Supplementary information for:
‘A downscaling approach to compare COVID-19 count data from databases aggregated at different spatial scales’

Andre Python, Andreas Bender, Marta Blangiardo, Janine B. Illian, Ying Lin, Baoli Liu, Tim Lucas, Siwei Tan, Yingying Wen, Davit Svanidze, Jianwei Yin

1. Introduction

This Section provides details on the data extraction and cleaning processes that has been performed to compare and select data on coronavirus. In addition, it provides further detail on the cross-validation procedure used to compare the predictive performance of disaggregated models. All operations (both computing and graphics) are carried out in the statistical software R [1] and can be downloaded from the Harvard dataverse platform (`https://doi.org/10.7910/DVN/SMGG9R`).

2. COVID-19 data selection

Data on coronavirus-infected cases are provided by various sources, including individual-level data from national, provincial, and municipal health reports, as well as additional information from online reports. A reference source of province-level data is provided by John Hopkins University CSSE (JHU) [2] through an interactive web platform, whose data can be downloaded using GitHub [3]. For cases in China, the sources used include Twitter feeds, online news services. The reported cases are confirmed with regional and local health departments, including centers for disease control and prevention of China, Taiwan, and Europe, the Hong Kong Department of Health, the Macau Government, and the World Health Organization (WHO), as well as city-level and state-level health authorities.

For the disaggregated datasets, we extracted city-level observations from the GitHub platform of the news agency Pengpai, which we refer to as "The Paper" [4,5] (09 April 2020). We added city names in English, spatial coordinates of the cities (centroid), and dates in English. Furthermore, we removed events without or with wrong spatial coordinates. The cleaned dataset counts observations from 11 January 2020 to the end of February 2020.

In addition to The Paper dataset, we extracted data from healthmap.org's GitHub (09 April 2020), which includes COVID-19 cases in China in Hubei province [6] and outside Hubei province [7]). We refer to this second dataset as Xu et al. data. The entire dataset (events without dates are removed) contains observations from January 18 to end of February 2020. The authors have provided the geolocalization of the data using google map and a variable that indicates the level of spatial accuracy, which can be used to subset the data. Xu et al. data has been aggregated through official government sources, peer-reviewed papers, and online reports. Various procedures have been applied by the authors to increase accuracy and comprehensiveness of the data. Therefore, we did not clean the original dataset since the provider used various robust procedures to ensure that the data is accurate and comprehensible, which include checking records and potential duplicates with peer-reviewed research articles.

3. Model validation

3.1 Out-of-sample predictions

Good in-sample performance may be the results of overfitting and are therefore not necessarily informative about the true performance of models. In order to assess the external validity of our model and to obtain more realistic estimations of performance metrics, we performed an out-of-sample cross-validation procedure by taking into account the spatial nature of the data and the model specifications[8–10]. While methods to evaluate out-of-sample performance of downscaling approaches have been developed (e.g.[11,12]), their relevance has been essentially drawn from the results of studies based on simulated data, which may not necessarily apply to our case study. Also, since province-level is the finer spatial resolution of our reference data (JHU data[2]), a cross-validation approach based on spatial blocks[13] would not be applicable.

As a result, we fit the downscaling models using data in all provinces except one (hold-out province), predicted the expected cases in the hold-out province, and reported performance metrics (RMSE, MAE) based on the expected cases from JHU. Through an iterative process, the data from the hold-out province is not used during the fitting process to ensure that the predictive performance of the models is assessed exclusively on new data. We performed the out-of-sample procedure on various model specifications, similarly to the in-sample procedure except that we did not consider models with i.i.d random polygon-specific effects since we opted for a procedure that remove data from province in an iterative way. In this case, province-specific effects cannot be used to inform the model.

The model specifications (*Model spec*) are the following: (1) include all covariates, (2) include only anthropogenic covariates (*Socio. cov.*), (3) include only environmental covariates (*Env. cov.*), (4-7) using alternative penalised complexity (PC) priors on the spatial parameters ρ *min* and σ_ζ *max*. In total, this cross-validation computes predictive performance metrics for a total of 33 provinces multiplied by 7, which corresponds to a total of 231 runs. The results are illustrated in Table S1.

In Table S1, we report the MAE and RMSE (*MAE all, RMSE all*) for each model specification, based on the out-of-sample predictions in all provinces. Since the out-of-sample predictions in some provinces appear far more challenging in some provinces (see Xingiang and Hubei in Fig S5), we also present the MAE and RMSE results with predictions excluded in Xingjiang (*MAE no Xingjiang, RMSE no Xingjiang*) and Hubei (*MAE no Hubei, RMSE no Hubei*) to compare the model performance in different areas.

To select the model we favor the model performance metric RMSE over MAE since large errors are particularly undesirable in our context. Table S1 shows that the model using only anthropogenic covariates (mod spec.: *soc*) has a better out-of-sample predictive ability. It has the lowest RMSE value overall (when all provinces are included) and also when we exclude predictions in Hubei province. If we exclude predictions in Xingjiang province, model *sigma2* with all covariates and alternative penalised complex-

mod. spec.	MAE all	RMSE all	MAE no Xingjiang	RMSE no Xingjiang	MAE no Hubei	RMSE no Hubei
all	10974.7	48166.9	2910.9	11443.3	9338.3	47614.8
soc	4480.0	13392.6	3761.6	12703.9	2620.6	7552.5
env	5515.6	19581.3	2803.7	11366.5	3705.8	16422.0
rho1	10456.4	45445.9	2878.5	11418.8	8803.7	44771.5
rho2	8665.0	36137.1	2766.3	11345.0	6956.1	34947.0
sigma1	17617.4	82994.3	3421.0	12008.9	16188.2	83533.7
sigma2	3805.9	13860.3	2423.4	11223.7	1945.9	8531.8

Table S1: **Performance metrics computed at province-level from an out-of-sample procedure.** Performance metrics computed from an out-of-sample procedure (iteratively remove each of the 33 provinces) to compare different model specifications used to predict COVID-19 counts using a downscaling approach. We compare a total of seven model specifications (without i.i.d random effects): (1) include all covariates (*all*), (2) include only anthropogenic covariates (*soc*), (3) include only environmental covariates (*env*), and models that include all covariates but using alternative penalised complexity (PC) priors on the spatial parameters (*rho1,rho2,sigma1,sigma2*).

ity (PC) prior on the spatial parameter $sigma_{max} = 10$ shows the lowest RMSE values. Therefore, we select model *soc*, which has the best predictive performance overall.

Model spec.	Socio. cov.	Env. cov.	iid. random	ρ min.	σ_ζ max.	Nobs Xu	Nobs Paper	MAE Xu	MAE Paper	RMSE Xu	RMSE Paper
1	yes	yes	yes	1	5	301	234	183.3	242.6	980.3	2644.5
2	yes	no	yes	1	5	301	234	186.0	254.0	861.0	2889.8
3	no	yes	yes	1	5	301	234	200.3	359.4	801.7	3733.4
4	yes	yes	no	1	5	301	234	192.8	149.7	2382.4	945.1
5	yes	yes	yes	10	5	301	234	183.0	240.0	995.8	2620.2
6	yes	yes	yes	5	5	301	234	182.9	239.8	996.6	2619.0
7	yes	yes	yes	1	10	301	234	182.9	240.0	996.3	2619.4
8	yes	yes	yes	1	2	301	234	183.6	248.2	942.7	2705.6

Table S2: **Performance metrics computed at district-level with several model specifications.** Performance metrics computed to compare the discrepancies on the COVID-19 counts of the spatially disaggregated datasets from Xu et al. and The Paper with predicted (mean) values from the downscaling approach at district-level (340 districts) using different model specifications.

3.2 In-sample predictions

Note that the out-of-sample predictions approach that remove in an iterative way each province (hold-out province) cannot be used to assess the predictive performance of models with province-specific i.i.d random effects. However, random effects can account for important differences we observe among provinces that cannot be informed by the covariates. Therefore, our final selected model is based on the selected specification (model using only anthropogenic covariates (mod spec.: *soc*)), by including an i.i.d random effects to account for province-specific characteristics.

Province	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Obs.rate	0.016	0.020	0.019	0.009	0.003	0.014	0.005	0.004	0.021	0.004	0.012	0.013	0.053	1.098	0.015	0.000
Pred.rate	0.016	0.020	0.019	0.009	0.003	0.014	0.005	0.004	0.021	0.004	0.012	0.013	0.052	1.098	0.015	0.000

Province	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
Obs.rate	0.008	0.020	0.003	0.003	0.011	0.003	0.006	0.008	0.015	0.004	0.006	0.002	0.010	0.003	0.000	0.004	0.022
Pred.rate	0.008	0.020	0.003	0.003	0.011	0.003	0.006	0.008	0.015	0.004	0.006	0.002	0.010	0.003	0.000	0.004	0.022

Table S3: **Results of in-sample predictions of the incidence rates (per 1000) at province-level.** Comparison of the observed incidence rate (per 1000) at province-level (*second row*) with the average predicted incidence rate (per 1000) (*third row*). The predictions are carried out in a total of 33 Chinese provinces are included.

To assess the in-sample predictive ability of the model, we compare the predicted incidence rate at province-level with the observed incidence rate computed as the number of reported COVID-19 cases from JHU[2] divided by the population size[14] of the province. For each province, the average incidence rate predicted by the downscaling is given by the sum of the mean predicted incidence rate multiplied by population size for each grid-cell within the province. A brief look at the results in Table S3 indicates that in all provinces, the predicted incidence rate (per 1000) is equal to the observed incidence rate (per 1000) within 2 decimals. We can conclude that the in-sample predictions of the model are of sufficient accuracy.

3.3 Model convergence assessment

The disaggregation model necessitates relatively few evaluations of the posterior density compared to MCMC methods since the former only requires to maximise the posterior and identify the posterior mode. However, a faster approximation may potentially reduce the accuracy of the estimation of the posterior samples[15]. Here, we assess the convergence of the selected downscaling model by comparing the parameter estimation with a self-tuning variant of Hamiltonian Monte Carlo (HMC) method, called no-U-turn sampler (NUTS)[16] implemented in the R package `tmbstan`[17]. The main principle is to run multiple Markov chains that are randomly initialised, remove warm-up samples, split into half the remainder of each chain to detect non-stationarity in each individual chain, and assess model convergence using the potential scale reduction statistic $\hat{R}$. This metric computes the ratio of the average variance of samples within each chain to the variance of the pooled samples across chains. A value of $\hat{R}$ below 1.1 suggests that the distribution of the chains has converged[18].

NUTS is applied to the `TMB` output object resulting from the downscaling model with a number of iterations set to 800 (half of them used as warm-up) for 6 chains. Based on the Hamiltonian dynamics, a new momentum vector is sampled for a given number of iterations, which updates the current state of the parameters using the leapfrog integrator with discretization time (step size) and number of leapfrog steps (nb.leapfrog). To

account for numerical errors during integration, a Metropolis acceptance step is required to to either accept or reject a particular step. The standard Metropolis acceptance probability (acceptance) has values close to 1 to improve sampling efficiency and is computed from multinomial sampling over the states for each Hamiltonian trajectory. A cap on the tree depth is set to 10 (default value), so that nb.leapfrog is capped to $2^{10} - 1 = 1023$ in order to reduce the work while keeping a sufficiently long trajectory for each iteration.[16].

	NUTS			Disaggregation	
Parameters	mean	std	$\hat{R}$	mean	std
$\beta_{W.access}$	-0.52	0.22	1.000	-0.55	0.21
β_{access}	-0.37	0.33	1.000	-0.39	0.32
intercept	-4.69	2.10	1.001	-4.74	2.08
spatial effect ($\log(\sigma)$)	0.30	0.024	1.001	0.34	0.48
spatial effect ($\log(\rho)$)	5.55	0.83	1.001	5.11	0.54
i.i.d. effects ($\log(\tau)$)	-0.66	0.32	1.002	-0.54	0.31

Table S4: **Convergence assessment: summary.** We compare the parameter estimation (mean and standard deviation) on the final model specifications between a Hamiltonian Monte Carlo (HMC) method: no-U-turn sampler (NUTS)[16] and the posterior approximation method used in the disaggregation model. The disaggregation model took about 25 minutes to run. NUTS took about 24 hours to run for six chains with 800 iterations each (Win 10, CPU i5/3GHz, 6 cores (1 chain per core)). NUTS convergence metric is provided by $\hat{R}$.

Table S4 compares results of the selected disaggregation model with the results of the NUTS (800 iterations, 6 chains). The sign and order of magnitude of the mean and standard deviation of the parameters are consistent between the NUTS and the posterior approximation method used in the disaggregation model. Fig S6 shows the histogram of the NUTS estimation of the parameters. The convergence of NUTS is reached, with $\hat{R} < 1.003$ obtained for the estimation of the intercept, covariate coefficients ($\beta_{W.access}$ and β_{access}), the i.i.d. effects ($\log(\tau)$), the standard deviation ($\log(\sigma)$) and range ($\log(\rho)$) of the spatial effect. Fig S7 illustrates the trace plot for the NUTS parameter estimation. It took about 24 hours to run 6 chains with 800 iterations per chain using the following computer specifications: Windows 10, CPU i5/3GHz, 6 cores (1 chain per core).

References

[1] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2008. URL: http://www.R-project.org.

[2] E. Dong, H. Du, L. Gardner, An interactive web-based dashboard to track COVID-19 in real time, The Lancet infectious diseases 20 (2020) 533–534.

[3] Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE), Coronavirus data at province level provided via GitHub, https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv, 2020. Accessed: 2020-04-09.

[4] Pengpai News agency, The Paper & Sixth Tone data, https://www.thepaper.cn, 2020. Accessed: 2020-02-01.

[5] Pengpai News agency, Coronavirus (COVID-19) data provided via GitHub, https://github.com/839Studio/Novel-Coronavirus-Updates/blob/master/Updates_NC.csv, 2020. Accessed: 2020-04-09.

[6] Xu et al., Coronavirus data for Hubei province provided via GitHub, https://github.com/beoutbreakprepared/nCoV2019/blob/master/ncov_hubei.csv, 2020. Accessed: 2020-04-09.

[7] Xu et al., Coronavirus data for all regions in the world except Hubei province provided via GitHub, https://github.com/beoutbreakprepared/nCoV2019/blob/master/ncov_outside_hubei.csv, 2020. Accessed: 2020-04-09.

[8] R. K. Heikkinen, M. Marmion, M. Luoto, Does the interpolation accuracy of species distribution models come at the expense of transferability?, Ecography 35 (2012) 276–288.

[9] S. J. Wenger, J. D. Olden, Assessing transferability of ecological models: an underappreciated aspect of statistical validation, Methods in Ecology and Evolution 3 (2012) 260–267.

[10] M. Trachsel, R. J. Telford, Technical note: Estimating unbiased transfer-function performances in spatially structured environments, Climate of the Past 12 (2016) 1215–1223.

[11] K. Wilson, J. Wakefield, Pointless spatial modeling, Biostatistics 21 (2020) e17–e32.

[12] H. C. Law, D. Sejdinovic, E. Cameron, T. Lucas, S. Flaxman, K. Battle, K. Fukumizu, Variational learning on aggregate outputs with Gaussian processes, in: Advances in Neural Information Processing Systems, 2018, pp. 6081–6091.

[13] D. R. Roberts, V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schröder, W. Thuiller, et al., Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, Ecography 40 (2017) 913–929.

[14] A. J. Tatem, Worldpop, open data for spatial demography, Scientific Data 4 (2017) 1–4.

[15] A. K. Nandi, T. C. D. Lucas, R. Arambepola, P. Gething, D. J. Weiss, disaggregation: An R Package for Bayesian Spatial Disaggregation Modelling, arXiv preprint (2020).

[16] M. D. Hoffman, A. Gelman, The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo, J. Mach. Learn. Res. 15 (2014) 1593–1623.

[17] C. C. Monnahan, K. Kristensen, No-U-turn sampling for fast Bayesian inference in ADMB and TMB: Introducing the adnuts and tmbstan R packages, PloS one 13 (2018) e0197954.

[18] A. Gelman, D. B. Rubin, et al., Inference from iterative simulation using multiple sequences, Statistical Science 7 (1992) 457–472.

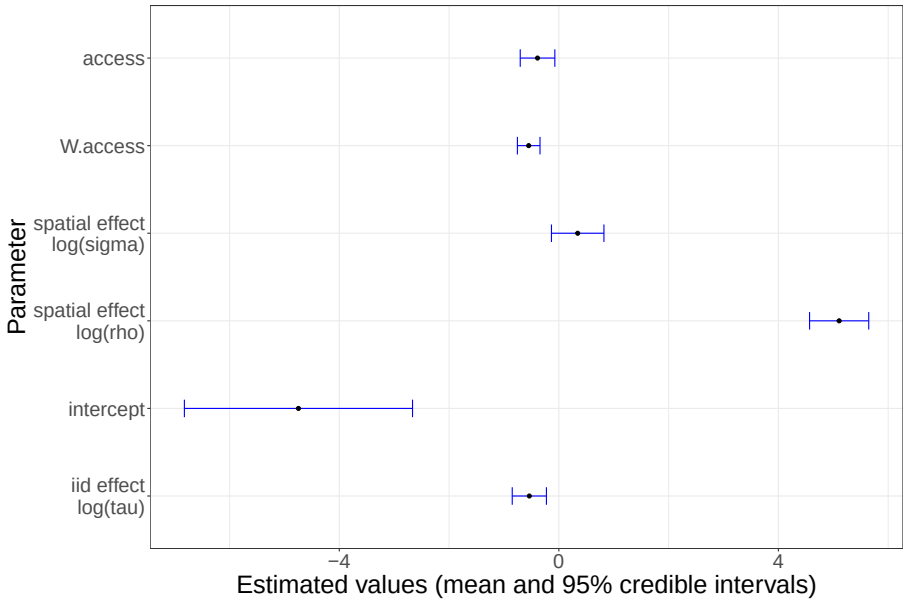

Figure S1: **Parameter estimation of the downscaling model.** The graphic shows the mean and 95% credible intervals (x-axis) of the estimated parameters (fixed-effects and random effects) in the model (y-axis). It includes (from bottom to top), the log precision of the i.i.d effects (polygon-level), intercept, spatial hyperparameters log ρ and log σ_ζ, and the β coefficients associated with each covariate (*access* and *W. access*).

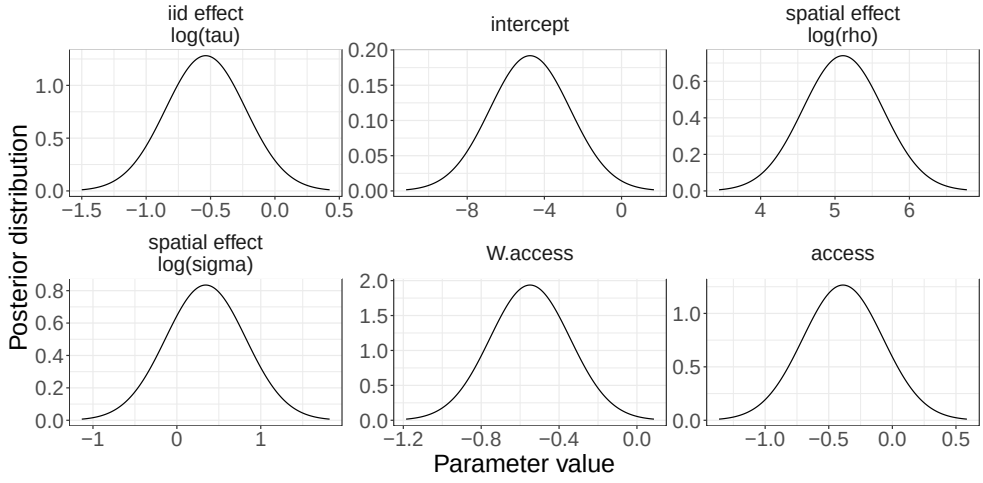

Figure S2: **Posterior distribution of the model parameters estimated by the downscaling model.** The graphic shows the posterior distribution (y-axis) of the estimated parameters (fixed-effects and random effects) (x-axis) in the model. It includes the log precision of the i.i.d effects (polygon-level), intercept, spatial hyperparameters log ρ and log σ_ζ, and the β coefficients associated with each covariate (*access* and *W. access*).

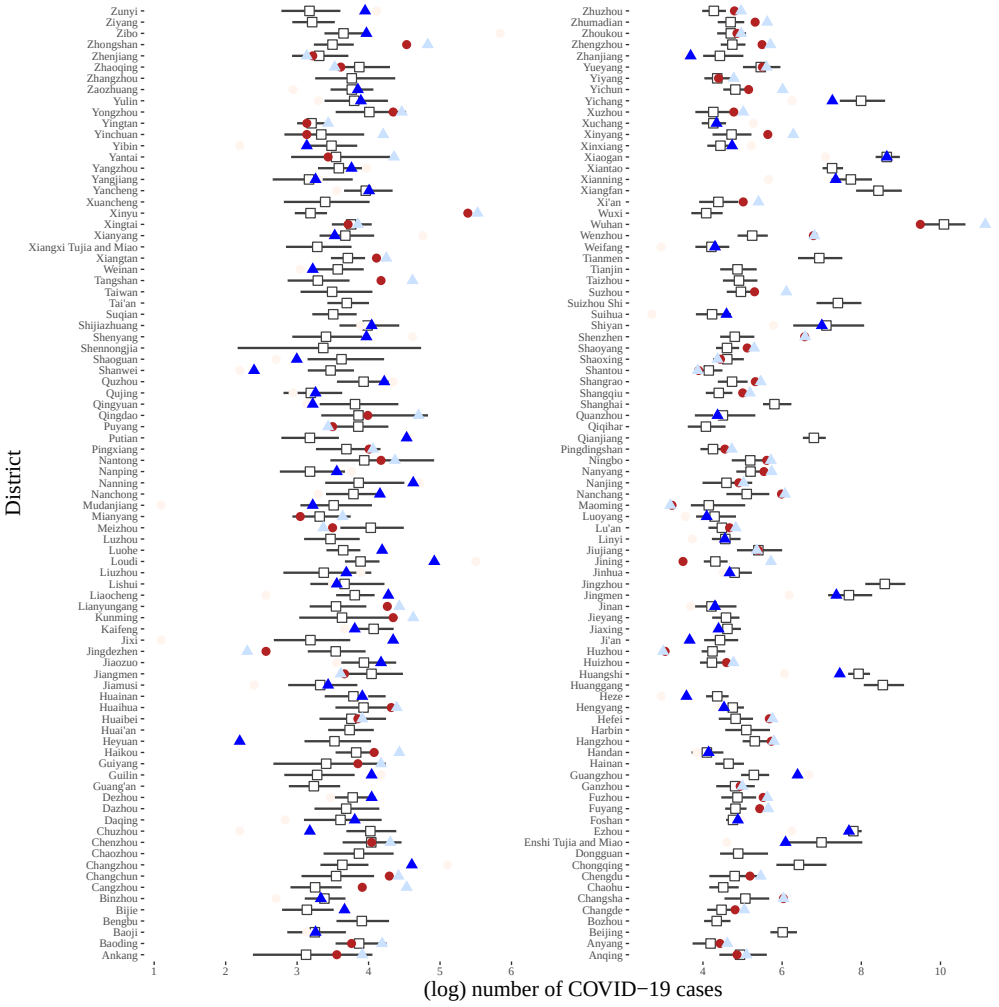

Figure S3: **Assessment of consistency of reported cases from The Paper and Xu et al. with estimated cases from downscaling (natural log scale) JHU data in high-impacted districts.** The figure shows the number of (natural log) COVID-19 infected cases for January and February 2020 at district-level (340 districts) in China from the Paper (*blue triangles*), Xu et al. (*red points*), along with an estimation of the (natural log) mean (*white squares*), 95% credible intervals (*grey segments*) of the infected cases based on the downscaling approach. For each district, the color of the symbol is faded for the corresponding dataset (Xu et al. or The Paper) that exhibit values less close to the predicted (natural log) mean.

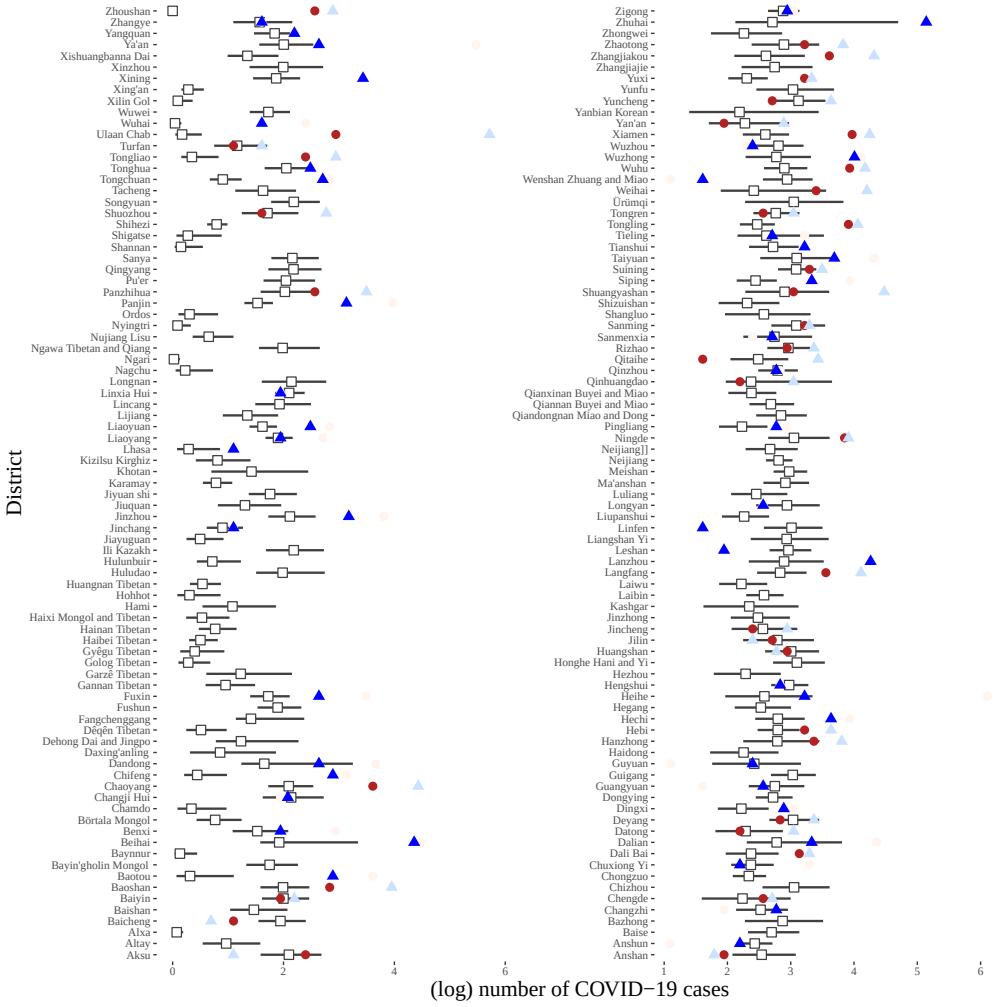

Figure S4: **Assessment of consistency of reported cases from The Paper and Xu et al. with estimated cases from downscaling (natural log scale) JHU data in low-impacted districts.** The figure shows the number of (natural log) COVID-19 infected cases for January and February 2020 at district-level (340 districts) in China from the Paper (*blue triangles*), Xu et al. (*red points*), along with an estimation of the (natural log) mean (*white squares*), 95% credible intervals (*grey segments*) of the infected cases based on the downscaling approach. For each district, the color of the symbol is faded for the corresponding dataset (Xu et al. or The Paper) that exhibit values less close to the predicted (natural log) mean.

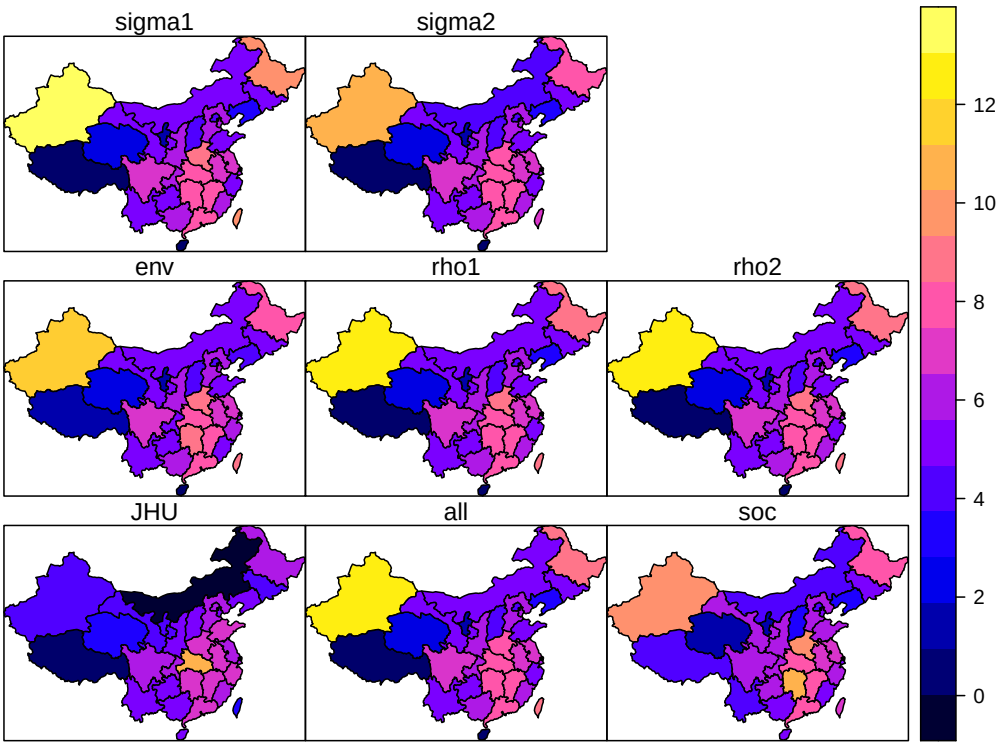

Figure S5: **JHU original COVID-19 data and out-of-sample estimated cases (January-February 2020) with downscaling (natural logarithm scale).** The maps show the original dataset of JHU of (natural log) COVID-19 infected cases for January and February 2020 at province-level in China (*bottom-right*), along with an estimation of the (natural log) infected cases based on the aggregation of all out-of-sample predictions in each of the 33 hold-out provinces from seven model specifications without i.i.d random effects: (1) include all covariates (*all*), (2) include only anthropogenic covariates (*soc*), (3) include only environmental covariates (*env*), and models that include all covariates but using alternative penalised complexity (PC) priors on the spatial parameters (*rho1,rho2,sigma1,sigma2*).

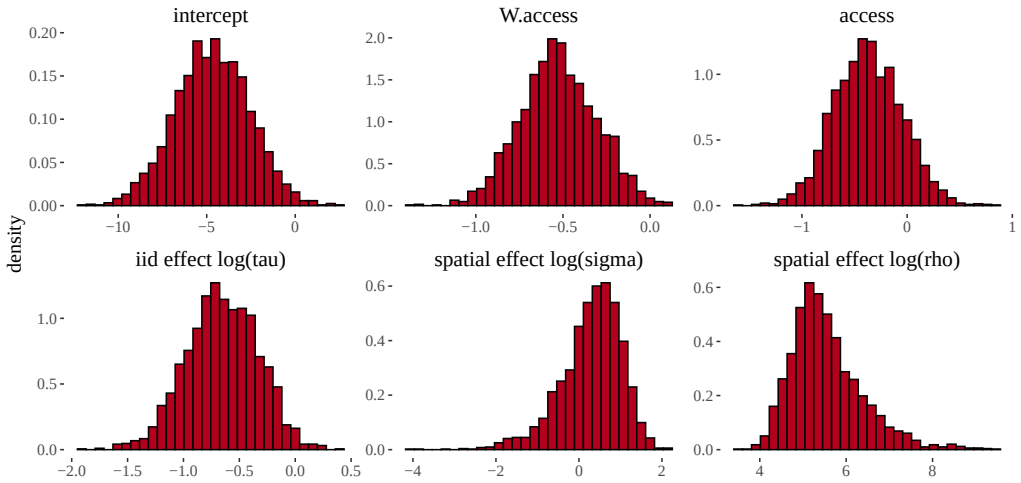

Figure S6: **Histogram of the NUTS parameter estimation.** The NUTS parameter estimation from MCMC draws using 800 (400 warm-up) iterations and 6 chains. The parameters are: the intercept, the covariate coefficients (*W.access* and *access*), the i.i.d. effects (log(τ)), and the parameters of the spatial effect (standard deviation log(σ) and range log(ρ). Warm-up samples are excluded.

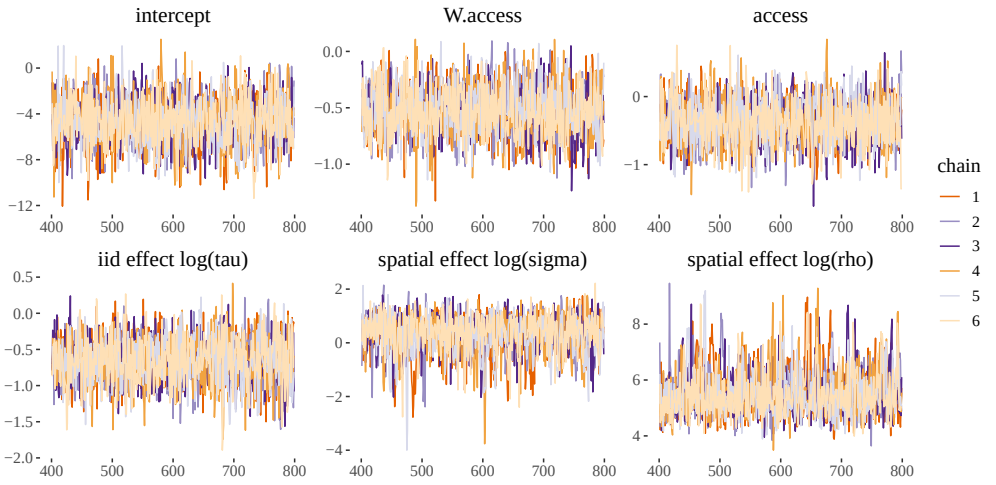

Figure S7: **Trace plot NUTS parameter estimation.** The NUTS parameter estimation from MCMC draws using 800 (400 warm-up) iterations and 6 chains. The parameters are: the intercept, the covariate coefficients (*W.access* and *access*), the i.i.d. effects (log(τ)), and the parameters of the spatial effect (standard deviation log(σ) and range log(ρ). Warm-up samples are not shown.