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

2 **data integration**

4 Yohei Sawada[1], Risa Hanazaki[2]

5 [1] Institute of Engineering Innovation, School of Engineering, the University of Tokyo,

6 Tokyo, Japan

7 [2] Institute of Industrial Science, the University of Tokyo, Tokyo, Japan

10 Corresponding author: Y. Sawada, Institute of Engineering Innovation, the University of

11 Tokyo, Tokyo, Japan, 2-11-6, Yayoi, Bunkyo-ku, Tokyo, Japan, yohei.sawada@sogo.t.u-

12 tokyo.ac.jp

14 **Abstract**

15 In socio-hydrology, human-water interactions are simulated by mathematical models.

16 Although the integration of these socio-hydrologic models and observation data is

17 necessary to improve the understanding of the human-water interactions, the

18 methodological development of the model-data integration in socio-hydrology is in its

19 infancy. Here we propose to apply sequential data assimilation, which has been widely

20 used in geoscience, to a socio-hydrological model. We developed particle filtering for a

21 widely adopted flood risk model and performed an idealized observation system

22 simulation experiment and a real-data experiment to demonstrate the potential of the

23 sequential data assimilation in socio-hydrology. In these experiments, the flood risk

24 model's parameters, the input forcing data, and empirical social data were assumed to be

25 somewhat imperfect. We tested if data assimilation can contribute to accurately

26 reconstructing the historical human-flood interactions by integrating these imperfect

27 models and imperfect and sparsely distributed data. Our results highlight that it is

28 important to sequentially constrain both state variables and parameters when the input

29 forcing is uncertain. Our proposed method can accurately estimate the model's unknown

30 parameters even if the true model parameter temporally varies. The small amount of

31 empirical data can significantly improve the simulation skill of the flood risk model.

32   Therefore, sequential data assimilation is useful to reconstruct historical socio-

33   hydrological processes by the synergistic effect of models and data.

1. Introduction

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

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

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.

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).

91

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.

101

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.

108

109

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:

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

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

(2)

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

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

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

127 $$\frac{dH}{dt} = \Delta(Y(t))R - \kappa_T H \qquad (6)$$

128 $$\frac{dM}{dt} = \Delta(Y(t))S - \mu_S M \qquad (7)$$

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), $\Delta\big(\Upsilon(t)\big) = 1$ with integral only when time t

146   passes the time of the flooding event (F>0), otherwise, $\Delta\big(\Upsilon(t)\big) = 0$. 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

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

163 model.

164

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

166 $$\boldsymbol{x}(t+1) = f\big(\boldsymbol{x}(t), \boldsymbol{\theta}, \boldsymbol{u}(t)\big) + \boldsymbol{q}(t) \tag{8}$$

167 where $\boldsymbol{x}(t)$ is the state variables (i.e. G, D, H, and M), $\boldsymbol{\theta}$ is the model parameters, $\boldsymbol{u}(t)$

168 is the external forcing (i.e., the high water level), and $\boldsymbol{q}(t)$ is the noise process which

169 represents the model error. In data assimilation, it is useful to formulate an observation

170 process as follows:

171 $$\boldsymbol{y}^f(t) = h\big(\boldsymbol{x}(t)\big) + \boldsymbol{r}(t) \tag{9}$$

172 where $\boldsymbol{y}^f(t)$ is the simulated observation, h is the observation operator which maps the

173 model's state variables into the observable variables, and $\boldsymbol{r}(t)$ is the noise process which

174 represents the observation error.

175

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

177 parameters:

178 $$p(\boldsymbol{x}(t), \boldsymbol{\theta}|\boldsymbol{y}^o(1{:}t)) \propto p(\boldsymbol{y}^o(t)|\boldsymbol{x}(t), \boldsymbol{\theta})p(\boldsymbol{x}(t), \boldsymbol{\theta}|\boldsymbol{y}^o(1{:}t-1)) \tag{10}$$

179   where $p(\boldsymbol{x}(t), \boldsymbol{\theta}|\boldsymbol{y}^o(1:t))$ is the posterior probability of the state variables $\boldsymbol{x}(t)$ and

180   parameters $\boldsymbol{\theta}$ given all observations up to time t $\boldsymbol{y}^o(1:t)$. The prior knowledge,

181    $p(\boldsymbol{x}(t), \boldsymbol{\theta}|\boldsymbol{y}^o(1:t-1))$, based on the model integration is updated using the likelihood

182   which includes the new observation at time t $p(\boldsymbol{y}^o(t)|\boldsymbol{x}(t), \boldsymbol{\theta})$. In this study, we assumed

183   that our observation error follows Gaussian distribution so that the likelihood can be

184   formulated as follows:

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

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

187   where **R** is the covariance matrix of the observation error process $\boldsymbol{r}(t)$. The prior

188   knowledge of the state variables is approximated by the ensemble simulation:

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

190   where N is the ensemble size, $\boldsymbol{x}^i, \boldsymbol{\theta}^i, \boldsymbol{u}^i$ are the realizations of the ensemble member i,

191   and $\delta[.]$ is the Direc delta function.

192

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

194   follows:

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

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

197   where $w(i)$ is the normalized weight for the realization of the ensemble member i and

198   is calculated using the likelihood (see also equation (11)).

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

200   Note that equations (13) and (14) update all state variables and parameters of the model

201   although the weight is calculated using only observable variables. Therefore, it is not

202   necessary to observe all state variables in order to update all system variables.

203

204   The implementation of SIRPF is the following:

205   1. Model state variables are updated from time t-1 to t using ensemble

206   simulation (equations (8) and (12)).

207   2. Simulated observations are calculated for all ensembles (equation (9)).

208   3. The likelihood for each ensemble member is calculated (equation (11))

209   4. The weights are obtained for all ensembles (equation (15))

210   5. We applied a resampling procedure according to the normalized weights.

211   The normalized weights of ensemble i, $w(i)$, can be recognized as the

212   probability that the ensemble i is selected after resampling. Resampled state

213   variables and parameters are defined as x_{resamp}^i and θ_{resamp}^i, respectively.

214   6. Since there are no mechanisms to increase the variance of parameters of

215   ensemble members, Moradkhani et al. (2005) proposed to perturb the

216   ensembles of parameters:

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

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

219   where $N(.)$ is the Gaussian distribution, Var^θ is the variance of $\boldsymbol{\theta}^i$, $\boldsymbol{\omega}$

220   is the fixed hyperparameter (see Table 1 for its variable) which guarantees

221   that the ensembles of parameters do not converge into a single value. s is

222   an adaptively changed factor according to the effective ensemble size, N_{eff}.

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

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

225   where $s_0 = 0.05$. The effective ensemble size is the measure of the

226   diversity of ensembles. If the effective ensemble size becomes small,

227   ensembles should be strongly perturbed in order to maintain the diversity of

228   ensembles. Similar strategy has been used in many SIRPF systems (e.g.,

229   Moradkhani et al. 2005; Poterjoy et al. 2019).

230

231

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:

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

v follows the Gumbel distribution:

$$p(v) = \frac{\exp{(-\frac{v-\mu}{\beta})}}{\beta} \exp{(-\exp(-(v-\mu)\beta))} \qquad (21)$$

249   where $\mu = 9, \beta = 2.5$. Although our high water level is not identical to Di Baldassarre

250   et al. (2013), the estimated trajectory of the state variables is similar to Di Baldassarre et

251   al. (2013).

252

253   Synthetic observations were generated by adding the Gaussian white noise to the F, G, D,

254   H, and M (see section 2.1) of the synthetic truth. The mean of the Gaussian white noise

255   was 0. The observation error, the standard deviation of the Gaussian white noise, was

256   firstly set to 10% of the synthetic true variables. Although this observation error is

257   generally larger than that used in meteorology and hydrology, we further increased the

258   observation error and tested the sensitivity of the observation error to the SIRPF's

259   performance. We firstly assumed that all of the F, G, D, H, and M can be observed every

260   10 years or every 10 model integration steps. Then, we evaluated the sensitivity of the

261   observation network (i.e. the observable variables and the observation intervals) to the

262   SIRPF's performance. Although it is not straightforward to observe social memory M,

263   several previous studies obtained the proxy of the social memory by interview data

264   (Barendrecht et al. 2019) and the number of Google searches (Gonzales and Ajami 2017).

265

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

267   metrics:

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

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

270   where $RMSE^i$ is root-mean-square-error for i th ensemble, T is the computational period,

271    $x^i(t)$ is the simulated state variables of ensemble i at time t, $z(t)$ is the synthetic truth

272   at time t.

273

274

275   **3.1.1. Experiment 1: Perfect model with uncertain high water levels**

276   In the first OSSE, we assumed that there is no uncertainty in model parameters. We used

277   the same parameter variables as the synthetic truth run and we did not perform the

278   estimation of parameters. Our SIRPF updated only state variables. Although the model

279   had no uncertainty, it was assumed that the input data, the timeseries of the high water

280   level, were uncertain. Lognormal multiplicative noise was added to the synthetic true high

281   water level so that different ensemble members have different high water levels in the

282   data assimilation experiment. The two parameters of the lognormal distribution,

283   commonly called μ and σ, were set to 0 and 0.15, respectively.

284

285

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

In the second OSSE, we assumed that some of the synthetic true parameter values were

unknown. The unknown parameters in the experiment 2 were 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, and

memory loss rate μ_S (see Table 1). We selected these unknown parameters one by one

from four equations of economy, politics, technology, and social to discuss how each state

variable's observation affects the estimation of parameters across these four equations

(see section 2.1). We have no unknown parameters related to F (equation (1)) since it is

unlikely that the parameters in equation (1) are much more inaccurate than the other

parameters. The parameters related to flood are mainly determined by the topography of

the flood plain so that the process described in equation (1) can be replaced by more

accurate hydrodynamic models in the real-world case study. The initial parameter

variables were assumed to be distributed in the bounded uniform distributions whose

ranges were found in Table 1. The uncertainty of the simulation induced by these

parameters' uncertainty is large enough to demonstrate the potential of data assimilation

to minimize the simulation's uncertainty (see Results). Our SIRPF sequentially

302 assimilated observations and estimated both state variables and parameters in the

303 experiment 2. The high water level data were uncertain as the experiment 1.

304

305

306 **3.1.3. Experiment 3: Unknown and time-variant model parameters and uncertain**

307 **high water levels**

308 To further demonstrate the potential of sequential data assimilation in socio-hydrology,

309 we assumed that the description of the model was biased in the experiment 3. Here we

310 assumed that two of the model parameters were temporally varied by the unknown

311 dynamics. Specifically, the rate by which new properties can be built, φ_P, and the

312 memory loss rate, μ_S, were temporally varied in the experiment 3:

313 $$\varphi_P(t) = \begin{cases} 5000 \ (t < 250) \\ 5000 + (t - 250) \times \frac{40000 - 5000}{500} \ (250 \leq t < 750) \\ 40000 \ (750 \leq t) \end{cases} \quad (24)$$

314 $$\mu_S(t) = \begin{cases} 0.01 \ (t < 250) \\ 0.01 + (t - 250) \times \frac{0.10 - 0.01}{500} \ (250 \leq t < 750) \\ 0.10 \ (750 \leq t) \end{cases} \quad (25)$$

315 In the data assimilation experiment, we assumed that the dynamics of φ_P and μ_S was

316 unknown, and we integrated the flood risk model with time-invariant φ_P and μ_S. We

317 evaluated if SIRPF could track this time-variant parameter and reveal the bias of the

318 model's description. The cost of levee raising γ_E, and the rate of decay of levees κ_T

319   were assumed to be time-invariant unknown parameters as they were in the experiment

320   2. The cost of levee raising γ_E affects the state variables of the flood risk model mainly

321   in the initial early years and the gradual change of the rate of decay of levees κ_T has few

322   impacts on the state variables. Therefore, we found that it is difficult to track the temporal

323   change of these two parameters. The input forcing data, high water level, were uncertain

324   as described in the experiment 1.

325

326

327   **3.2. Real-data experiment**

328   In addition to the OSSEs, we performed the real-world experiment in the city of Rome,

329   Italy. Ciullo et al. (2017) collected real-world data and calibrated their flood risk model.

330   Using the data collected by Ciullo et al. (2017), we performed the data assimilation

331   experiment. It should be noted that the flood risk model of Ciullo et al. (2017) is different

332   from our model (i.e. Di Baldassarre et al. 2013), although they are conceptually similar.

333

334   All the data were collected from Figure 1 of Ciullo et al. (2017) by WebPlotDigitizer

335   (https://automeris.io/WebPlotDigitizer/). The observed high water level of Tiber River

336   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).

444

445   As we found in the experiment 1, the SIRPF's performance declines with the increased

446   observation error (Figure S6). However, it is promising that our SIRPF can improve the

447   simulation skill with larger observation errors up to 25% of the synthetic truth considering

448   that the observations in the socio-hydrologic domain are often inaccurate.

449

450   In contrast to the experiment 1, the larger ensemble size is required to stably estimate both

451   state variables and parameters (Figure S7). The increased degree of freedom and the

452   nonlinear relationship between parameters and observations increase the necessary

453   ensemble size.

454

455

456   **4.1.3. Experiment 3: Unknown and time-variant model parameters and uncertain**

457   **high water levels**

458   In addition to the experiment 2, two of the unknown parameters (φ_P and μ_S) temporally

459   vary in the synthetic truth of the experiment 3. We found that a larger spread of φ_P is

460   required to stably track the time-variant synthetic true φ_P so that we increased s_0 in

461   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**

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

533   First, the sequential data assimilation can mitigate the negative impact of the uncertainty

534   in the input forcing on the simulation of socio-hydrologic state variables. We found that

535   the small perturbation of high water levels greatly affects the long-term trajectory of the

536   socio-hydrologic state variables as Viglione et al. (2014) found. It is necessary to

537   sequentially constrain the state variables and parameters by sequential data assimilation

538   if the input forcing is uncertain although previous studies on the model-data integration

539   in socio-hydrology mainly focused on parameter calibration assuming no uncertainty in

540   the input forcing (e.g., Barendrecht et al. 2019; Roobavannan et al. 2017; Ciullo et al.

541   2017; van Emmerik et al. 2014; Gonzales and Ajami 2017). To deeply understand the

542   socio-hydrologic processes, the long-term historical analysis should be performed.

543   Although there are many studies on the accurate reconstruction of the historical weather

544   condition (e.g., Toride et al. 2017), it may be necessary to tackle with the uncertainty in

545   hydrometeorological datasets used for the input forcing of the socio-hydrologic models.

546

547   Second, our SIRPF can efficiently improve the simulation of the socio-hydrologic state

548   variables using the sparsely distributed data. All model variables should not necessarily

549   be observed to constrain the model's state variables and parameters. In some cases,

550   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

587 into a model, the (non-linear) observation operator (see equation (9)), the assignment of

588 the observation error, and assimilation methods should be carefully designed as

589 previously discussed in the context of numerical weather prediction (e.g., Sawada et al.

590 2019; Okamoto et al. 2019; Minamide and Zhang 2017). Future work will focus on the

591 methodological development to efficiently assimilate observations in the social domain

592 with complicated structure of observation operators and errors.

593

594 **6. Conclusion**

595 In this study, we proposed to apply the sequential data assimilation to the socio-

596 hydrologic models. By several OSSEs and the real-data experiment in the flood risk

597 modeling, we found that our proposed SIRPF is robust to the imperfect input forcing and

598 the imperfect model. The sequential data assimilation is useful to reconstruct the socio-

599 hydrologic conditions from the inaccurate and sparsely distributed data and the imperfect

600 simulation.

601

602

603 **Acknowledgements**

604   We thank Di Baldassarre for sharing the original source code of the flood risk model. We

605   thank two anonymous referees for their constructive comments. Data Integration and

606   Analysis System (DIAS) provided us the computational resources.

607

608   **Code/Data availability**

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

610

611   **Author Contribution**

612   YS designed the study. RH and YS jointly developed the data assimilation system for the

613   flood risk model and performed the numerical experiments. YS and RH contributed to

614   interpreting the results. YS wrote the first draft of the paper and RH contributed to editing

615   the paper.

616

617   **Competing interests**

618   The authors declare that they have no conflict of interest.

619

620   **References**

621 Barendrecht, M. H., Viglione, A., Kreibich, H., Merz, B., Vorogushyn, S., and Blöschl,

622 G.: The Value of Empirical Data for Estimating the Parameters of a

623 Sociohydrological Flood Risk Model. *Water Resources Research*.

624 https://doi.org/10.1029/2018WR024128, 2019

625 Bauer, P., Thorpe, A., and Brunet, G.: The quiet revolution of numerical weather

626 prediction. *Nature*, *525*(7567), 47–55. https://doi.org/10.1038/nature14956, 2015

627 Ciullo, A., Viglione, A., Castellarin, A., Crisci, M., and Di Baldassarre, G.: Socio-

628 hydrological modelling of flood-risk dynamics: comparing the resilience of green

629 and technological systems. *Hydrological Sciences Journal*, *62*(6), 880–891.

630 https://doi.org/10.1080/02626667.2016.1273527, 2017

631 Dang, Q., and Konar, M.: Trade Openness and Domestic Water Use. *Water Resources*

632 *Research*, *54*(1), 4–18. https://doi.org/10.1002/2017WR021102, 2018

633 Di Baldassarre, G., Viglione, A., Carr, G., Kuil, L., Salinas, J. L., and Blöschl, G.: Socio-

634 hydrology: Conceptualising human-flood interactions. *Hydrology and Earth*

635 *System Sciences*, *17*(8), 3295–3303. https://doi.org/10.5194/hess-17-3295-2013,

636 2013

637   Di Baldassarre, G., et al.: Socio-hydrology: Scientific Challenges in Addressing a Societal

638   Grand Challenge. *Water Resources Research*, 1–29.

639   https://doi.org/10.1029/2018wr023901, 2019

640   Gonzales, P., and Ajami, N.: Social and Structural Patterns of Drought-Related Water

641   Conservation and Rebound. *Water Resources Research*, *53*(12), 10619–10634.

642   https://doi.org/10.1002/2017WR021852, 2017

643   Hersbach, H. et al.: Global reanalysis: goodbye ERA-Interim, hello ERA5, *ECMWF*

644   *Newsletter, 159,* 17-24, doi: 10.21957/vf291hehd7, 2019

645   Kobayashi, S., et al.: The JRA-55 Reanalysis: General Specifications and Basic

646   Characteristics. *Journal of the Meteorological Society of Japan, 93,* 5-48.

647   https://doi.org/10.2151/jmsj.2015-001, 2015

648   Kreibich, H., et al.: Adaptation to flood risk: Results of international paired flood event

649   studies. *Earth's Future*. https://doi.org/10.1002/eft2.232, 2017

650   Lievens, H., et al.: Joint Sentinel-1 and SMAP data assimilation to improve soil moisture

651   estimates. *Geophysical Research Letters*, *44*(12), 6145–6153.

652   https://doi.org/10.1002/2017GL073904, 2017

653   Minamide, M., and Zhang, F: Adaptive Observation Error Inflation for Assimilating All-

654   Sky Satellite Radiance., *Monthly Weather. Review*, *145,* 1063–1081,

655   https://doi.org/10.1175/MWR-D-16-0257.1, 2017

656   Miyoshi, T., and Yamane, S.: Local Ensemble Transform Kalman Filtering with an

657   AGCM at a T159/L48 Resolution. *Monthly Weather Review*, *135*(2002), 3841–

658   3861. https://doi.org/10.1175/2007MWR1873.1, 2007

659   Moradkhani, H., Hsu, K. L., Gupta, H., and Sorooshian, S.: Uncertainty assessment of

660   hydrologic model states and parameters: Sequential data assimilation using the

661   particle filter. *Water Resources Research*, *41*(5), 1–17.

662   https://doi.org/10.1029/2004WR003604, 2005

663   Mostert, E.: An alternative approach for socio-hydrology: Case study research. *Hydrology*

664   *and Earth System Sciences*, *22*(1), 317–329. https://doi.org/10.5194/hess-22-317-

665   2018, 2018

666   Mount, N., J., et al.: Data-driven modelling approaches for sociohydrology: opportunities

667   and challenges within the Panta Rhei Science Plan. *Hydrological Sciences Journal,*

668   *61(7),* 1192-1208. https://doi.org/10.1080/02626667.2016.1159683, 2016

669 Okamoto, K, Sawada, Y, Kunii, M. Comparison of assimilating all-sky and clear-sky

670   infrared radiances from Himawari-8 in a mesoscale system. *Q J R Meteorol Soc.*,

671   *145*, 745-766. https://doi.org/10.1002/qj.3463, 2019

672 Pande, S., and Savenije, H. H. G.: A sociohydrological model for smallholder farmers in

673   Maharashtra, India. *Water Resources Research*, *52*(3), 1923–1947.

674   https://doi.org/10.1002/2015WR017841, 2016

675 Pathiraja, S., Anghileri, D., Burlando, P., Sharma, A., Marshall, L., and Moradkhani, H.:

676   Time-varying parameter models for catchments with land use change: the

677   importance of model structure, Hydrol. Earth Syst. Sci., *22*, 2903–2919,

678   https://doi.org/10.5194/hess-22-2903-2018, 2018.

679 Penny, S. G., and Miyoshi, T.: *A local particle filter for high-dimensional geophysical*

680   *systems*. 391–405. https://doi.org/10.5194/npg-23-391-2016, 2016

681 Poterjoy, J., Wicker, L., and Buehner, M.: Progress toward the application of a localized

682   particle filter for numerical weather prediction. *Monthly Weather Review*, *147*(4),

683   1107–1126. https://doi.org/10.1175/MWR-D-17-0344.1, 2019

684 Qin, J., Liang, S., Yang, K., Kaihotsu, I., Liu, R., and Koike, T.: Simultaneous estimation

685   of both soil moisture and model parameters using particle filtering method through

686   the assimilation of microwave signal. *Journal of Geophysical Research*, *114*(D15),

687   1–13. https://doi.org/10.1029/2008JD011358, 2009

688   Rasmussen, J., Madsen, H., Jensen, K. H., and Refsgaard, J. C.: Data assimilation in

689   integrated hydrological modeling using ensemble Kalman filtering: evaluating the

690   effect of ensemble size and localization on filter performance. *Hydrology and Earth*

691   *System Sciences*, *19*(7), 2999–3013. https://doi.org/10.5194/hess-19-2999-2015,

692   2015

693   Roobavannan, M., Kandasamy, J., Pande, S., Vigneswaran, S., and Sivapalan, M.: Role

694   of Sectoral Transformation in the Evolution of Water Management Norms in

695   Agricultural Catchments: A Sociohydrologic Modeling Analysis. *Water Resources*

696   *Research*, *53*(10), 8344–8365. https://doi.org/10.1002/2017WR020671, 2017

697   Sawada, Y., Koike, T., and Walker, J. P.: A land data assimilation system for simultaneous

698   simulation of soil moisture and vegetation dynamics. *J. Geophys. Res. Atmos.*, *120*,

699   5910– 5930. doi: 10.1002/2014JD022895, 2015

700   Sawada, Y., Nakaegawa, T. and Miyoshi, T.: Hydrometeorology as an inversion problem:

701   Can river discharge observations improve the atmosphere by ensemble data

702   assimilation? *Journal of Geophysical Research: Atmospheres*, 123, 848– 860.

703   https://doi.org/10.1002/2017JD027531, 2018

704 Sawada, Y., Okamoto, K., Kunii, M., and Miyoshi, T.: Assimilating every-10-minute

705   Himawari-8 infrared radiances to improve convective predictability. *Journal of*

706   *Geophysical   Research: Atmospheres*,   124,   2546–2561.

707   https://doi.org/10.1029/2018JD029643, 2019

708 Sivapalan, M., Savenije, H.H.G. and Blöschl, G.: Socio-hydrology: A new science of

709   people and water. Hydrol. Process., 26: 1270-1276. doi:10.1002/hyp.8426, 2012

710 Sivapalan, M., Konar, M., Srinivasan, V., Chhatre, A., Wutich, A., Scott, C. A., and

711   Wescoat, J. L.: Socio-hydrology: Use-inspired water sustainability science for the

712   Anthropocene, *Earth's Future*, *2*, 225–230. https://doi.org/10.1002/2013EF000164,

713   2014.

714 Toride, K., Neluwala, P., Kim, H. and Yoshimura, K.: Feasibility Study of the

715   Reconstruction of Historical Weather with Data Assimilation. *Mon. Wea. Rev.,* **145**,

716   3563–3580, https://doi.org/10.1175/MWR-D-16-0288.1, 2017

717 Van Emmerik, T. H. M., et al.: Socio-hydrologic modeling to understand and mediate the

718   competition for water between agriculture development and environmental health:

719   Murrumbidgee River basin, Australia. Hydrology and Earth System Sciences,

720   18(10), 4239–4259. https://doi.org/10.5194/hess-18-4239-2014, 2014

721   Viglione, A., et al.: Insights from socio-hydrology modelling on dealing with flood risk -

722   Roles of collective memory, risk-taking attitude and trust. *Journal of Hydrology*,

723   *518*(PA), 71–82. https://doi.org/10.1016/j.jhydrol.2014.01.018, 2014

724   Vrugt, J. A., ter Braak, C. J. F., Diks, C. G. H., and Schoups, G.: Hydrologic data

725   assimilation using particle Markov chain Monte Carlo simulation: Theory, concepts

726   and applications. *Advances in Water Resources*, *51*, 457–478.

727   https://doi.org/10.1016/j.advwatres.2012.04.002, 2013

728   Yu, D. J., Sangwan, N., Sung, K., Chen, X., and Merwade, V.: Incorporating institutions

729   and collective action into a sociohydrological model of flood resilience. *Water*

730   *Resources Research*, *53*(2), 1336–1353. https://doi.org/10.1002/2016WR019746,

731   2017

732

733

734

735

Table 1. Parameters of the flood risk model

736

737

	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

738

739

740   **Table 2.** RMSE of the no data assimilation experiment (NoDA) and the data
741   assimilation experiment (DA) in which all observations are assimilated every 10 years
742   with 5000 ensembles in the experiment 1 (see section 3.1).
743

	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}

744
745

746 **Table 3.** RMSE of the no data assimilation experiment (NoDA) and the data
747 assimilation experiment (DA) in which all observations are assimilated every 10 years
748 with 5000 ensembles in the experiment 2 (see section 3.2).
749

	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}

750
751

752 **Table 4.** RMSE of the no data assimilation experiment (NoDA) and the data
753 assimilation experiment (DA) in which all observations are assimilated every 10 years
754 with 5000 ensembles in the experiment 3 (see section 3.3).
755

	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}

756
757

758

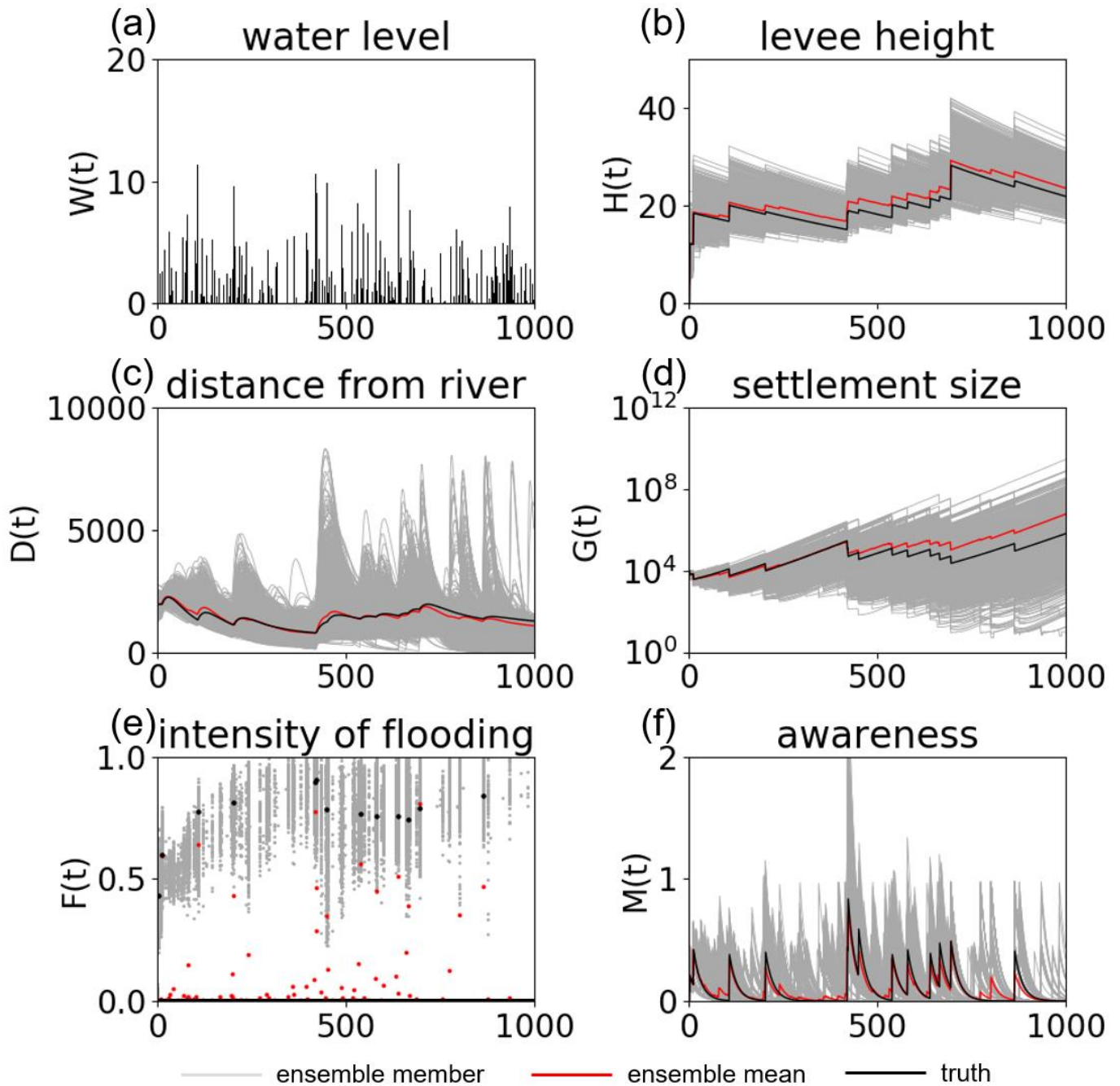



Figure 1. Timeseries of (a) high water level W(t), (b) the flood protection level (or levee height) H(t), (c) the

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

763   3.1.1). The time step is annual. Grey, red, and black lines are the ensemble members, their mean, and the

764   synthetic truth, respectively.

765

766

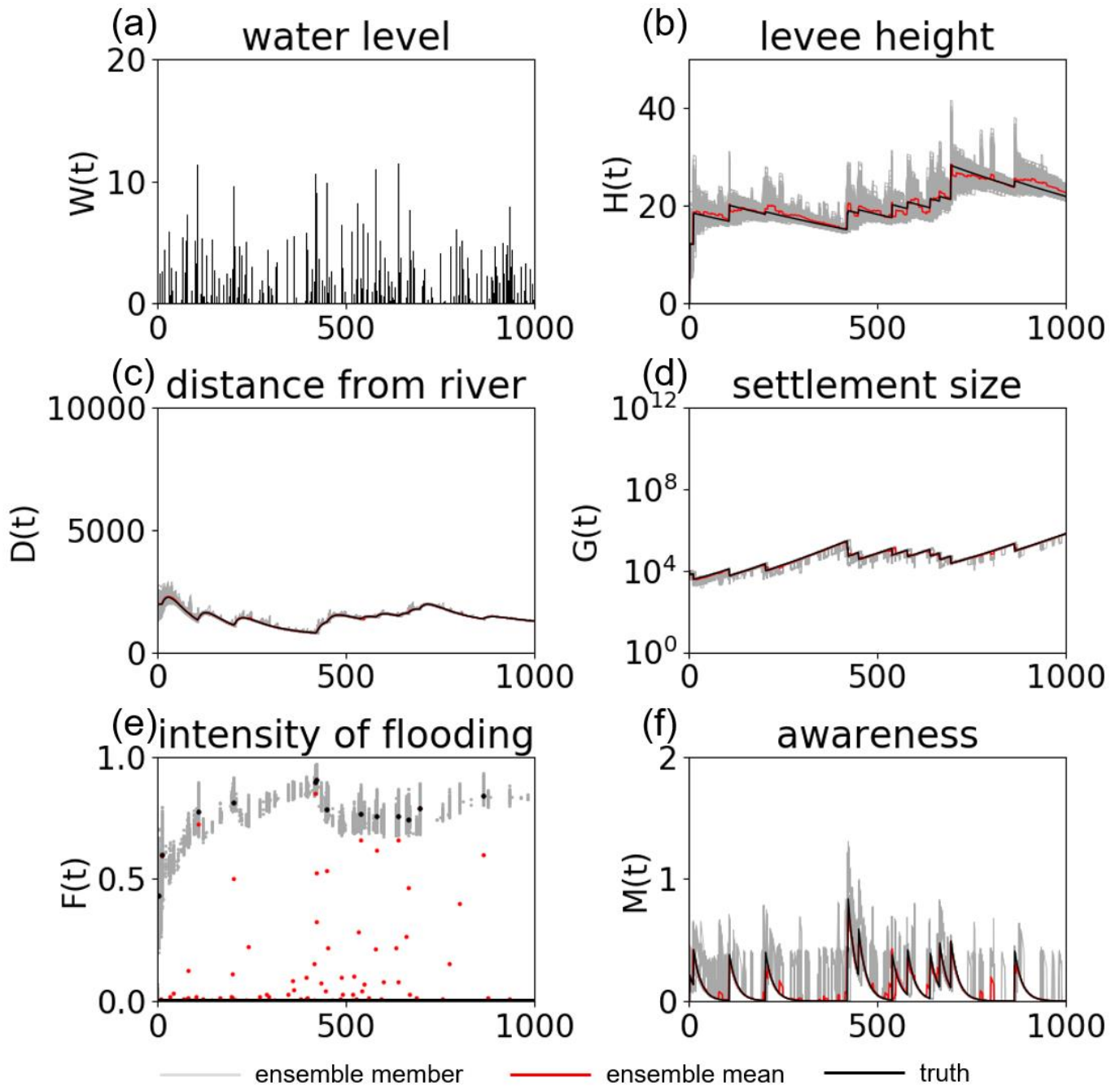



Figure 2. Timeseries of (a) high water level W(t), (b) the flood protection level (or levee height) H(t), (c) the

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

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

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

773

774

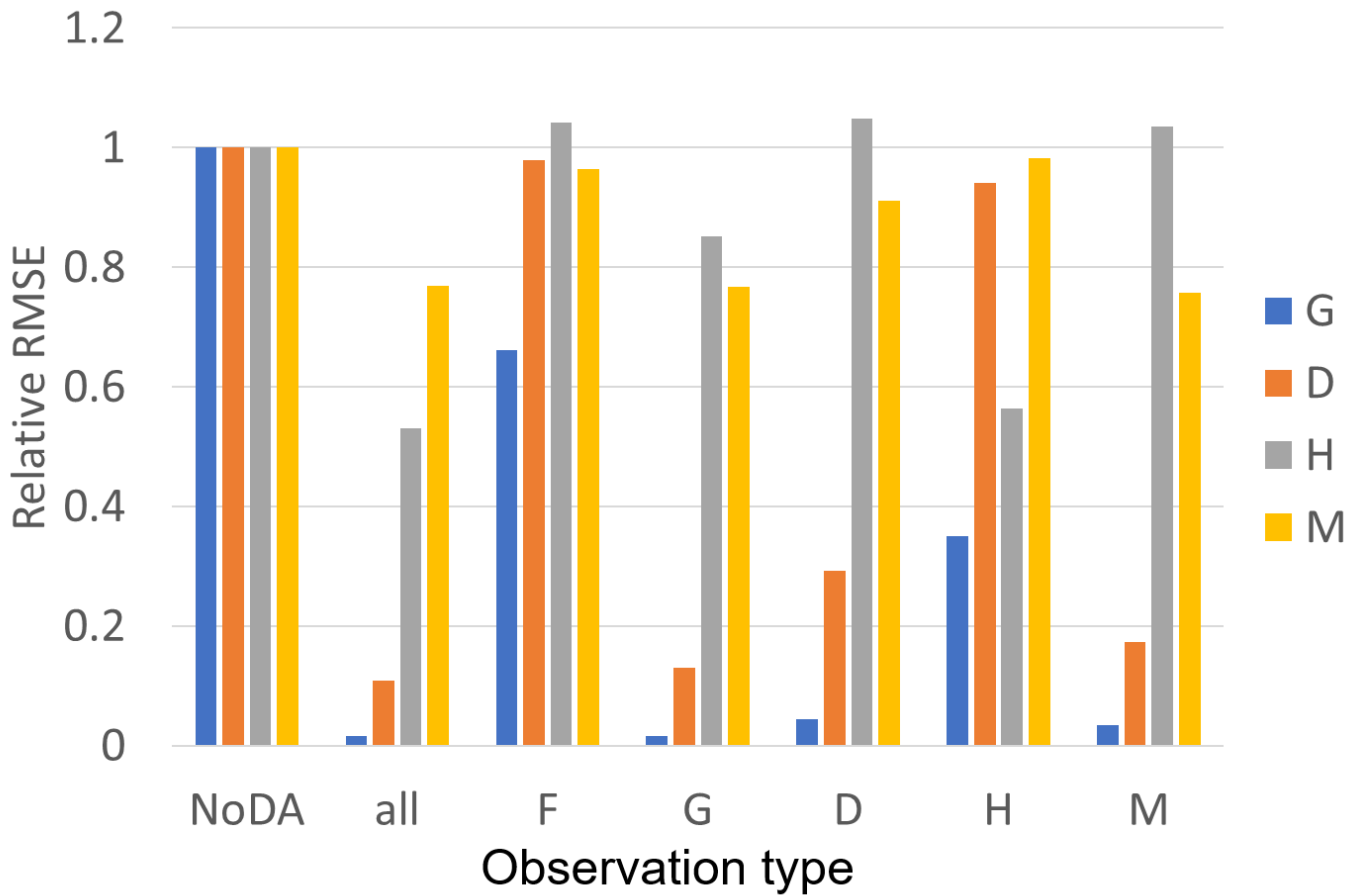



Figure 3. The ratio of RMSEs of the no data assimilation experiment (NoDA) to those of the data assimilation

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 1 (see section 3.1.1). 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 $M(t)$.

780

781

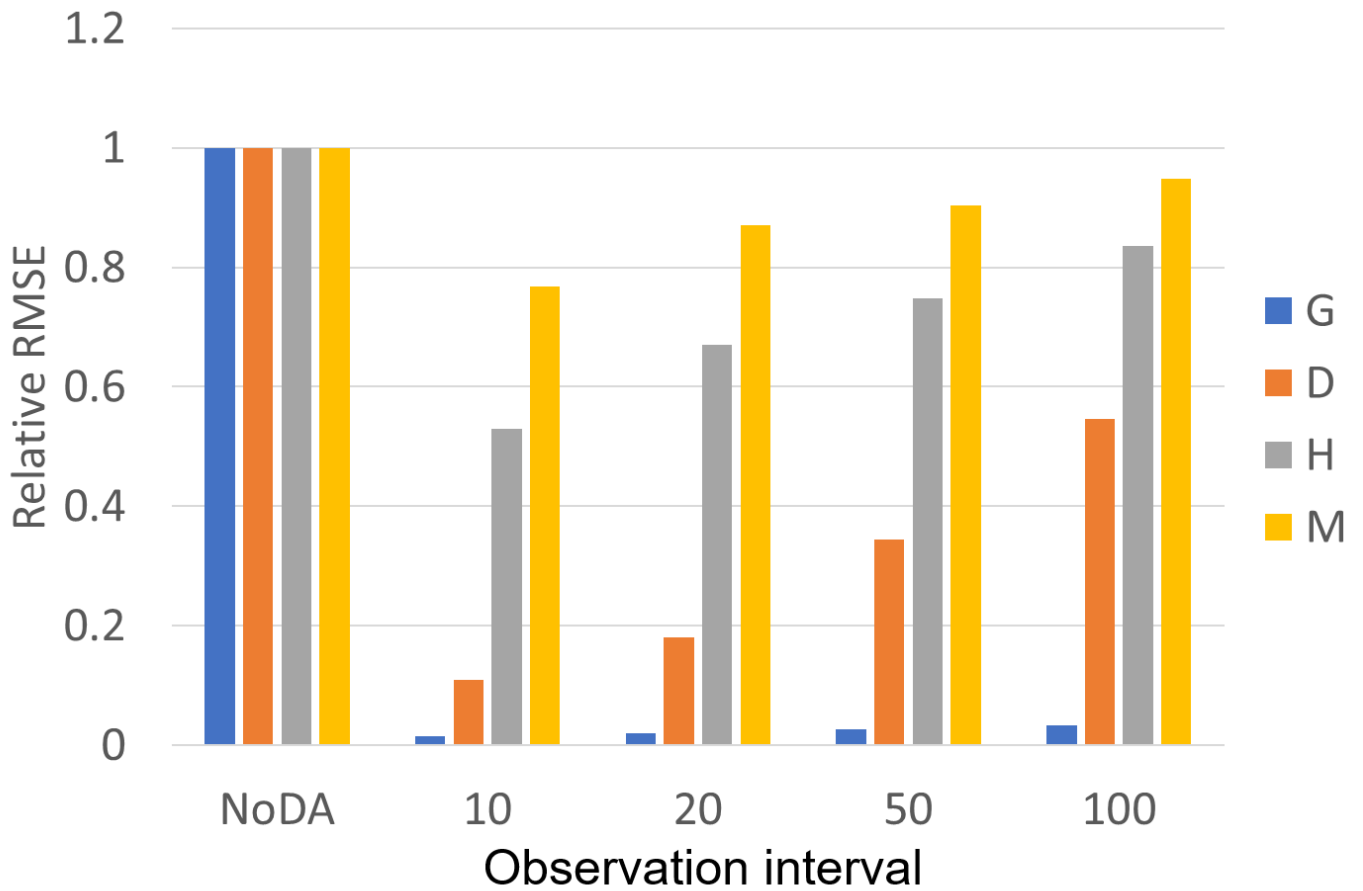



Figure 4. The ratio of RMSEs of the no data assimilation experiment (NoDA) to those of the data assimilation

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

in the experiment 1 (see section 3.1.1). 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).

787

788

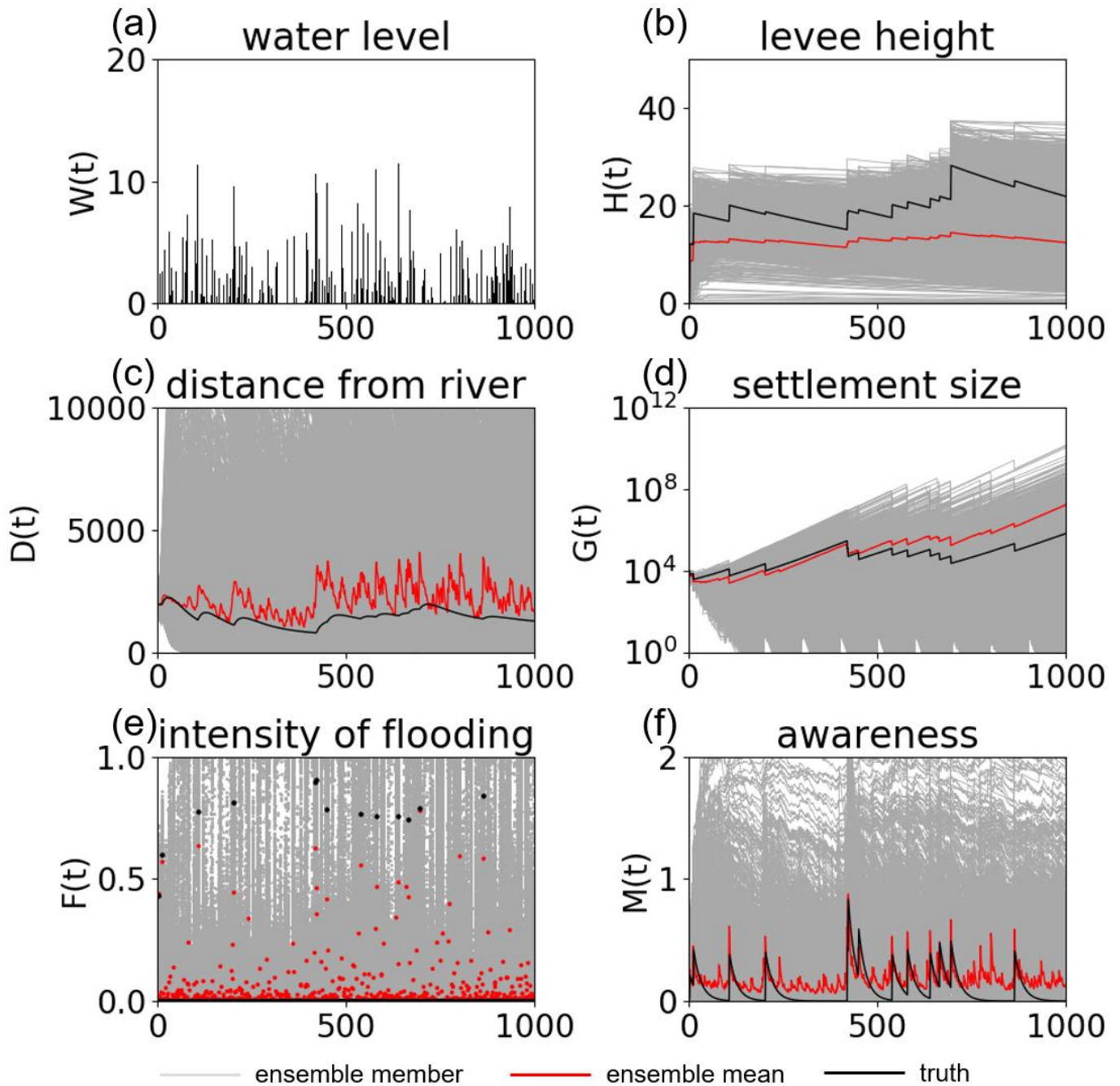



Figure 5. Timeseries of (a) high water level W(t), (b) the flood protection level (or levee height) H(t), (c) the

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 2 (see section

793 3.1.2). The time step is annual. Grey, red, and black lines are the ensemble members, their mean, and the

794 synthetic truth, respectively.

795

796

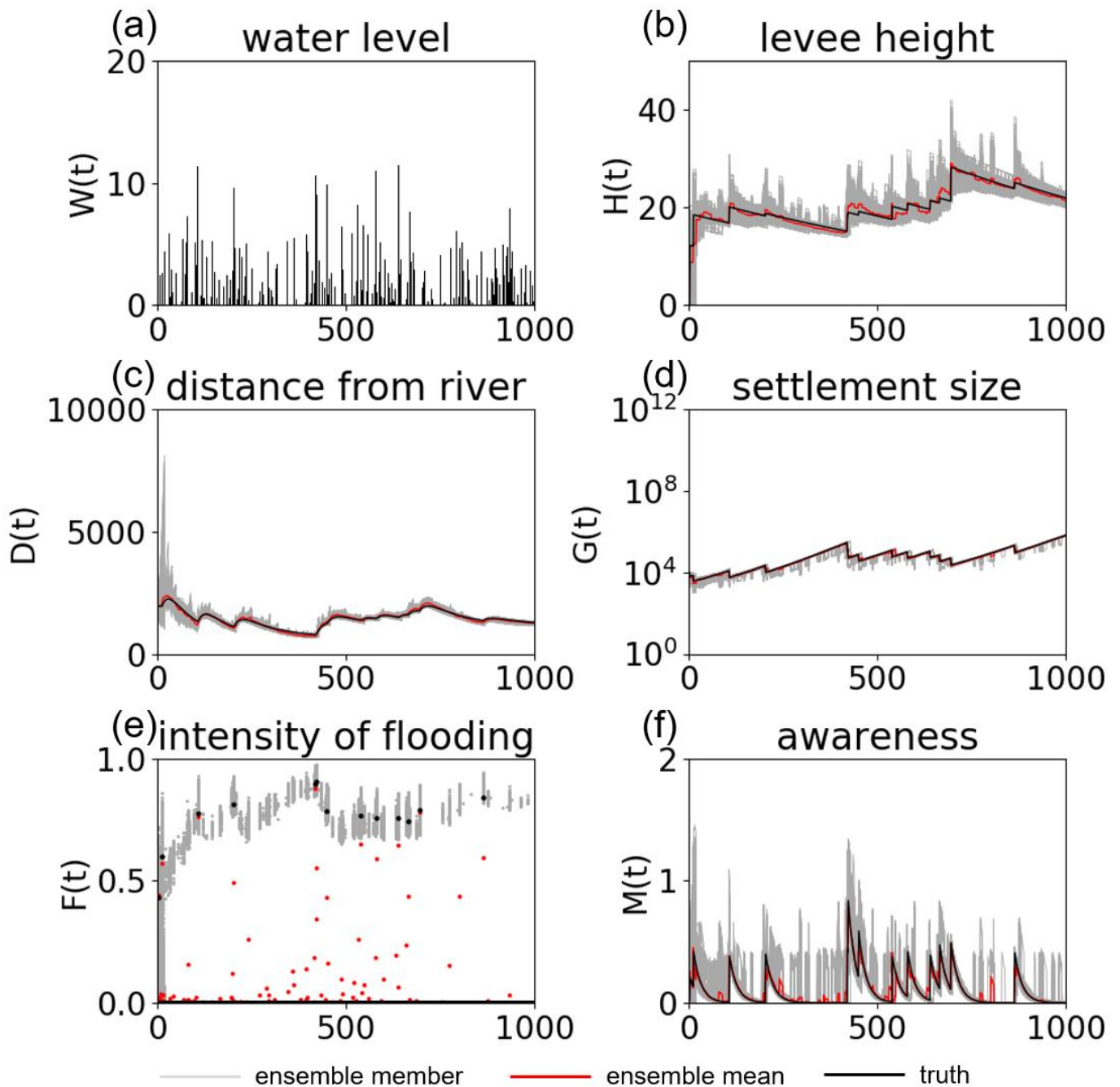



Figure 6. Timeseries of (a) high water level W(t), (b) the flood protection level (or levee height) H(t), (c) the

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

802   every 10 years with 5000 ensembles in the experiment 2 (see section 3.1.2). The time step is annual. Grey, red,

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

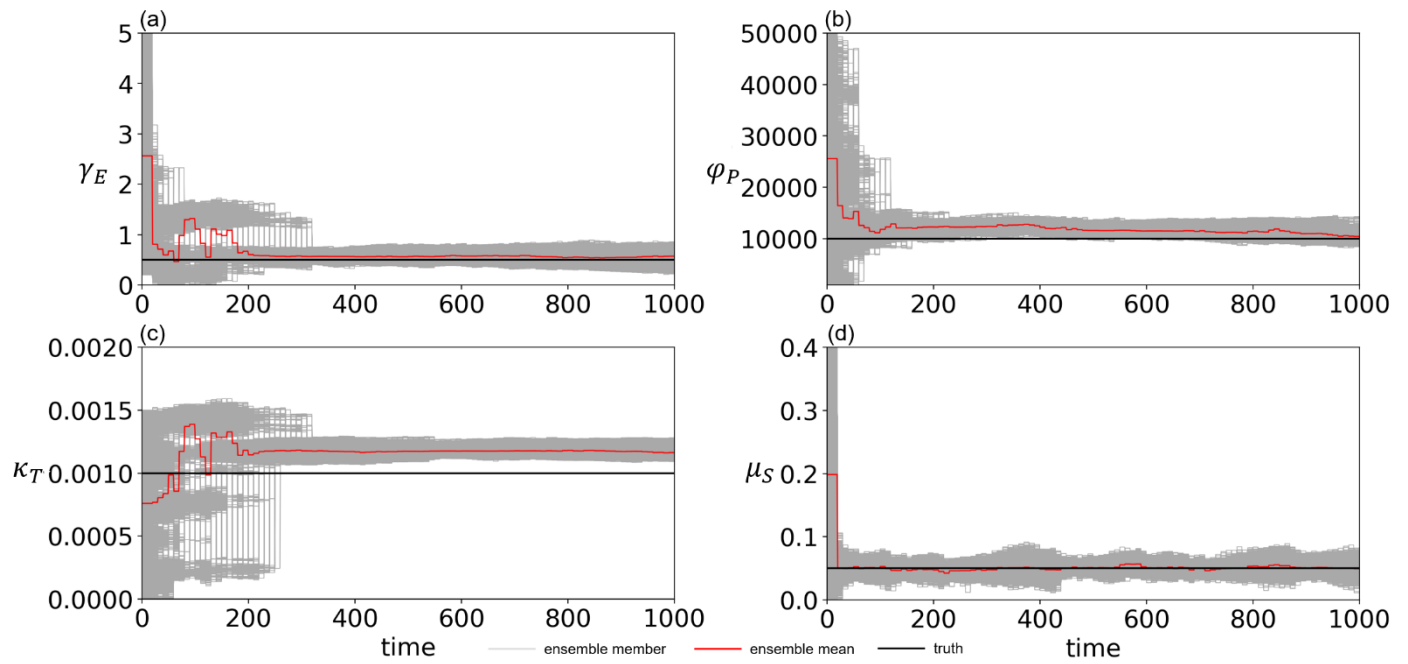



Figure 7. Timeseries of (a) the cost of levee raising γ_E, (b) the rate by which new properties can be built φ_P,

(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.

811

812

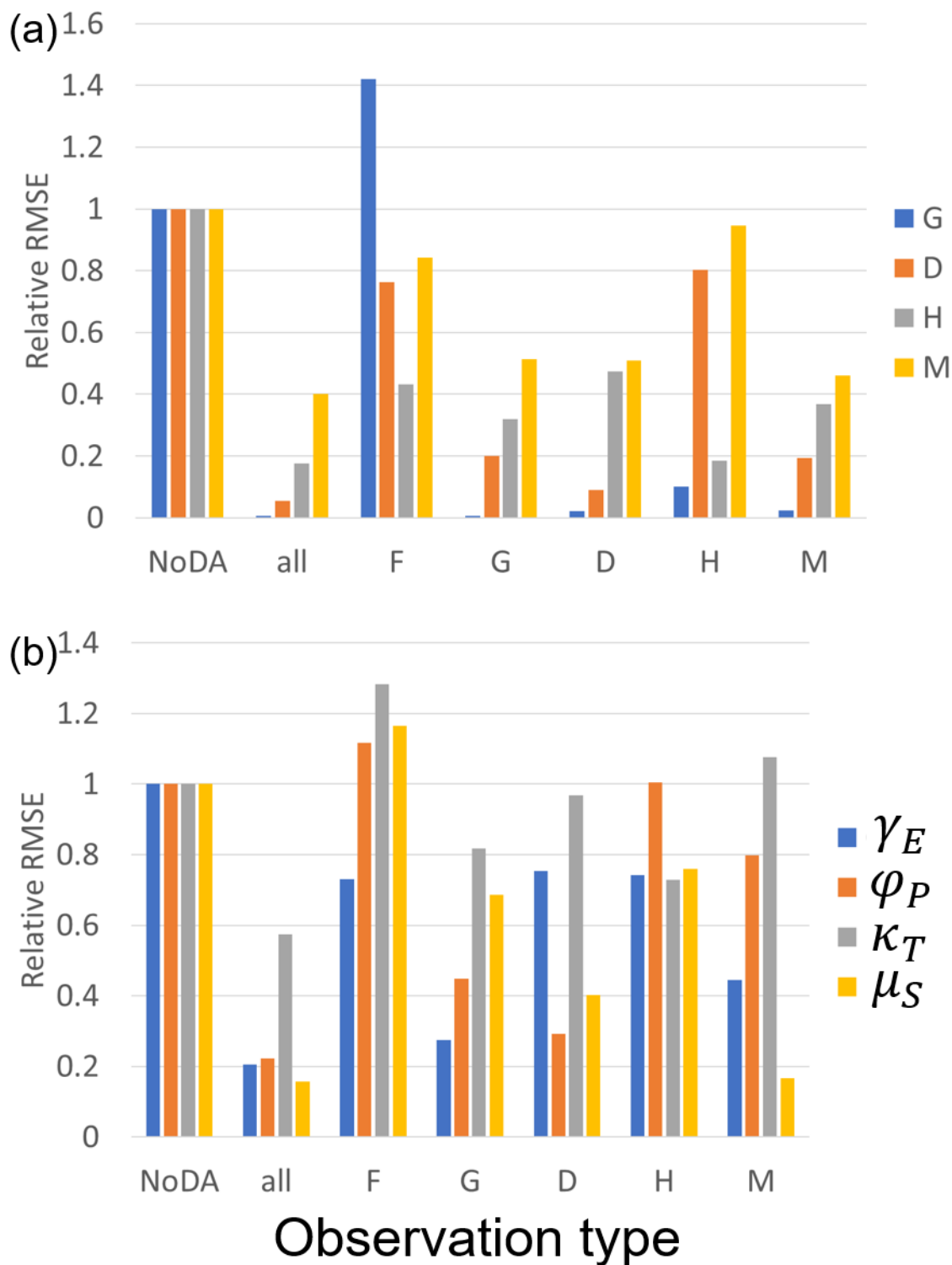



Figure 8. The ratio of RMSEs of the no data assimilation experiment (NoDA) to those of the data assimilation

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

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

817   protection level (or levee height) H(t), and the social awareness of the flood risk M(t). (b) Blue, orange, gray,

818   and yellow bars are RMSEs of the cost of levee raising γ_E, the rate by which new properties can be built φ_P,

819   the rate of decay of levees κ_T, memory loss rate μ_S.

820

821

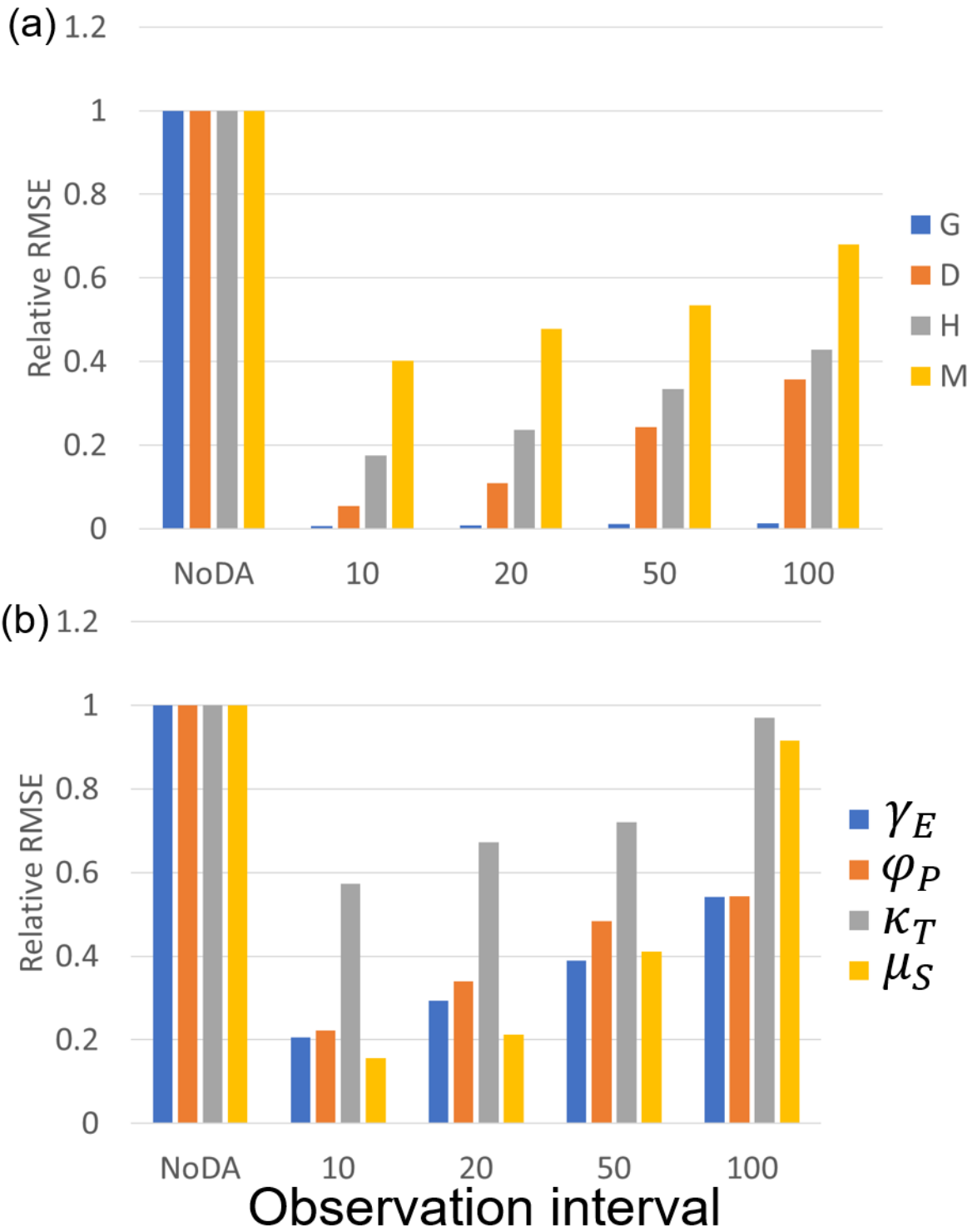



Figure 9. The ratio of RMSEs of the no data assimilation experiment (NoDA) to those of the data assimilation

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

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

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

827 level (or levee height) H(t), and the social awareness of the flood risk M(t). (b) Blue, orange, gray, and yellow

828 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

829 decay of levees κ_T, memory loss rate μ_S.

830

831

832

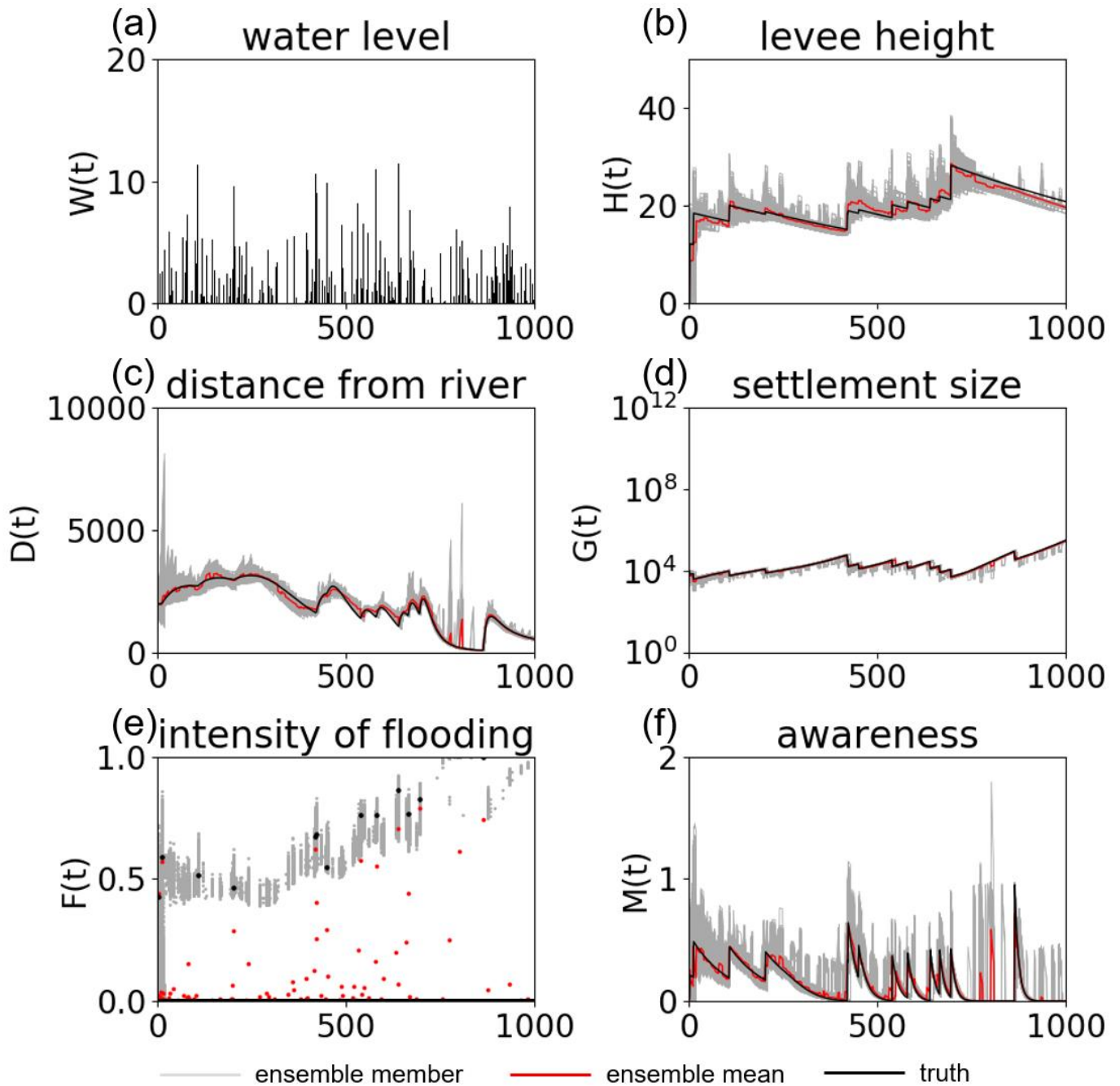



Figure 10. Timeseries of (a) high water level W(t), (b) the flood protection level (or levee height) H(t), (c) the

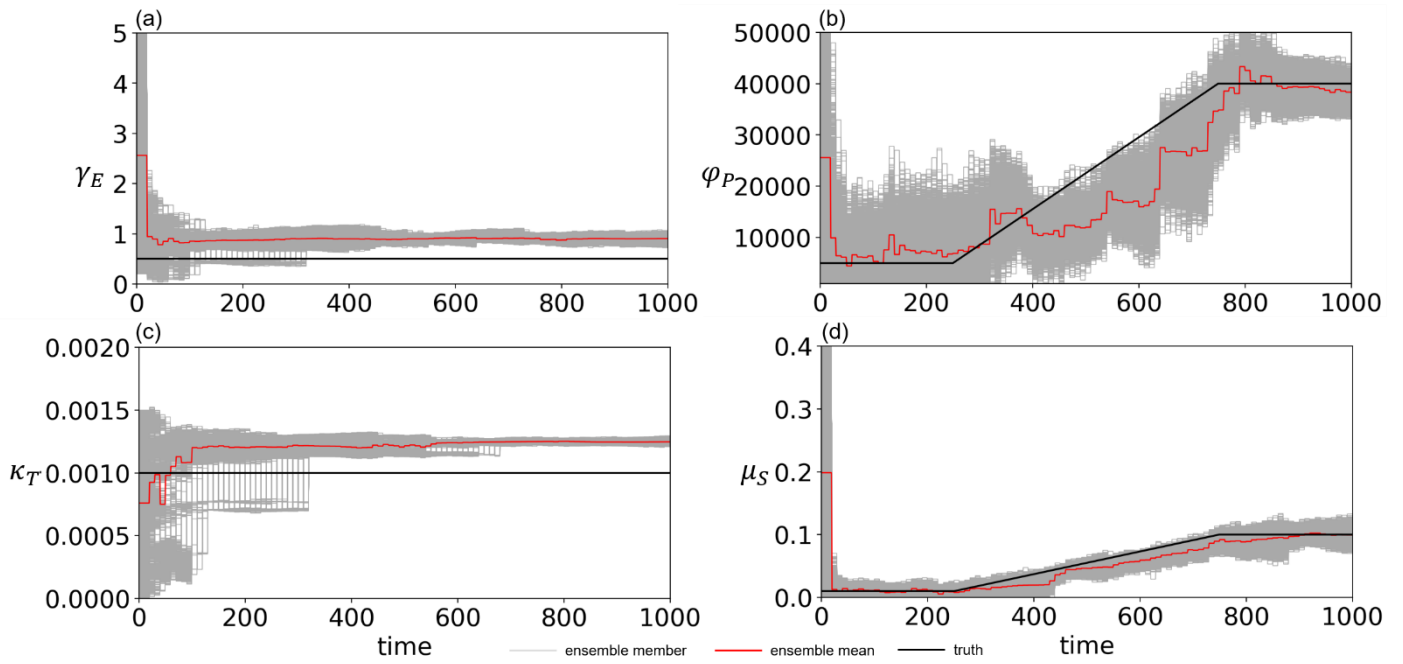
distance of the center of mass of the human settlement from the river D(t), (d) the size of the human settlement

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

837 every 10 years with 5000 ensembles in the experiment 3 (see section 3.1.3). The time step is annual. Grey, red,

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

839

840 **Figure 11.** Timeseries of (a) the cost of levee raising γ_E, (b) the rate by which new properties can be built

841 φ_P, (c) the rate of decay of levees κ_T, (d) memory loss rate μ_S estimated by the data assimilation of all

842 observations (F, G, D, H, and M) with 5000 ensembles every 10 years in the experiment 3 (see section 3.1.3).

843 The time step is annual. Grey, red, and black lines are the ensemble members, their mean, and the synthetic

844 truth, respectively.

845

846

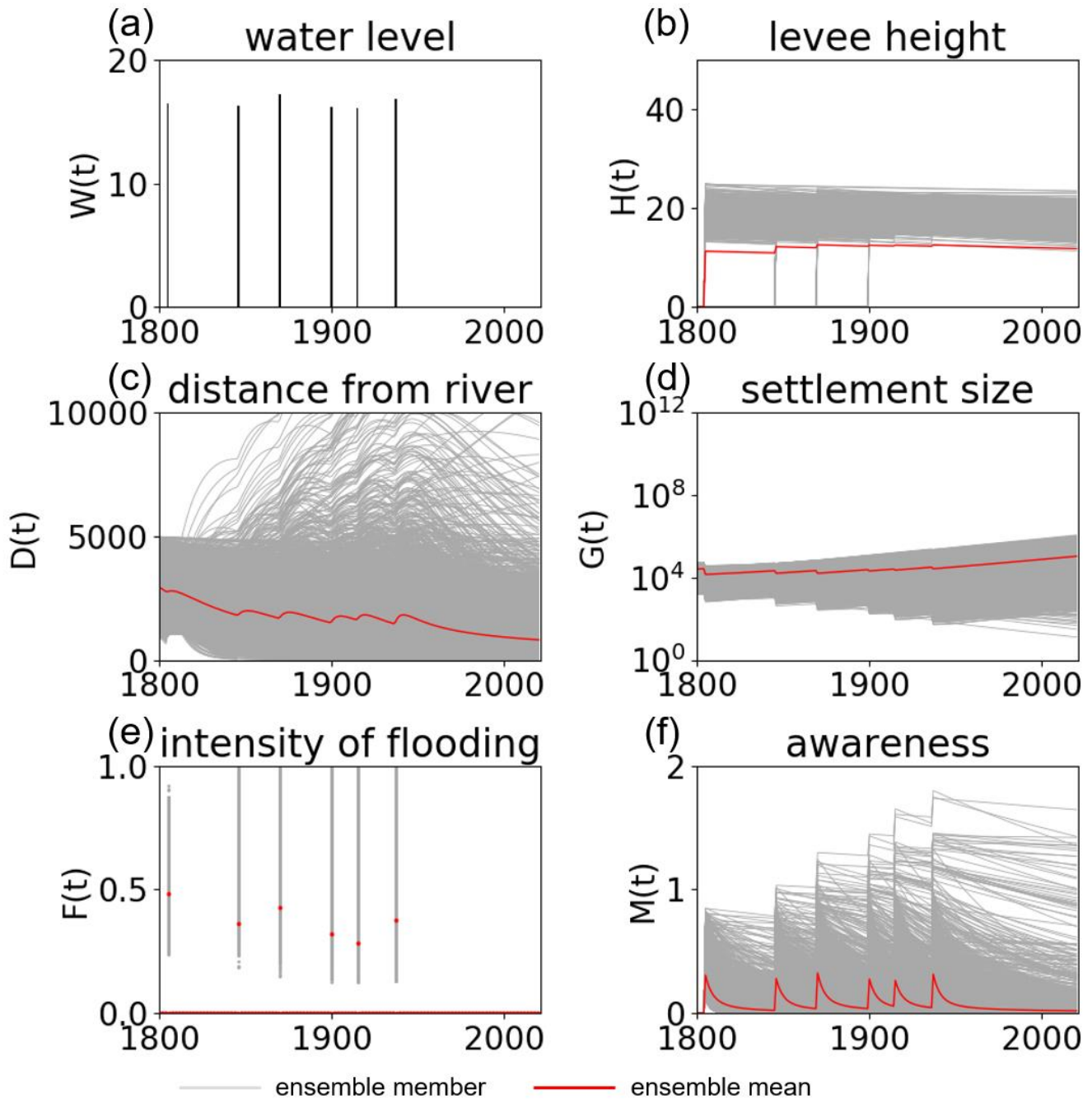



Figure 12. Timeseries of (a) high water level W(t), (b) the flood protection level (or levee height) H(t), (c) the

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

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

853 respectively.

854

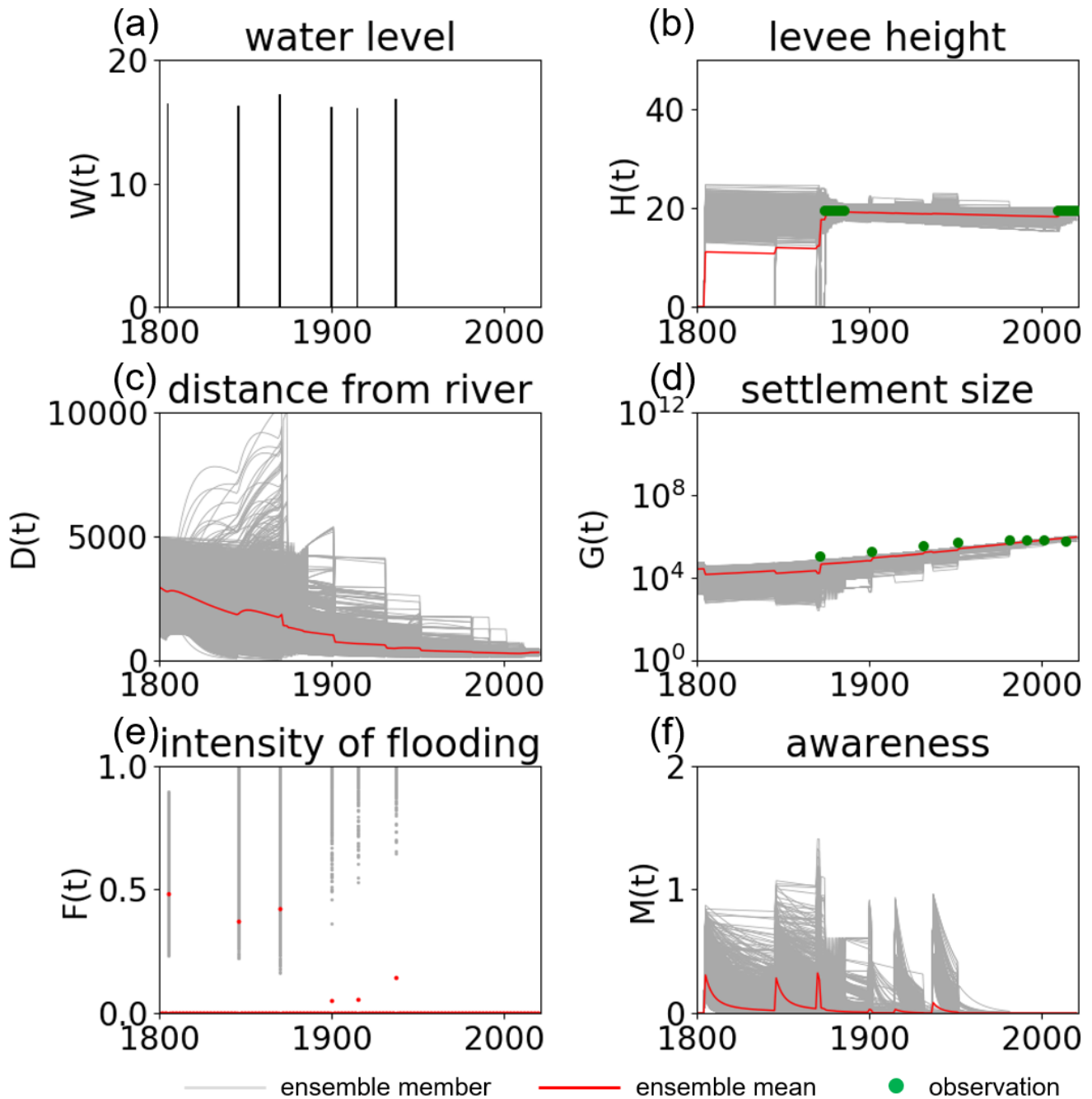



Figure 13. Timeseries of (a) high water level W(t), (b) the flood protection level (or levee height) H(t), (c) the

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 real-world observations of G and H (green dots) are assimilated

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

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

862

863

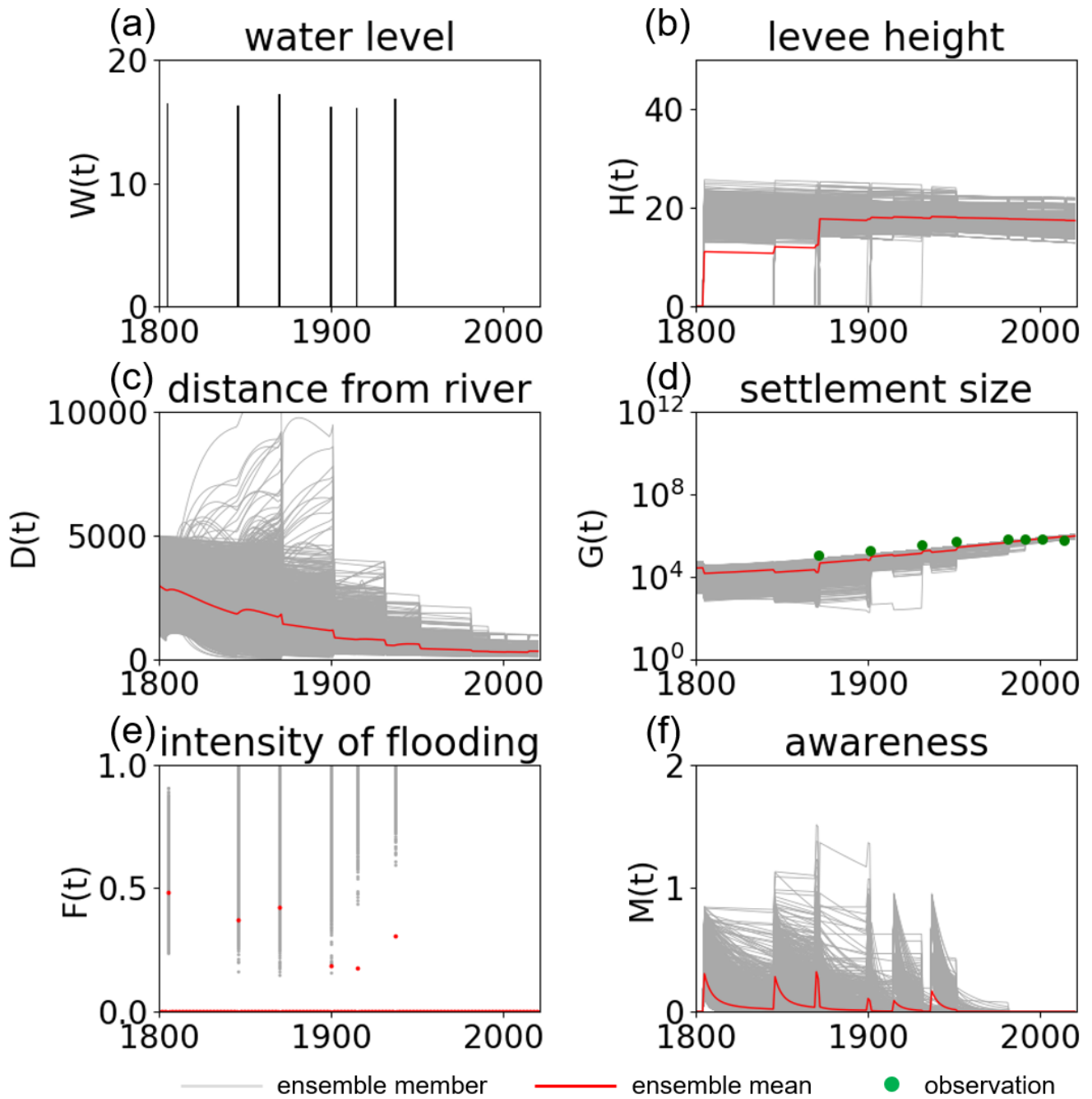



Figure 14. Same as Figure 13 but only real data of G are assimilated.

865

866

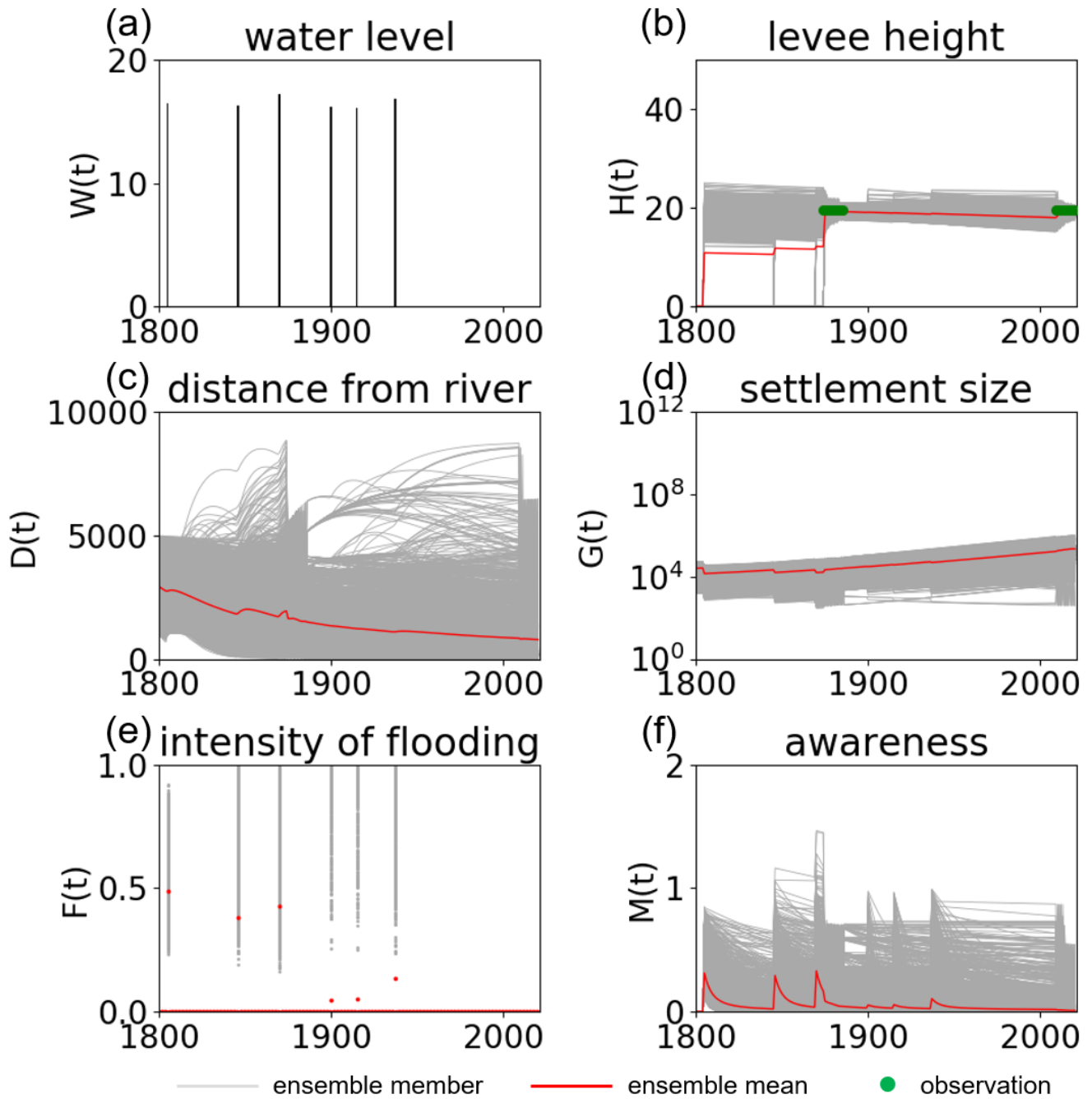



Figure 15. Same as Figure 13 but only real data of H are assimilated.

868