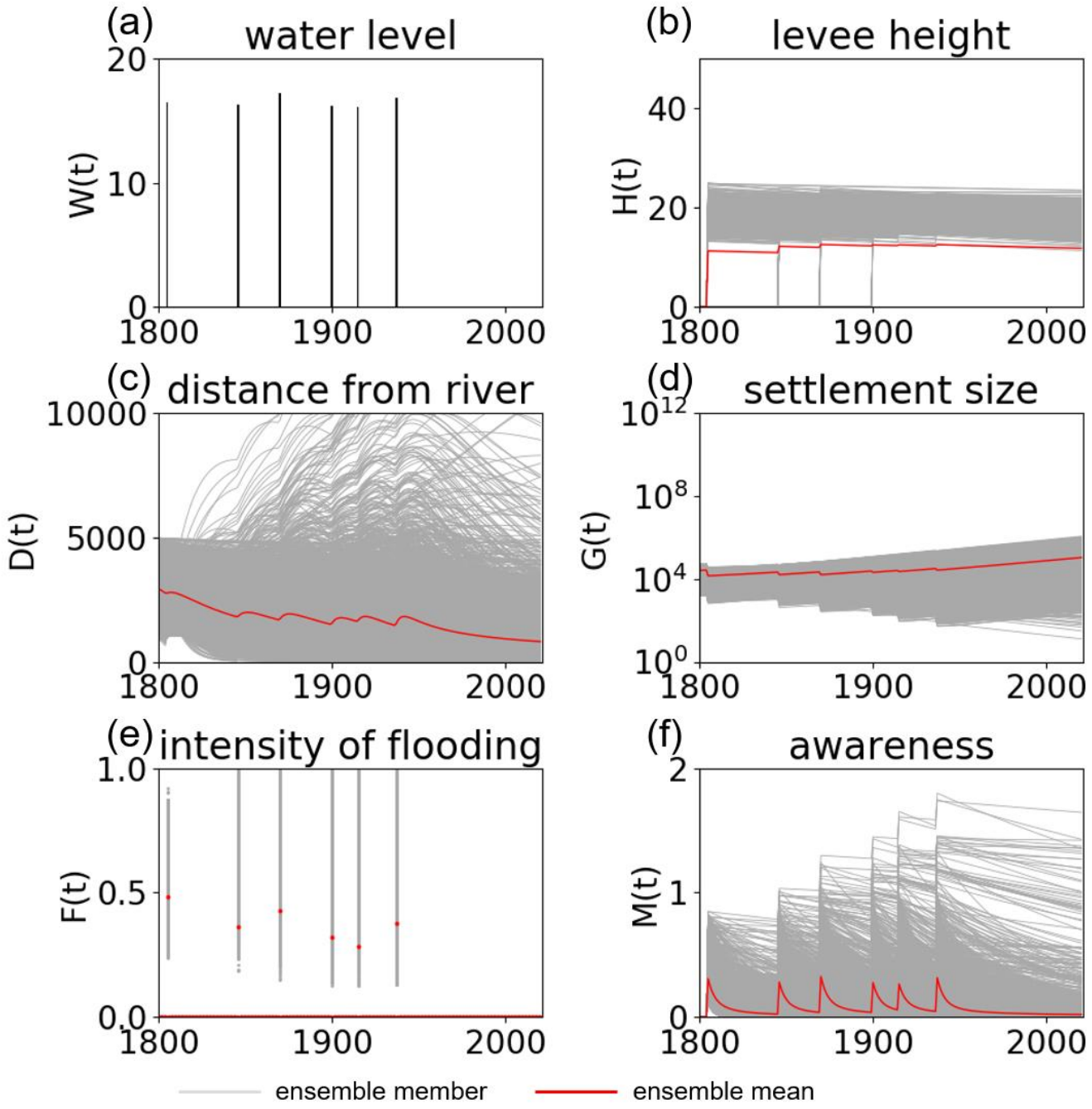
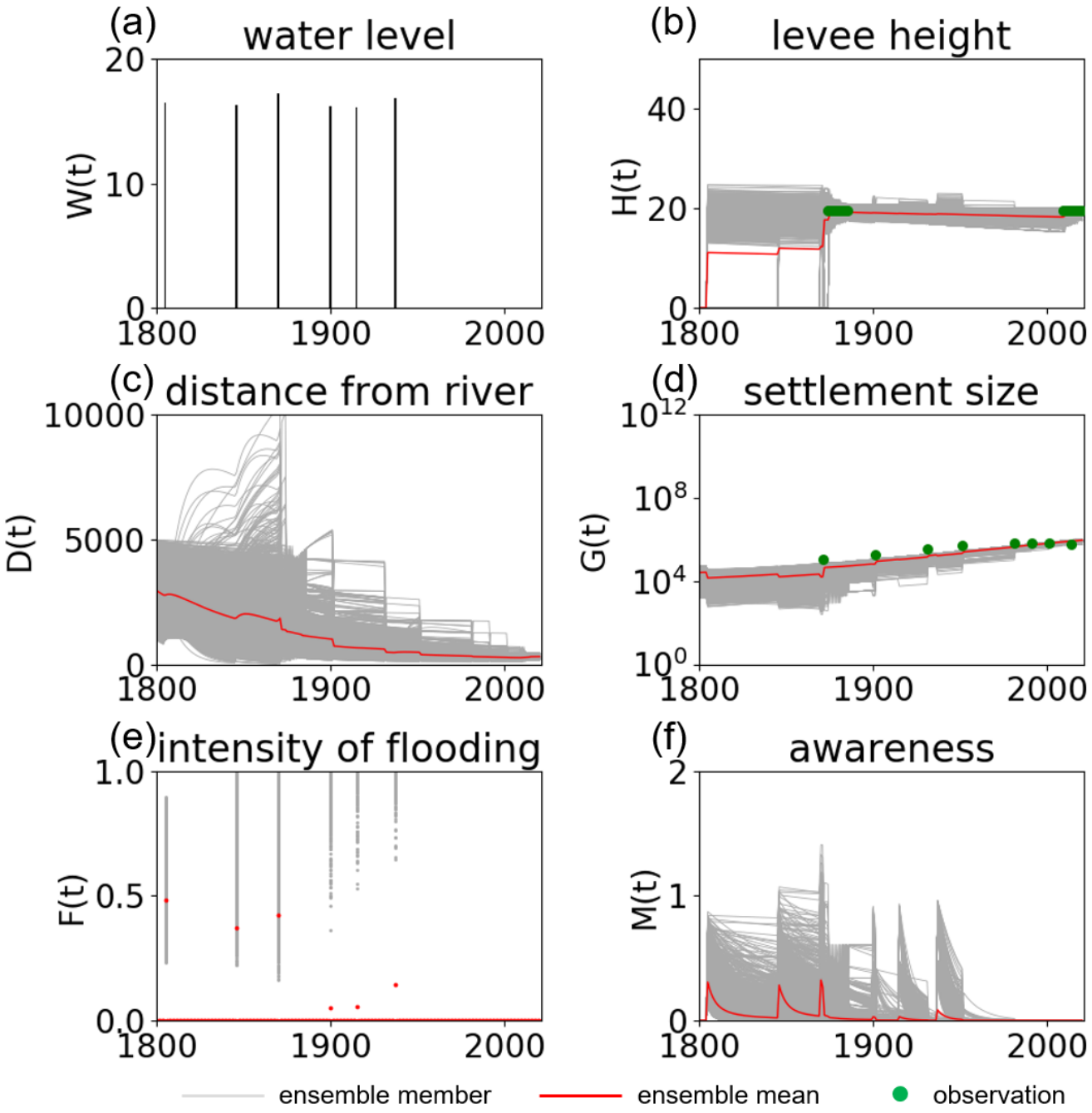


Figure 12. Timeseries of (a) high water level W(t), (b) the flood protection level (or levee height) H(t), (c) the distance of the center of mass of the human settlement from the river D(t), (d) the size of the human settlement G(t), (e) the intensity of flooding events F(t), and (f) the social awareness of the flood risk M(t) simulated by 5000 ensembles with uncertain high water levels and no data assimilation in the real-world experiment in the

844 city of Rome. The time step is annual. Grey, and red lines are the ensemble members and their mean,
845 respectively.
846



847

848 **Figure 13**. Timeseries of (a) high water level W(t), (b) the flood protection level (or levee height) H(t), (c) the

849 distance of the center of mass of the human settlement from the river D(t), (d) the size of the human settlement

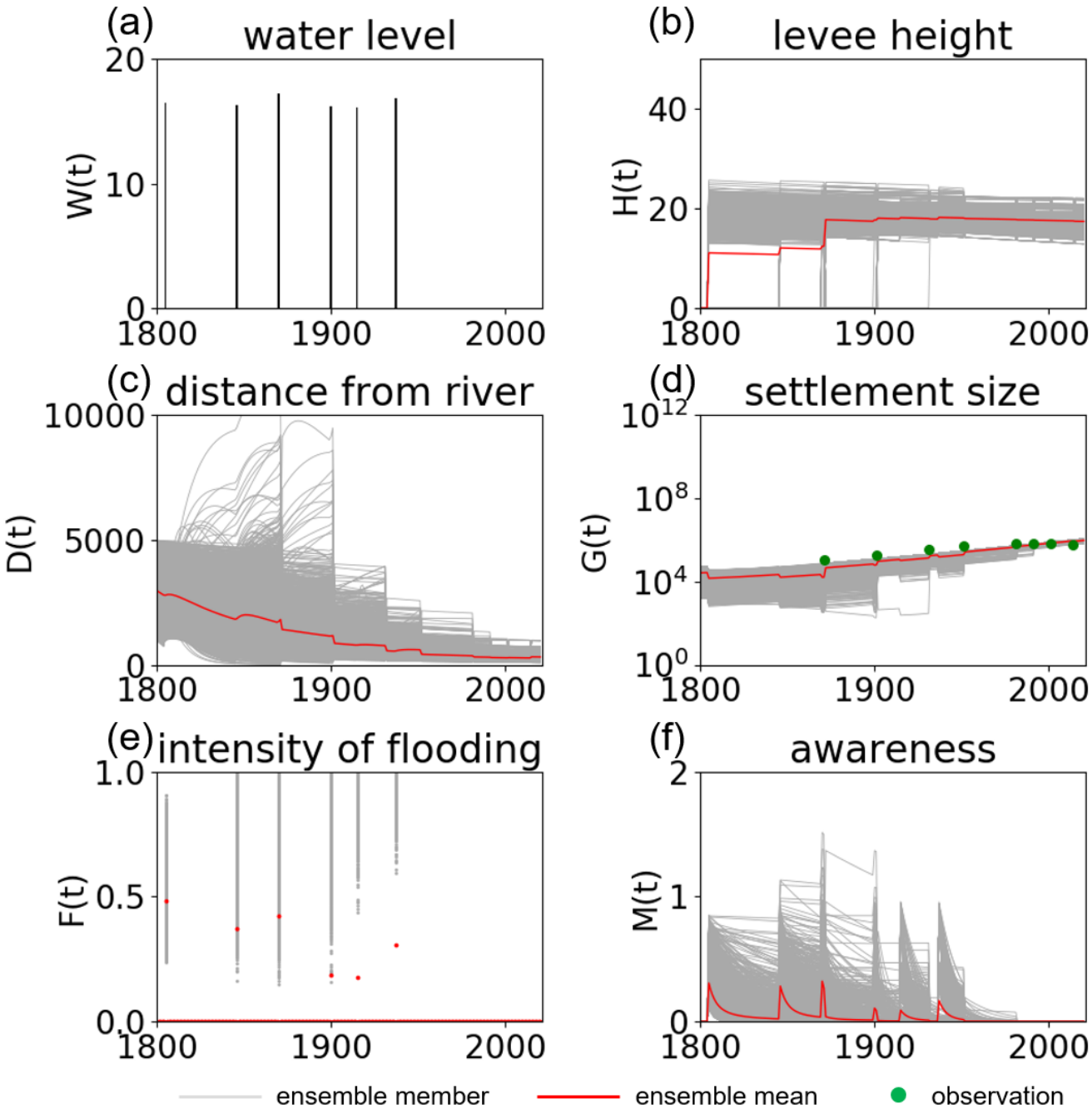
850 G(t), (e) the intensity of flooding events F(t), and (f) the social awareness of the flood risk M(t) simulated by

851 the data assimilation experiment in which the real-world observations of G and H (green dots) are assimilated

852 into the model with 5000 ensembles in the real-world experiment in the city of Rome. The time step is annual.

853 Grey, and red lines are the ensemble members and their mean, respectively.

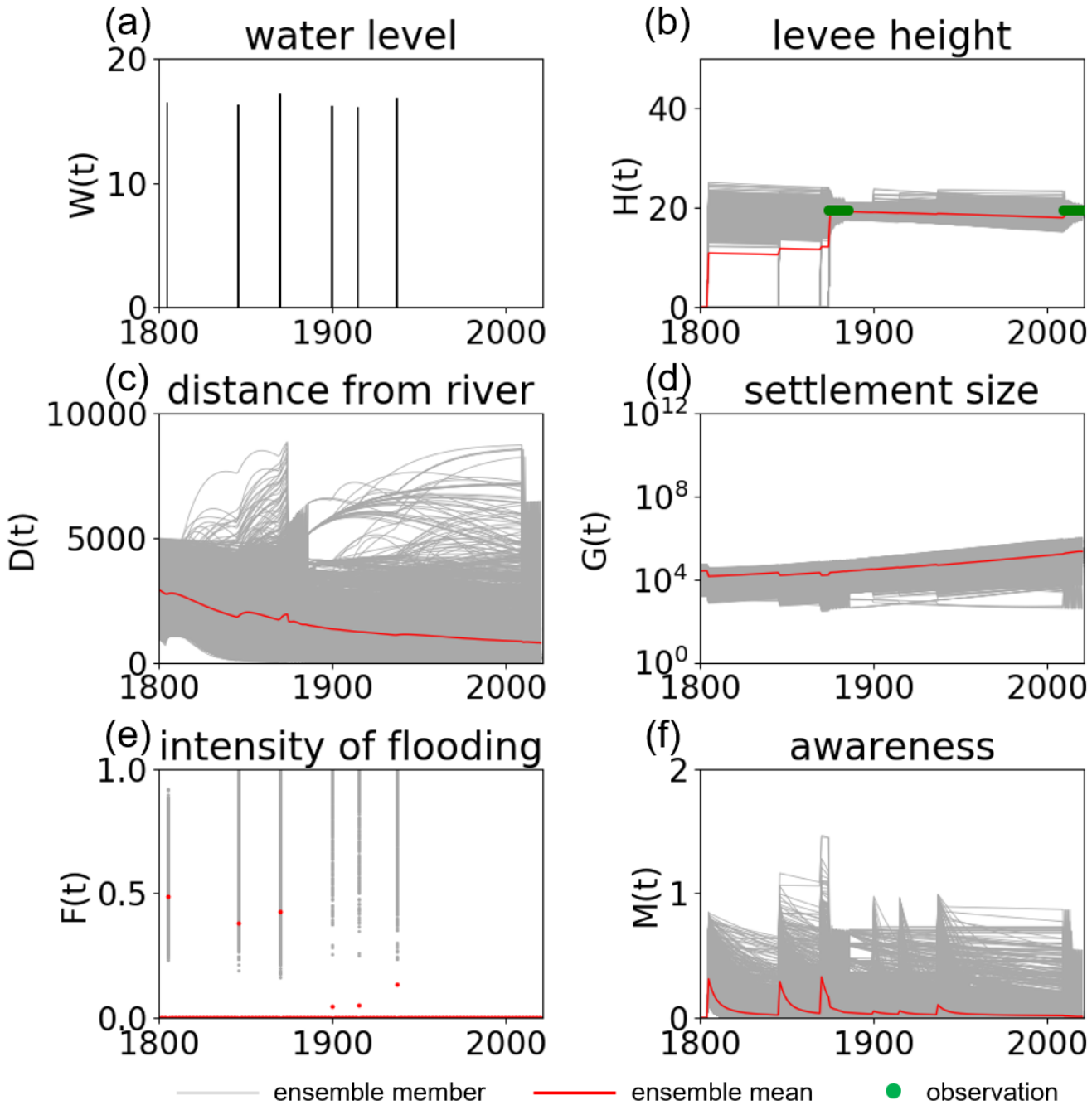
854



855

856 **Figure 14.** Same as Figure 13 but only real data of G are assimilated.

857



858

859 **Figure 15.** Same as Figure 13 but only real data of H are assimilated.

860