COVID-19 Ensemble Models Using Representative Clustering

Joon-Seok Kim[1], Hamdi Kavak[2], Andreas Züfle[1], Taylor Anderson[1]
[1]Department of Geography and Geoinformation Science, George Mason University, USA
[2]Department of Computational and Data Sciences, George Mason University, USA
{jkim258,hkavak,azufle,tander6}@gmu.edu

Abstract

In response to the COVID-19 pandemic, there have been various attempts to develop realistic models to both predict the spread of the disease and evaluate policy measures aimed at mitigation. Different models that operate under different parameters and assumptions produce radically different predictions, creating confusion among policy-makers and the general population and limiting the usefulness of the models. This newsletter article proposes a novel ensemble modeling approach that uses representative clustering to identify where existing model predictions of COVID-19 spread agree and unify these predictions into a smaller set of predictions. The proposed ensemble prediction approach is composed of the following stages: (1) the selection of the ensemble components, (2) the imputation of missing predictions for each component, and (3) representative clustering in application to time-series data to determine the degree of agreement between simulation predictions. The results of the proposed approach will produce a set of ensemble model predictions that identify where simulation results converge so that policy-makers and the general public are informed with more comprehensive predictions and the uncertainty among them.

1 Introduction

SARS-CoV-2 is a highly contagious human respiratory coronavirus resulting in mortality across the United States and worldwide [2]. Researchers have made considerable efforts to understand the virus' infection dynamics and develop various models to shed light on the future. Forecasts obtained from the models are used to predict the number of cases and deaths to support the development of effective policy interventions and the public health response. However, the wide range of COVID-19 models employ different parameter settings, are designed based on various assumptions, and are inherently uncertain. As a result, existing models produce a range of radically different predictions making it difficult for decision-makers and the broader public to understand, compare-between, and validate them, creating barriers to their use. Therefore, there is an urgent need to cross-evaluate the wide-range of existing COVID-19 models, find a consensus among their predictions, and increase the transparency of model assumptions and their inherent uncertainty.

Ensemble modeling is a term that describes the wide range of approaches used to combine predictions from multiple models, also known as components [28]. Components can be mathematical, curve-fitting, or agent-based models and typically operate under a range of different assumptions and use different data sources. The ensemble components can be combined using various algorithms, one

of which is referred to as stacked generalization [39] or stacking. In this approach, a single ensemble is generated by simply averaging predictions derived from equally weighted components. In variations of this approach, ensemble components may be weighted based on whether they meet a specific condition. These weights may be assigned statically or may change adaptively over time [22]. Aside from averaging, some ensembles are generated using the median, the trimmed mean to exclude extreme predictions, voting, Bayesian model averaging, multiple linear regression, and principal component regression [38].

Ensemble models often have been found to outperform any single model by offsetting component biases [33, 41]. If the components are diverse and independent, ensemble approaches can generate predictions with increased prediction accuracy and reduced error variance [17]. Thus, ensemble modeling has been utilized extensively to make predictions about weather and climate [18, 21], hydrologic processes [38], species distributions [10], and more recently infectious disease including influenza [31], Ebola hemorrhagic fever [37], dengue [14, 40], and COVID-19 [1, 23, 30].

Traditionally, ensemble approaches summarize the various predictions between components into one single prediction. However, the reliance on ensemble means without critical examination of the ensemble components can be dangerous. Mackenzie [21] illustrates this concept using three models, each of which indicates that a river is unsafe to cross at some point. Yet the average of the models says otherwise. In other words, acknowledging the assumptions and the resulting variation and bias among component predictions is important and can hold key information that explains future conditions otherwise ignored by their ensembles.

Therefore, we propose the development and implementation of an ensemble approach using representative clustering [32, 43] that is capable of exploring the various dimensions of agreement between ensemble components and thus is not limited to combining the component predictions into a single prediction. The novel representative clustering approach is proposed as follows: (1) selection of the ensemble components, (2) imputation of the missing predictions for each model, and (3) application of representative clustering to develop ensembles.

In this newsletter article, we begin by introducing the wide range of existing COVID-19 models that are available as potential ensemble components, their predictions, and their uncertainty . Next, we propose the novel ensemble prediction approach that will be used to unify selected components as ensembles. Finally, we present some initial results before describing our next steps.

2 Ensemble Clustering Approach

This study proposes the development of a novel ensemble prediction approach (see Figure 1) that is capable of exploring the various dimensions of agreement between ensemble components and thus is not limited to combining the component predictions into a single prediction. The proposed ensemble prediction approach is composed of the following stages: (1) the selection of the ensemble components, (2) the imputation of missing predictions for each model, and (3) the application of representative clustering so that it can be applied to time-series data and thus determine the degree of agreement between simulation predictions. Model cross-comparison is a rare practice in the modeling community. Therefore, our efforts to bring together different COVID-19 models and cross-compare them is also a novel contribution.

2.1 Model Integration

Selection of Model Components and Parameters. With the rapid spread of SARS-CoV-2, researchers have been designing simulation models to predict new cases and deaths as well as to understand the impact of different mitigation measures such as social distancing and mandatory lockdowns.

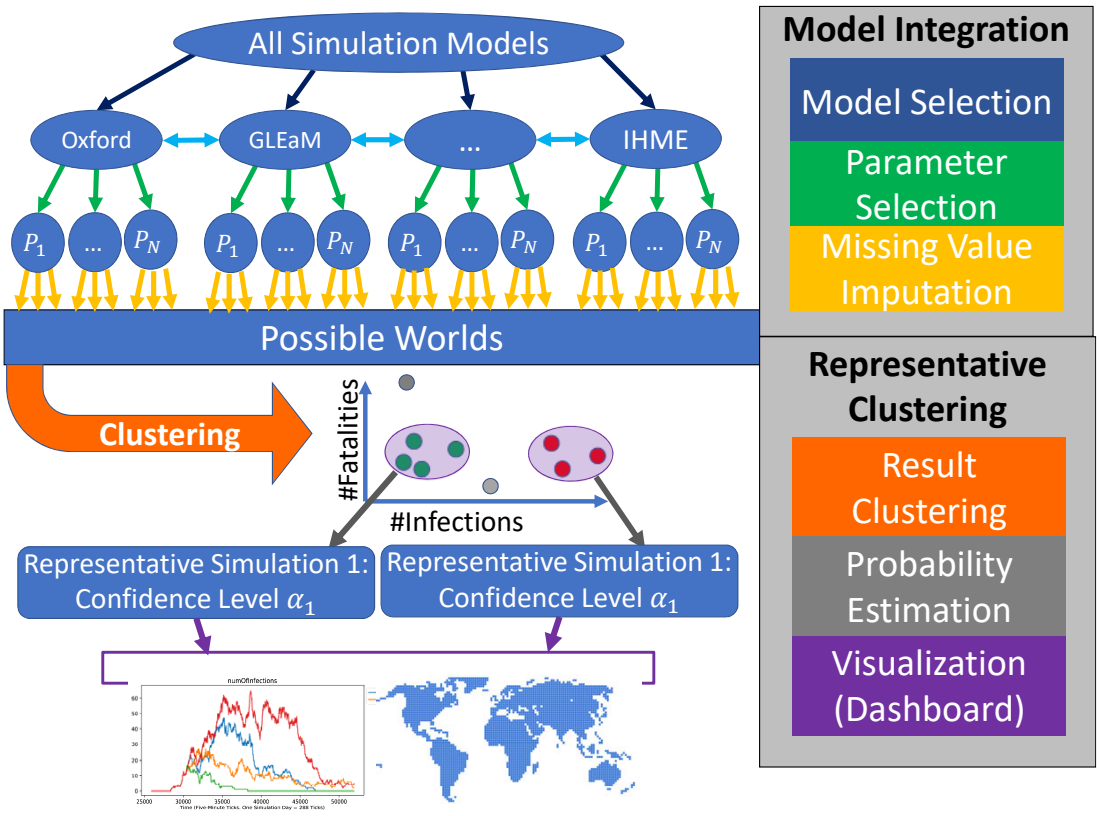

Figure 1: Ensemble clustering approach overview.

These models develop and implement approaches ranging from metapopulation, curve-fitting and statistical, as well as agent-based and in many cases the source code and the prediction data has been made publicly available. For example, the Global Epidemic and Mobility Model (GLEaM) is a metapopulation model that combines geographic mobility and population data with disease dynamics [36]. This effort was adapted and calibrated to model many outbreaks, including most recently the COVID-19 pandemic [4]. Another team of experts from the Los Alamos National Laboratory utilized their expertise in disease modeling and developed a statistical model to make new case and death predictions for COVID-19. Predictions from this model are publicly available [5]. The Imperial College COVID-19 Response Team adapted an established agent-based epidemic model [8, 12] to COVID-19 as well as developed a new mathematical model to estimate disease spread [7]. The Institute for Health Metrics and Evaluation (IHME) has developed a curve-fitting type of statistical model to project new cases and hospital beds needed [24], which is publicly available to use [13].

The models described above are just a few of the many that have been developed and implemented to predict COVID-19 trajectories of spread (see also Table 1 for more examples). In general, the CDC splits existing models into two categories [3]. One category includes models that make predictions under the assumption of business as usual meaning that existing control methods will remain in place [29, 34, 35]. The other category includes models that make predictions under different possible scenarios, usually with respect to testing the effect of different policy measures or the degree to which the population follows these guidelines [6, 11, 25, 27].

In the model integration stage, we will select a number of models as ensemble components. Table 1 presents some examples of existing models with open and available data that can be used as potential

Table 1: Examples of potential ensemble components.

Team Name and Reference	Model Name	Model Type
Auquan Data Science [34]	MLOptimized ModifiedSEIR	Modified SEIR model with compartments for reported and unreported infections. Non-linear mixed effects curve-fitting
Carnegie Mellon Delphi Group [9]	TimeSeries	A basic AR-type time series model fit using case counts and deaths as features
Columbia University [27]	Select	County-level SEIR model
CovidAnalytics at MIT [20]	DELPHI	SEIR model
Discrete Dynamical Systems [15]	Negative Binomial Dynamical System	Jointly modeling daily deaths and cases using a negative binomial distribution
GT [29]	DeepCOVID	Deep learning
Institute for Health Metrics and Evaluation [25]	CurveFit	Non-linear mixed effects curve-fitting
Los Alamos National Labs [26]	GrowthRate	Statistical dynamical growth model accounting for population susceptibility
MOBS Lab at Northeastern [35]	GLEAM COVID-19	Metapopulation, age structured SLIR model
NotreDame-FRED [6]	NotreDame-FRED	Agent-based model developed for influenza with parameters modified to represent the natural history of COVID-19
Youyang Gu (YYG) [11]	ParamSearch	SEIR model with machine learning layer

ensemble components. The goal is to select models that employ a range of modeling approaches and use a variety of assumptions. This is an important feature of the ensemble modeling approach, which relies on the diversity and independence between the ensemble components. Based on our selection, we will obtain each model's prediction data, made open and available by the COVID-19 Forecast Hub [30] as well as each model's respective repository or web pages. Although it varies from model to model, most model prediction data includes a start date, a prediction date, the predicted number of cumulative cases, the predicted number of incident and cumulative deaths, the predicted number of incident hospitalizations, the corresponding location for the prediction, and the confidence interval. We consider each prediction to be a *"Possible (future) World"*, analogous to work in uncertain database management [42, 44, 45]. The difference in uncertain database management is that current and past data is uncertain, whereas for disease prediction, it is data from the future that is uncertain. But in both cases, the challenge is to find a consensus among possible worlds (different database instances or different predictions) and enrich this consensus with reliability information.

Imputation of Missing Predictions. Due to the nature of independence of each model's development, the temporal resolution of the prediction data that is available for each model may be inconsistent and asynchronous. Imagine that two different models that make predictions starting from May 1st and onward. *Model X* might estimate the number of deaths and cases each day for the next four weeks. *Model Y* might estimate the number of deaths and cases each day for the next twenty weeks. In another scenario, imagine that another model, *Model Z* begins making predictions that start on May 5 and onward. Thus, there are no predictions available from *Model Z* from May 1st to May 4th. With the assumption that all of the models are calibrated to the most recent ground truth, the more recent the date of the forecast is, the more accurate the model is.

The inconsistent temporal resolution of the prediction data presents a challenge for the inclusion of

important components into the ensemble. As a result, we propose the use of imputation algorithms to fill the prediction gaps. In a sense, we aim here to predict the missing predictions of the models. We represent predictions in a three-mode tensor $\mathcal{P}_{i,j,m}$ such that a cell $p_{i,j,m}$ corresponds to a prediction made on Day i, made for Day j, by model m. For example, if one mode m predicts on Day $i = 10/02/2020$ that there will be 5000 deaths on Day $j = 10/09/2020$, then we will have $p_{i,j,m} = 5000$. This tensor is sparse, as existing models publish their predictions sparsely (often once per week), and predictions are made only for a short time window (often 14 or 28 days). As part of the model integration stage, we will test and evaluate various imputation algorithms and determine which are most accurate in predicting the missing model predictions. Some of the imputation algorithms we aim to test include linear interpolation, linear regression, non-negative matrix factorization [19] for individual prediction models, and tensor factorization [16]. We hope that more complex imputation algorithms are able to leverage collaborative filtering to fill missing model prediction by assessing that "other models had relatively high predictions for this day" and "this model had relatively low predictions for this day made on earlier days."

2.2 Representative Clustering

Once we have imputed the data, we will have obtained a broad set of predictions, each corresponding to different "possible worlds" generated from different models, different parameters, and under different assumptions. Each possible world consists of time series data, corresponding to the predicted number of incident and cumulative cases, the predicted number of incident and cumulative deaths, and the predicted number of incident hospitalizations. We want to use an approach that has been published for clustering of uncertain data [43] by mapping possible worlds into a reduced feature space then clustering possible worlds (in our case predictions) as depicted on the bottom half of Figure 1. Each cluster then corresponds to a set of mutually similar predictions which may stem from different models. We will then select the median among these predictions (defined as the prediction that minimizes the pairwise distance to other predictions in the same cluster) as a cluster representative. Assuming that each model has the same likelihood to correctly capture the unknown future, we can apply inductive statistics to estimate the probability that a cluster represents the unknown true future and provide an error bound using the radius of the cluster (the maximum distance between the cluster representative and other predictions in the same cluster). As the project continues, we will use supervised learning to reinforce the weights of those models and parameters, yielding the most accurate predictions. Each cluster representative can be considered an ensemble of models with a high degree of agreement between predictions. These representative will then be visualized on a dashboard which, instead of exploring the plethora of existing predictions, allows us to visually analyze a small number of representative predictions together with their confidence values. For example, the user may be presented with *Model X* and given the information that 40% of all predictions agree with this prediction up to an error which will be visualized using error bounds. This condensed representation takes the burden from users to interpret an overwhelming number of predictions and allows them to focus on only a small number of representative predictions.

3 An Online Medium to Disseminate Our Results

It is a likely scenario that the COVID-19 health emergency will continue into the coming several months and perhaps years. In a time of such uncertainty, policy-makers are right now relying on existing COVID-19 models to anticipate future conditions. Leveraging simulation models' predictive capabilities is critical to rapidly inform the public about what's likely to come and help policy-makers plan for those conditions. Therefore, there is an urgent need to compare and synthesize the wide-range

of existing COVID-19 models and their resulting simulations, disseminate this information clearly, and increase transparency about model assumptions and uncertainty. To address this need, we have designed a COVID-19 Ensemble Dashboard as a medium for which the broader public, decision-makers, the modeling community, and key organizations can explore and compare between existing models. The prototype for our proposed dashboard is presented in Figure 2.

Figure 2: COVID-19 Ensemble Dashboard

Our dashboard aims at providing two main functionalities regarding the status of COVID-19 at the US country-level and state-levels: (1) Giving users the option to examine dozens of existing COVID-19 models' predictions and over time, including both weekly and daily predictions as well as incident and cumulative metrics. Besides this, we provide a "stacked" time-series visualization to see all US states in the same picture. Users can also display individual predictions plotted against the ground-truth numbers, which facilitates examining model performance against real-world results; (2) Allowing users to examine the results from our representative clustering approach, as outlined in section 2. With this functionality, the users will be able to examine the agreement and disagreement between various models periodically. This functionality is currently being implemented and will be integrated as periodic reports into our dashboard. The dashboard will be updated as needed to incorporate new data and models as they become available, facilitating the opportunity to rapidly cross-compare new predictions and disseminate this information to the public.

4 Conclusion and Future Work

In this newsletter, we propose a novel ensemble modeling approach that leverages representative clustering to both examine the degree of agreement between models of COVID-19 spread and their predictions as well unify predictions into a smaller subset. The novel ensemble clustering approach begins with the process of data integration and thus the selection of ensemble components and the imputation of each component's missing predictions. Next, clustering is used to find a set of ensembles that are representative of groups of models with predictions that have a high degree of agreement for the same forecasting horizon. This research is still in early stages. Future steps include implementation and testing of the proposed approach. The proposed approach has the advantage of not being limited to the generation of one ensemble and thus acknowledges the unique assumptions of the components while removing the burden from policy makers, the general public, as well as other researchers to interpret an overwhelming number of COVID-19 model predictions.

Acknowledgements

This research is supported by National Science Foundation under Grant No. DEB-203068, "RAPID: An Ensemble Approach to Combine Predictions from COVID-19 Simulations."

References

[1] S. Abbott, J. Hellewell, R. N. Thompson, K. Sherratt, H. P. Gibbs, N. I. Bosse, J. D. Munday, S. Meakin, E. L. Doughty, J. Y. Chun, et al. Estimating the time-varying reproduction number of sars-cov-2 using national and subnational case counts. *Wellcome Open Research*, 5(112):112, 2020.

[2] P. Auwaerter. Johns Hopkins' ABX Guide. `https://www.hopkinsguides.com/hopkins/view/Johns_Hopkins_ABX_Guide/540747/all/Coronavirus_COVID_19__SARS_CoV_2_`. Accessed: 2020-04-11.

[3] Centers for Disease Control and Prevention. Covid-19 Forecasts: Deaths (`https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html`).

[4] M. Chinazzi et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (covid-19) outbreak. *Science*, 2020.

[5] S. Del Valle. Los Alamos COVID-19 Confirmed and Forecasted Case Data. `https://covid-19.bsvgateway.org/`. Accessed: 2020-04-11.

[6] G. Espana, R. Oidtman, S. Cavany, A. Costello, A. Wieler, A. Lerch, C. Barbera, M. Poterek, Q. Tran, S. Moore, and A. Perkins. NotreDame-FRED (`https://github.com/confunguido/covid19_ND_forecasting`).

[7] N. Ferguson et al. Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand, 2020.

[8] N. M. Ferguson, D. A. Cummings, C. Fraser, J. C. Cajka, P. C. Cooley, and D. S. Burke. Strategies for mitigating an influenza pandemic. *Nature*, 442(7101):448–452, 2006.

[9] A. Green, A. Hu, M. Jahja, V. Ventura, L. Wasserman, R. Tibshirani, V. Shankar, J. Bien, L. Brooks, B. Narasimhan, S. Rajanala, A. Rumack, N. Simon, J. Sharpnack, and R. Tibshirani. Carnegie Mellon Delphi Group (`https://delphi.cmu.edu`).

[10] G. Grenouillet, L. Buisson, N. Casajus, and S. Lek. Ensemble modelling of species distribution: the effects of geographical and environmental ranges. *Ecography*, 34(1):9–17, 2011.

[11] Y. Gu. Youyang Gu (YYG) (`https://covid19-projections.com`).

[12] M. E. Halloran, N. M. Ferguson, S. Eubank, I. M. Longini, D. A. Cummings, B. Lewis, S. Xu, C. Fraser, A. Vullikanti, T. C. Germann, et al. Modeling targeted layered containment of an influenza pandemic in the united states. *Proceedings of the National Academy of Sciences*, 105(12):4639–4644, 2008.

[13] IHME. COVID-19 Projections. `https://covid19.healthdata.org/united-states-of-america`. Accessed: 2020-04-18.

[14] M. A. Johansson, K. M. Apfeldorf, S. Dobson, J. Devita, A. L. Buczak, B. Baugher, L. J. Moniz, T. Bagley, S. M. Babin, E. Guven, et al. An open challenge to advance probabilistic forecasting for dengue epidemics. *Proceedings of the National Academy of Sciences*, 116(48):24268–24274, 2019.

[15] R. Kalantari and M. Zhou. Discrete Dynamical Systems (`https://dds-covid19.github.io/`).

[16] A. Karatzoglou, X. Amatriain, L. Baltrunas, and N. Oliver. Multiverse recommendation: n-dimensional tensor factorization for context-aware collaborative filtering. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 79–86, 2010.

[17] V. Kotu and B. Deshpande. *Predictive analytics and data mining: concepts and practice with rapidminer*. Morgan Kaufmann, 2014.

[18] T. Krishnamurti, C. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendran. Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, 285(5433):1548–1550, 1999.

[19] M. Kurucz, A. A. Benczúr, and K. Csalogány. Methods for large scale svd with missing values. In *Proceedings of KDD cup and workshop*, volume 12, pages 31–38. Citeseer, 2007.

[20] M. L. Li, H. T. Bouardi, O. S. Lami, T. A. Trikalinos, N. K. Trichakis, and D. Bertsimas. CovidAnalytics at MIT (`https://www.covidanalytics.io/`).

[21] D. Mackenzie. *Mathematics of climate change: a new discipline for an uncertain century*. Mathematical Sciences Research Institute, 2007.

[22] T. McAndrew and N. G. Reich. Adaptively stacking ensembles for influenza forecasting with incomplete data. *arXiv preprint arXiv:1908.01675*, 2019.

[23] P. Melin, J. C. Monica, D. Sanchez, and O. Castillo. Multiple ensemble neural network models with fuzzy response aggregation for predicting covid-19 time series: the case of mexico. In *Healthcare*, volume 8, page 181. Multidisciplinary Digital Publishing Institute, 2020.

[24] C. J. Murray et al. Forecasting covid-19 impact on hospital bed-days, icu-days, ventilator-days and deaths by us state in the next 4 months. *medRxiv*, 2020.

[25] C. Murry. IHME (`https://covid19.healthdata.org/united-states-of-america`).

[26] D. Osthus, S. D. Valle, C. Manore, B. Weaver, L. Castro, C. Shelley, M. M. Smith, J. Spencer, G. Fairchild, T. Pitts, D. Gerts, L. Dauelsberg, A. Daughton, M. Gorris, B. Hornbein, D. Israel, N. Parikh, D. Shutt, and A. Ziemann. Los Alamos National Labs (`https://covid-19.bsvgateway.org/`).

[27] S. Pei, T. Yamana, S. Kandula, W. Yang, M. Galanti, and J. Shaman. Columbia University (`https://blogs.cuit.columbia.edu/jls106/publications/covid-19-findings-simulations/`).

[28] R. Polikar. Ensemble based systems in decision making. *IEEE Circuits and systems magazine*, 6(3):21–45, 2006.

[29] A. Prakash, A. Rodriguez, J. Cui, A. Tabassum, and B. Adhikari. GT (`https://deepcovid.github.io/`).

[30] E. L. Ray, N. Wattanachit, J. Niemi, A. H. Kanji, K. House, E. Y. Cramer, J. Bracher, A. Zheng, T. K. Yamana, X. Xiong, et al. Ensemble forecasts of coronavirus disease 2019 (covid-19) in the us. *medRxiv*, 2020.

[31] N. G. Reich, C. J. McGowan, T. K. Yamana, A. Tushar, E. L. Ray, D. Osthus, S. Kandula, L. C. Brooks, W. Crawford-Crudell, G. C. Gibson, et al. Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the us. *PLoS computational biology*, 15(11):e1007486, 2019.

[32] K. A. Schmid and A. Züfle. Representative query answers on uncertain data. In *SSTD'19*, pages 140–149, 2019.

[33] C. Tebaldi and R. Knutti. The use of the multi-model ensemble in probabilistic climate projections. *Philosophical transactions of the royal society A: mathematical, physical and engineering sciences*, 365(1857):2053–2075, 2007.

[34] V. Tomar and C. Jain. Auquan Data Science (`https://covid19-infection-model.auquan.com/`).

[35] A. Vespignani, M. Chinazzi, J. T. Davis, K. Mu, A. P. y Piontti, N. Samay, X. Xiong, M. E. Halloran, I. M. Longini, N. E. Dean, K. Sun, M. Litvinova, C. Gioannini, L. Rossi, and M. Ajelli. MOBS Lab at Northeastern (`https://covid19.gleamproject.org/`).

[36] A. Vespignani et al. COVID-19 MODELING IN THE UNITED STATES. `https://covid19.gleamproject.org/`. Accessed: 2020-04-11.

[37] C. Viboud, K. Sun, R. Gaffey, M. Ajelli, L. Fumanelli, S. Merler, Q. Zhang, G. Chowell, L. Simonsen, A. Vespignani, et al. The rapidd ebola forecasting challenge: Synthesis and lessons learnt. *Epidemics*, 22:13–21, 2018.

[38] N. R. Viney, H. Bormann, L. Breuer, A. Bronstert, B. F. Croke, H. Frede, T. Gräff, L. Hubrechts, J. A. Huisman, A. J. Jakeman, et al. Assessing the impact of land use change on hydrology by ensemble modelling (luchem) ii: Ensemble combinations and predictions. *Advances in water resources*, 32(2):147–158, 2009.

[39] D. H. Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

[40] T. K. Yamana, S. Kandula, and J. Shaman. Superensemble forecasts of dengue outbreaks. *Journal of The Royal Society Interface*, 13(123):20160410, 2016.

[41] T. K. Yamana, S. Kandula, and J. Shaman. Individual versus superensemble forecasts of seasonal influenza outbreaks in the united states. *PLoS computational biology*, 13(11):e1005801, 2017.

[42] A. Zuefle. Uncertain spatial data management: An overview. *arXiv preprint arXiv:2009.01121*, 2020.

[43] A. Züfle, T. Emrich, K. A. Schmid, N. Mamoulis, A. Zimek, and M. Renz. Representative clustering of uncertain data. In *ACM KDD'14*, pages 243–252, 2014.

[44] A. Züfle, G. Trajcevski, D. Pfoser, M. Renz, M. T. Rice, T. Leslie, P. Delamater, and T. Emrich. Handling uncertainty in geo-spatial data. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, pages 1467–1470. Ieee, 2017.

[45] A. Züfle, G. Trajcevski, D. Pfoser, and J. S. Kim. Managing uncertainty in evolving geo-spatial data. In *2020 21st IEEE International Conference on Mobile Data Management (MDM)*, pages 5–8, June 2020.