Pandemic Resilience

Case studies of an AI-calibrated ensemble of models to inform decision making

November, 2024

GPAI | THE GLOBAL PARTNERSHIP ON ARTIFICIAL INTELLIGENCE

This report was planned prior to the integration of the Global Partnership on Artificial Intelligence (GPAI) and the Organisation for Economic Co-operation and Development (OECD) mid-2024. Consequently, the report was not subject to approval by GPAI and OECD members and should not be considered to reflect their positions.

Acknowledgements

This report was developed in the context of the Responsible AI for Pandemic Resilience project, with the steering of the project Co-Leads, supported by the Global Partnership on Artificial Intelligence (GPAI) Responsible AI Working Group. The GPAI Responsible AI Working Group agreed to declassify this report and make it publicly available.

Project lead: **Michael O'Sullivan**[*],University of Auckland

The report was prepared and compiled by: **Ahmed Farid**[‡], Sudanese Standards and Metrology Organization; **Sahar Bahrami**[‡], McGill University; **Michael O'Sullivan**[*], University of Auckland; **Jamieson Warner**[‡], The University of Texas at Austin.
The report is the result of a collaborative effort where the following individuals offered expertise, suggestions, reflections, and comments: **Nathaniel Hupert**[‡], Weill Cornell Medical College, Cornell University; **Patrick McSharry**[‡], Carnegie Mellon University Africa;**Risto Miikkulainen**[‡],The University of Texas at Austin; **Olivier Francon**[‡], Cognizant AI Labs; **Arnaud Quenneville-Langis**[†] and **Antoine Glory**[†] from CEIMIA; **Victoria Agyepong**[‡], Beyond3Generations; **Nathan Allen**[‡], University of Auckland; **Sobhan Chatterjee**[‡], University of Auckland.

GPAI is grateful for the meaningful contribution of Pandemic Resilience Project ex co-leads **Margarita Sordo**[*], Harvard Medical School, and **Paul Suetens**[*], KU Leuven.

GPAI would like to acknowledge the contribution of Associate Professors **Cameron Walker**[†] and **Ilze Ziedins**[†] for their prior contribution to development of the Continuous Time Markov Chain (CTMC) model. This model was developed as part of a COVID-19 modelling project within Te Pūnaha Matatini and funded by Aotearoa | New Zealand's Ministry of Business, Innovation and Employment (MBIE).

GPAI would like to thank **Celine Caira**[**], OECD AI, who gave valuable feedback and commented on drafts of this report. GPAI would also like to acknowledge the tireless efforts of the colleagues at the International Centre of Expertise in Montréal on Artificial Intelligence (CEIMIA) and of the GPAI Responsible AI Working Group. We are grateful, in particular, for the support of **Arnaud Quenneville-Langis**, **Stephanie King** and **Antoine Glory** from CEIMIA, and for the dedication of the Working Group Co-Chairs **Amir Banifatemi**[*], (AI Commons) and **Francesca Rossi**[*], (IBM Research).

[*] Expert of GPAI's Responsible AI Working Group
[**] Observer at GPAI's Responsible AI Working Group
[†] Invited specialist [‡] Contracted party by the CofEs

Citation

GPAI 2024. *Pandemic Resilience: Case studies of an AI-calibrated ensemble of models to inform decision making*. Report, November 2024, Global Partnership on Artificial Intelligence (GPAI).

Executive summary

Report overview

This report from Global Partnership on Artificial Intelligence (GPAI)'s Pandemic Resilience project follows its 2023 report and is focused on practically implementing the concepts previously developed by the project team. Indeed, the 2023 report laid the foundation for this research while presenting recommendations on various approaches that aligned pandemic modelling with responsible Artificial Intelligence (AI). The 2023 report showcased a calibration framework approach and an ensemble modelling concept, focusing on the added value and pertinence of both consistent calibration and ensembling; that is, ensuring models are consistent in shared parameter values while using the strengths of different models and creating a digital "task force". The combination of the calibration framework and ensemble model encourages and enables modellers from different locations and backgrounds to work together by using standardised versions of their work.

Although there has been substantial modelling activity of Non-Pharmaceutical Interventions (NPIs) for COVID-19, this activity has been fragmented across different countries, with mixed access and sharing of data and models. This report documents a prototype calibration framework – based on a multi-objective genetic algorithm – that simultaneously calibrates multiple models across different locations and ensures consistent parameter values across models. The resulting, calibrated models are then combined using an ensemble modelling concept that provides more accurate model results than any of the models do individually. Hence, consistent models for multiple locations are created and can be shared easily with these locations. In addition, diverse perspectives from the models can provide more accurate results for each location through the ensemble model.

Initial case study results show that long runs of the calibration framework improves model accuracy by approximately 60%. They also show the efficacy of the calibration framework over manual calibration. However, artefacts from the underlying models still present challenges for calibration at the beginning of a modelling horizon. Initial case study results for the ensemble model show reasonable improvements for prediction accuracy in locations with large numbers of COVID-19 cases, but the inclusion of models that work well for lower case numbers needs to be explored to fully investigate the benefits and limitations of ensemble modelling.

Key findings

- Standardisation of inputs/outputs for models is important for enabling new models to be easily added to the calibration framework and ensemble model.
- Ensemble models provide robustness and diverse perspectives when combining multiple models for making predictions. Ensemble models also provide further robustness by adapting how they combine the underlying models to improve prediction accuracy.
- By carrying out case studies across different countries, satisfying results have been obtained for multiple different locations and insights from all locations are used during calibration.
- Automated calibration via long runs of the calibration framework provide more accurate results than time consuming manual calibration.
- Ensemble models provide estimates of prediction uncertainty, a key innovation in this report that enhances the robustness of the resulting predictions.
- The combination of calibration framework and ensemble modelling enables informed, data-driven decision making by policymakers across multiple locations in a way that aligns with responsible AI principles. This enhances the democratisation of pandemic modelling, digital technology and AI.

Recommendations

- **Standardise datasets and models**
  - Standardise datasets across locations to: 1) enable models developed in one location to be used in other locations; and 2) enable data from multiple locations to be used to calibrate models. Enhance the ability of data and models to be shared.
- **Extend and refine the calibration framework and ensemble modelling and integrate them into pandemic preparedness initiatives/organisations**
  - Add models predicting economic impacts of NPIs. Complete further testing and refinement of both the calibration framework and ensemble modelling. Consult with pandemic preparedness initiatives/organisation to ascertain how the framework could be 1) integrated with systematic gathering of disease data for monitoring disease spread; and 2) the ensemble model could be integrated with disease spread prediction and public health response. Add other models as appropriate to ensure diversity of perspective in response and robustness of modelling to different disease progression.
- **Test the efficacy of the calibration framework and ensemble modelling within a tabletop exercise**

- Pilot the use of the calibration framework and ensemble modelling to provide solutions to policymakers from different locations in a tabletop simulation of a pandemic outbreak.

Glossary

AI An Artificial Intelligence (AI) system is “a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments.” (OECD, 2019). 1, 3, 8

API Application Programming Interface (API) defines the inputs and outputs to an application, i.e., piece of software. 7

CTMC A Continuous Time Markov Chain (CTMC) is “a continuous stochastic process in which, for each state, the process will change state according to an exponential random variable and then move to a different state as specified by the probabilities of a stochastic matrix.” (Chen and Mao, 2021). 2, 8

Cyber-Physical The Cyber-Physical model in this report is based on the plant-controller concept from Cyber-Physical Systems Lee, 2008 – see 3.2.2 for details.. 8, 17–21, 26, 28, 34, 37–39, 41, 47, 49

DevOps Development Operations (DevOps) is end-to-end automation of development and delivery of software. (Ebert et al., 2016). 50

GA Mitchell, 1996 states that no rigorous definition of Genetic Algorithms (GAs) is “accepted by all in the evolutionary-computation community that differentiates GAs from other evolutionary computation methods. However, it can be said that most methods called “GAs” have at least the following elements in common: populations of chromosomes, selection according to fitness, crossover to produce new offspring, and random mutation of new offspring”. 26

GPAI The Global Partnership on Artificial Intelligence (GPAI) is “a multi-stakeholder initiative which aims to bridge the gap between theory and practice on AI by supporting cutting-edge research and applied activities on AI-related priorities.” (GPAI, 2021). 1, 2, 8

LSTM A Long Short-Term Memory network (LSTM) is a recurrent neural network approach that uses “a novel, efficient, gradient based method” (Hochreiter and Schmidhuber, 1997). 8

ML Machine Learning (ML) is a discipline of artificial intelligence (AI) and computer science that uses data and algorithms to allow AI to learn in the same way that people do, gradually improving its accuracy. (Beyerer, Kühnert, and Niggemann, 2019). 13

MOGA Multi-Objective Genetic Algorithms (MOGAs) (Fonseca and Fleming, 1993) use a standard genetic algorithm approach combined the multi-objective concept of non-dominated solutions when selecting individual solutions that are selected for subsequent generations of solutions. 8

NPI Non-Pharmaceutical Intervention (NPIs) are "actions, apart from getting vaccinated and taking medicine, that people and communities can take to help slow the spread of illnesses." (CDC, 2022). 1, 9

OxCGRT Oxford Covid-19 Government Response Tracker (OxCGRT) "provides a systematic cross-national, cross-temporal measure of how government responses have evolved over the full period of the disease's spread." (Hale et al., 2021). 7

pymoo Pymoo is "a multi-objective optimization framework in Python" (Kalyanmoy Deb et al., 2002) that offers state of the art single- and multi-objective optimization algorithms – see 3.4.1 for details.. 27, 28, 32

RIO Residual Estimation with I/O Kernels (RIO) is an ensemble modelling approach from Qiu, Meyerson, and Miikkulainen, 2020 – see §4 for more detail.. 8

SEIR Susceptible-Exposed-Infected-Recovered (SEIR) processes are compartmental models used for epidemiology with four compartments (a.k.a. state): 1) Susceptible: individuals that may become infected if they come into contact with infectious individuals; 2) Exposed: individuals that are infected, but have not become infectious yet; 3) : Infectious: individuals that are infected and may spread the disease, i.e., are infectious; 4) : Recovered: individual that have recovered, are no longer infectious and (in many cases) have immunity to the disease. 9

SVM Support Vector Machine (SVM), a class of algorithms for classification, regression and other applications. (Cristianini and Ricci, 2008). 13

Key Acronyms

AI Artificial Intelligence. 1–3, 8, 9, 13, 47, 48, 51–54, 57, *Glossary:* AI

API Application Programming Interface. 7, 15, 47, *Glossary:* API

CDC Centers for Disease Control and Prevention. *Glossary:* CDC

CFR Case Fatality Rate. *Glossary:* CFR

CTMC Continuous Time Markov Chain. 2, 8, 17, 19–21, 26, 28, 34, 37–39, 41, 46–49, *Glossary:* CTMC

DevOps Development Operations. 50, 51, *Glossary:* DevOps

GA Genetic Algorithm. 26, 27, 36, *Glossary:* GA

GPAI Global Partnership on Artificial Intelligence. 1, 2, 8, 46, 53, 57, *Glossary:* GPAI

ICU Intensive Care Unit. 19, 21

LSTM Long Short-Term Memory network. 8, 13, 20, 21, 26, 28, 34, 37, 39, 48, 49, *Glossary:* LSTM

MAE Mean Absolute Error. *Glossary:* MAE

ML Machine Learning. 3, 13, 48, *Glossary:* ML

MOGA Multi-Objective Genetic Algorithm. 8, 15, 27, 32–34, *Glossary:* MOGA

NPI Non-Pharmaceutical Intervention. 1, 2, 9, 10, 14, 17–19, 22, 24, 25, 49–54, *Glossary:* NPI

OxCGRT Oxford Covid-19 Government Response Tracker. 7, 12, 24, 47, 48, 50, *Glossary:* OxCGRT

RIO Residual Estimation with I/O Kernels. 8, 32, 35, 42–47, 49, 51, *Glossary:* RIO

RMSE Root Mean Square Error. *Glossary:* RMSE

SEIR Susceptible-Exposed-Infected-Recovered. 9, 10, 14, *Glossary:* SEIR

SVM Support Vector Machine. 13, *Glossary:* SVM

1 Introduction

Coronavirus Disease 2019 (COVID-19), caused by the novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), was a rapidly evolving global emergency that placed significant strain on healthcare systems throughout the world (Gallo Marin et al., 2021). In addition to severe health consequences across the globe, it had a substantial negative influence on various fields of human endeavor, including education, business, global travel, and especially health care. These economic consequences of the COVID-19 pandemic highlight the need for creative scenario planning techniques that balance the provision of effective public health interventions with the need for economic security on a local and international scale (Youn, Geismar, and Pinedo, 2022). This report discusses the difficulty of reacting to catastrophic public health events, which can lead to significant mortality and economic loss, and presents a model-based approach to aid decision-making that leverages existing work on epidemic modelling and standardising public data.

The COVID-19 pandemic provides a good opportunity to evaluate how policymakers tackle the issue of how to respond locally to a global pandemic as well as balancing different factors in their response. First, it highlights the necessity of frequent, standardised gathering of disease data for effective disease monitoring such as in the One World-One Health model (Hayman et al., 2023), as well as the need to exchange knowledge rapidly in order to better inform national and international decision-making. Creating public, standardised datasets is a key part of disease tracking/monitoring and knowledge sharing. This project uses the Oxford Covid-19 Government Response Tracker (OxCGRT) dataset that was acquired, standardised, and curated at Oxford University (Hale et al., 2021). Knowledge sharing, including tools with embedded expertise such as calibrated models, is also key for supporting those countries that do not have the necessary resources to develop models to inform their pandemic response. There was a plethora of modelling activity both regarding the spread of COVID-19 and how this spread was affected by various response options (see §2.1 for a summary). However, there were also challenges in sharing the knowledge and models developed between countries (see §2.2 for a summary). Rapid customisation of existing models to local data was rare during the pandemic (for one example of a locally customizable, template-running, web-based model, see Aguas et al., 2020). Such efforts were hampered by inconsistent data availability and lack of multiple modeling approaches. This project addresses this gap by combining standardised data with a standardised Application Programming Interface (API) for modelling, meaning that multiple models can be customised for different locations rapidly, and that these models and locations will use shared parameters consistently. In addition, models with multiple perspectives such as pandemic spread and economic performance can be used with the same integrated dataset to consider multiple metrics of performance for a given response policy. An ensemble of pandemic spread models is used within this report and the addition of (one or more) economic models is outlined as future work. Other

models, such as those on the effect of the pandemic on mental health and the inclusion of interventions to promote (physical and mental) health could also be included to provide a more holistic approach to pandemic response policy.

Pandemics disturb normal social, economic, (physical and mental) health care, and governmental activities, so the pandemic response must be carefully planned and implemented. Governments need to consider, for example, trade-offs between a centralized versus decentralized response strategies based on demographics, geography, and other factors. Centralized responses may appear to improve control of public health interventions but may discourage community participation and voluntary compliance with pandemic-related restrictions. Trade-offs between health outcomes and economic performance must also be carefully evaluated when planning a pandemic response.

This report is the second of two progress reports from the Global Partnership on Artificial Intelligence (GPAI) Pandemic Resilience project. The previous report (GPAI, 2023) presented recommendations on various approaches that aligned pandemic modelling with responsible Artificial Intelligence (AI). The research presented in this report provides an important step in the provision of ensembles of standardised models that can be rapidly customised, calibrated, and deployed to support pandemic responses across the globe.

This report is structured as follows. Following this introduction, Section 2 reviews the necessary background on COVID-19 modelling methods and efforts.

Section 3 describes the automatic calibration methodology using a Multi-Objective Genetic Algorithm (MOGA). The emphasis is on the simultaneous calibration of input parameters across multiple models and locations to account for uncertainty, provide a calibrated ensemble model, and include different perspectives such as economic performance. This section provides a detailed account of the calibration framework and the standardized interface for combining and comparing three distinct but complementary modelling approaches, namely: (1) a Cyber-Physical model; (2) a Long Short-Term Memory network (LSTM) model, which is a form of neural network; and (3) a stochastic model using Continuous Time Markov Chains (CTMCs). Defining a standard for inputs, outputs, and calibration, including data, enabled models to use the common parameters and data, make comparable predictions, and be calibrated simultaneously.

Section 4 describes the methodology for combining multiple models into an ensemble. Uncertainty in the predictions is estimated using Residual Estimation with I/O Kernels (RIO) method, and these estimates are then used to form a combined prediction as a mixture of Gaussians.

In section 5, the experiments and results for the calibration framework are presented through two case studies. The first one considers Aotearoa | New Zealand, Kenya, Sweden, and the United Kingdom from March to June 2020. The second case study calibrates models to compare the results of Aotearoa | New Zealand and Sudan. Comparisons with manually calibrated models and across the two case studies are also described.

Following the case studies, the report concludes in Section 6 with reflections on the research into the calibration framework and suggestions for future directions for this research. We consider the contributions of this research, namely the value of standardisation, how the use of ensemble models provides robustness and diversity, how informed data-driven decision-making is enabled, how pandemic preparedness is supported, and how the research's approach aligns with responsible AI principles. We then analyze limitations in the research so far and provide ideas to address difficulties. We next consider how this research could be used in practice via three use cases for decision-makers. We conclude both the section and the report as a whole by conclude this section by exploring how the contributions, limitations and use cases motivate future research using standardisation, the calibration framework, and ensemble modelling.

2 Background

This section summarises relevant background literature on COVID-19 modelling, knowledge sharing, ensemble models, and standardised data/models.

2.1 COVID-19 modelling

Table 2.1 summarises the different COVID-19 modelling approaches used throughout the world. Note that this table is not supposed to be comprehensive, but shows the breadth and diversity of COVID-19 initiatives globally. Although there was a significant level of activity throughout the world, with some common approaches, there was not a coordinated approach globally, i.e., the activity was very fragmented across countries and even across regions within countries. In addition, many developing countries had limited capacity to create models for their own populations.

One of the key outcomes from COVID-19 modelling was the ability of models to inform how the use of Non-Pharmaceutical Interventions (NPIs) would affect the spread of COVID-19. Models that included the effect of NPIs are summarised here along with the method for including NPI effects.

Liu, Thomadsen, and Yao (2020) highlighted the network effects and social distancing effects in the spread rate of COVID-19. They have used a modified version of the Susceptible-Infected-Recovered (SIR) model of the spread rate under different social distance levels. Chowdhury et al. (2020) used a standard Susceptible-Exposed-Infected-Recovered (SEIR) compartmental model to assess the impact of dynamic community-based NPIs, to control COVID-19 pandemic in 16 diverse economic regions: Western Europe (The Netherlands, Belgium), South America (Chile, Colombia), North America (Mexico), Africa (South Africa, Nigeria, Ethiopia, Tanzania, Uganda), South Asia (India, Bangladesh, Pakistan, Sri Lanka), West Asia (Yemen), and the Pacific (Australia). According to the results, dynamic suppres-

sion interventions will assist countries with reducing mortality rate and preventing critical care overload, reducing global economic burden and giving them time to develop clinical preventive strategies. Sarkar, Khajanchi, and Nieto (2020) developed a SEIR model to predict the spread of COVID-19 outbreak in 17 provinces of India. The model emphasized the effectiveness of social distancing and contact tracing between uninfected and infected individuals.

Banholzer et al. (2021) used a semi-mechanistic Bayesian hierarchical model to investigate the effectiveness of NPIs on the number of new infections across 20 countries (i.e., the United States, Canada, Australia, Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Portugal, Spain, Sweden, and the United Kingdom, Norway, and Switzerland). The bans on large gatherings reduced transmission considerably. The effect of stay-at-home orders and work-from-home orders were comparatively small. Brauner et al. (2021) estimated the NPI effectiveness on COVID-19 transmission in 41 countries using data-driven, cross-country modeling (Bayesian hierarchical model). Closing both schools and universities, gathering bans (limiting gatherings to 10 people or less), and face-to-face business closures were the most effective. The stay-at-home orders were less effective.

Country	SEIR	Machine Learning	Compartmental Modelling	Ensemble Modelling	Other
Aotearoa | New Zealand	Hendy et al., 2021		Ziedins, Walker, and O'Sullivan, N.D.		Hendy et al., 2021 (Statistical branching)
India			Sarkar, Khajanchi, and Nieto, 2020		
Kenya			Wangari et al., 2021		
Nigeria			Iboi et al., 2020		
Sudan	Elsheikh et al., 2020				
United Kingdom		Abdulaal et al., 2020			
United States					Tang and S. Wang, 2020 (Mathematical modeling)
Global	Cooper, Mondal, and Antonopoulos, 2020	Cognizant, 2021		**This research**	

Table 2.1: Summary of COVID-19 modelling throughout the world

2.2 Sharing knowledge/models

Giabbanelli et al. (2021) shared challenges, and suggested solutions, in the development of COVID-19 modeling through assessment performed by modelers in six COVID-19 research teams. Despite variation in the models, they encountered common challenges, including unknown or publicly unavailable data, along with the lack of public presence which limits a research team's impact. Therefore, the importance of rapidly emerging centralized data repositories was emphasized as a way to address this challenge. Moreover, the importance of relationships and communications between modelers, policymakers, and the general public was highlighted. According to Bertozzi et al. (2020), other challenges in COVID-19 modeling include uncertainty in parameters, length and severity of social distancing, and outcomes depending on the data or type of model that has been used. Aguas et al. (2020) established the COVID-19 Modelling (CoMo) Consortium and a dynamic infectious disease model to overcome the inaccessibility of COVID-19 mathematical models to policymakers in low-income and middle-income countries. This approach addressed the global need for technology, training, and effective communication.

Cosgriff, Ebner, and Celi (2020), mentioned lack of patient-level COVID-19 publicly available datasets. In this regard, having a unifying multinational, universally shared COVID-19 electronic health record would be necessary. The Medical Information Mart for Intensive Care (MIMIC), containing 50,000 patient admissions to Beth Israel Deaconess Medical Center (BIDMC), is an example of open data sharing for a wider geography of BIDMC's hospital (A. Johnson, Pollard, and Roger, 2016). Gao et al. (2020) proposed different integrated platform models for aiding data sharing and management of COVID-19 pandemic information at national and international levels and to overcome challenges such as data decentralization, standardization, and globalization. They arrived at these platform models after reviewing existing approaches, tools, and software in this domain. Their integrated approach will facilitate the global transfer of datasets and collaboration between researchers, scientists, and institutions to aid the prevention, and treatment of COVID-19 as well as research. Hale et al. (2021) introduced a global, publicly available dataset, OxCGRT, which includes 19 pandemic policy indicators related to closure and containment, health, and economic policies and covers 184 countries.

Given that the existence of a variety of models makes it difficult for public officials and government to select a model to use, one potential approach is the use of an ensemble of models that works across multiple locations and that uses standard global parameter values for consistency. Sherratt et al. (2023) showed ensemble forecasts maximize the predictive performance of COVID-19 cases and deaths in every forecast across Europe (32 countries). The median ensemble methods outperformed those based on means. Paireau et al. (2022) first evaluated 19 predictors and 12 individual models, then built an ensemble model averaging across the models, outperforming the baseline model to anticipate COVID-19 healthcare demand in France. The ensemble model performed best on average with epidemiological and mobility predictors as the most promising predictors to improve the forecast. Cramer et al. (2022) first evaluated the performance of 27 individual models to forecast COVID-19 deaths

in the US. According to the results, their ensemble model outperformed all the stand-alone models that contributed to it, providing a reliable and accurate forecast.

2.3 Ensemble models for AI and ML

In this section, we discuss how ensemble AI and Machine Learning (ML) models can improve the performance of COVID modeling. Rahman et al. (2021) carried out a literature review of ML approaches for COVID-19. They identified four different ML application methods to combat COVID-19, all focusing on improving decision-making processes in a healthcare context for physicians, and policymakers, as well as identifying potentially infected people. They highlight the applicability of ML as a useful tool for analyzing, screening, tracking, forecasting, and predicting trends and characteristics of COVID-19.

Shastri et al. (2021) used a deep-LSTM ensemble model using convolutional and bi-directional LSTM. The model was able to forecast COVID-19 confirmed cases and deaths (for one month ahead) in India with the accuracy of 97.59 % and 98.88% respectively. This emphasizes the utility of AI for estimating the spread of COVID-19, especially in countries with large populations like India which need accurate predictions of COVID-19 spread, even when there is a shortage of healthcare workers to monitor how the disease is spreading, e.g., in the middle of a pandemic. Maaliw et al. (2021) showed that their proposed ensemble ML model combining Autoregressive integrated moving averages (ARIMA) and stacked long short-term memory networks (S-LSTM) outperforms each of the single models to forecast COVID-19 infections and deaths with an average accuracy of 90.73%. The model was validated by analyzing time series data of four countries including the Philippines, United States, India, and Brazil.

Tayarani-Najaran (2022) proposed an evolutionary algorithm with a surrogate ensemble learning algorithm to optimize government policies against the spread of the virus. The ensemble ML algorithm consists of ten base learning algorithms to improve performance. An Support Vector Machine (SVM) algorithm is built to predict the accuracy of each learning algorithm. The resulting model is used as a fitness function for the evolutionary algorithm. Jin et al. (2022) proposed a novel data-driven TCN-GRU-DBN-Q-SVM ensemble hybrid model based on Temporal Convolutional Networks (TCN), Gated Recurrent Units (GRUs), Deep Belief Networks (DBNs), Q-learning, and SVM models for COVID-19 infection prediction. The model provides satisfactory results which were verified against three national infection datasets from the UK, India, and the US to ensure the generalization of the proposed model. Ibrahim, Tulay, and Abdullahi (2023) proposed an ensemble ML approach called ANN ensemble (ANN-E) and SVM ensemble (SVM-E) for the purpose of predicting COVID-19 pandemic. Three standalone ML models including artificial neural network (ANN), adaptive neuro-fuzzy inference system (ANFIS), and SVM were used. The proposed ANN-E and SVM-E approaches outperformed all the other standalone methods to predict daily COVID-19 cases in 10 African countries.

2.4 Standardised data, standardised models, and automated model calibration

Sass et al. (2020) introduced a uniform, standard dataset called “German Corona Consensus Dataset” (GECCO) to support the interoperability of a COVID-19 dataset by using international terminologies and health IT standards.

Annan and Hargreaves (2020) calibrated an epidemic SEIR model to daily reported case numbers in the UK using a Markov Chain Monte Carlo algorithm to perform Bayesian parameter estimation. The results show a superior performance of the calibrated model for forecasting the early growth of the COVID-19 outbreak in the UK. This method can also estimate the effectiveness of lockdown policies.Nastasi et al. (2022) have applied SIRD (susceptible S, infected I, recovered R, and dead D) and SIRDV (vaccinated compartment V) models with calibration on real a dataset of cases of the COVID-19 spread in Great Britain (GBR) and Israel (ISR). The results highlight the effectiveness of vaccination campaigns to reduce the number of infected people and deaths.

3 Automatic calibration using a multi-objective genetic algorithm

P. S. Oh and S. J. Oh (2011) state that “a model plays the roles of describing, explaining and predicting natural phenomena and communicating scientific ideas to others.” Models take data as input and use this data to generate new data as output. Models usually have some input parameters that help define how a model works. When dealing with an epidemic or pandemic, input data takes the form of how the pandemic has been spreading so far, e.g., current case numbers, along with decisions on strategies, e.g., NPIs, to put in place to combat disease spread. Output data defines how well the selected strategies have worked in terms of slowing down or stopping disease spread. Many models also have input parameters that help determine how the model transforms the input data to output data. The accuracy of a model, i.e., how well the model’s output data matches corresponding data fro the real world is affected by the model’s input parameters. The process of model calibration involves changing a model’s input parameters so that the output data produced by the model – for a given set of input data – is as accurate as possible in terms of matching real world output data (according to a pre-defined measure of accuracy).

Calibrating models, including Covid spread models, is often challenging to do manually (Hazelbag et al., 2020; Kong, McMahon, and Gazelle, 2009). The relationship between model input parameters and model accuracy, i.e., how close the model outputs are to known values, is often complex and, hence, not intuitive. Automatic calibration of models is also not straightforward as the underlying optimisation problem is usually non-linear, often with inte-

ger decision variables and almost always with multiple local optima, i.e., multiple, disparate input parameter configurations that result in reasonably accurate models (Q. Wang, 1997). The properties of automatic model calibration problems mean that metaheuristic approaches, such as genetic algorithms, are an appropriate approach to automatically calibration (Kong, McMahon, and Gazelle, 2009).

In this report, we present a MOGA approach to calibrating Covid spread models. Given the number of possible input parameters that need to be calibrated, this is a difficult problem for any solution approach including MOGA. However, as discussed in §3.1, by calibrating multiple models and locations simultaneously, we get more data to inform input parameter calibration for those parameters that are consistent across models and/or locations. After discussing the MOGA-based calibration concept in §3.1, we present the ensemble of models being calibrated in §3.2 before defining a standard API for the models and describing the MOGA set-up in detail in §3.3 and §3.4 respectively. We then discuss how model predictions are integrated into a unified prediction with confidence intervals in §4. We describe how a model of economic performance can be added to the ensemble and associated calibration framework in §6.4.3 before finishing with a description of how various solutions, i.e., sets of input parameters, are evaluated in §3.5.

3.1 Calibration framework concept

Figure 3.1 represents how different models (in this figure Model 1 and Model 2) were developed and calibrated to (input and output) data in different locations (in this figure Locations A and B respectively). However, models were not often calibrated across multiple locations so, for example, Model 1 would not be able to be used in Location B without further calibration and possibly more development. In addition, parameters across the different models may differ, even if they represent the same value, since models were developed and calibrated independently. For example, the reproduction number of the alpha variant of COVID-19 may be different in Model 1, calibrated for Location A, and Model 2, calibrated for Location B.

By developing and calibrating models independently in different locations, some shared knowledge is not used appropriately to the detriment of the transferability and veracity of the models. Consider the example of the reproduction number of COVID-19 variant alpha, denoted R_0^α. When Model 1 is calibrated using the data from Location A, R_0^α may be affected by an increased population density and estimate a higher R_0^α than that from Model 2 which is being calibrated in Location B with a lower population density. Both estimates would work well for their respective models in the respective calibration locations, but transferring Model 2 to Location A would likely result in estimates of population spread that are too low (due to a low R_0^α) with the reverse occurring when transferring Model 1 to Location B.

Two factors need to be addressed in this example. First, R_0^α should be consistent across both models and locations. Location and/or model-specific changes to spread should be specified as separate parameters, i.e., location (specific) parameters or location and model (specific) parameters. In this example, population density parameters for each location could

be used within the two models to adjust the "effective reproduction number" of the alpha Covid-19 variant so that R_0^α is consistent across both locations and models. Second, the calibration of parameters should happen simultaneously so that consistent values for global parameters, that work best for all models in all locations, can be found as can consistent location parameters, i.e. parameters that should be consistent across different models such as the location's population density.

Figure 3.2 depicts a process that calibrates parameters consistently across locations and models as appropriate and that simultaneously calibrates all models, as an ensemble, for each location with input and output data for each location shared across all models and the accuracy of the ensemble, i.e., all the models, used to inform the calibration process.

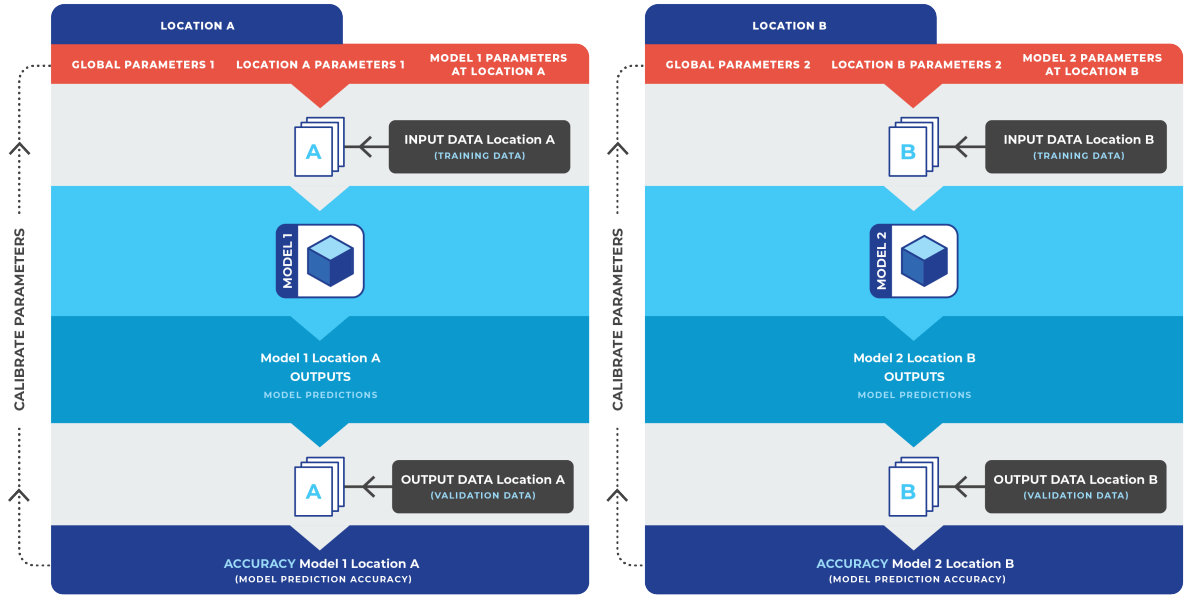


Figure 3.1: Models developed and calibrated individually

To achieve the approach depicted in figure 3.2, the input and output parameters for the models need to be standardised so that consistent calibration of the input parameters can be implemented and the ensemble accuracy of the models can be evaluated within this calibration. To determine these standards, we first summarise the models being considered in this research and identify which inputs are global, location (specific) and location/model (specific).

3.2 COVID-19 spread models within the calibration framework

For the ensemble calibration method presented in this research, three existing models for COVID-19 spread were considered. These three models were previously summarised by GPAI (2021). The revised summaries presented in this section have a change in notation

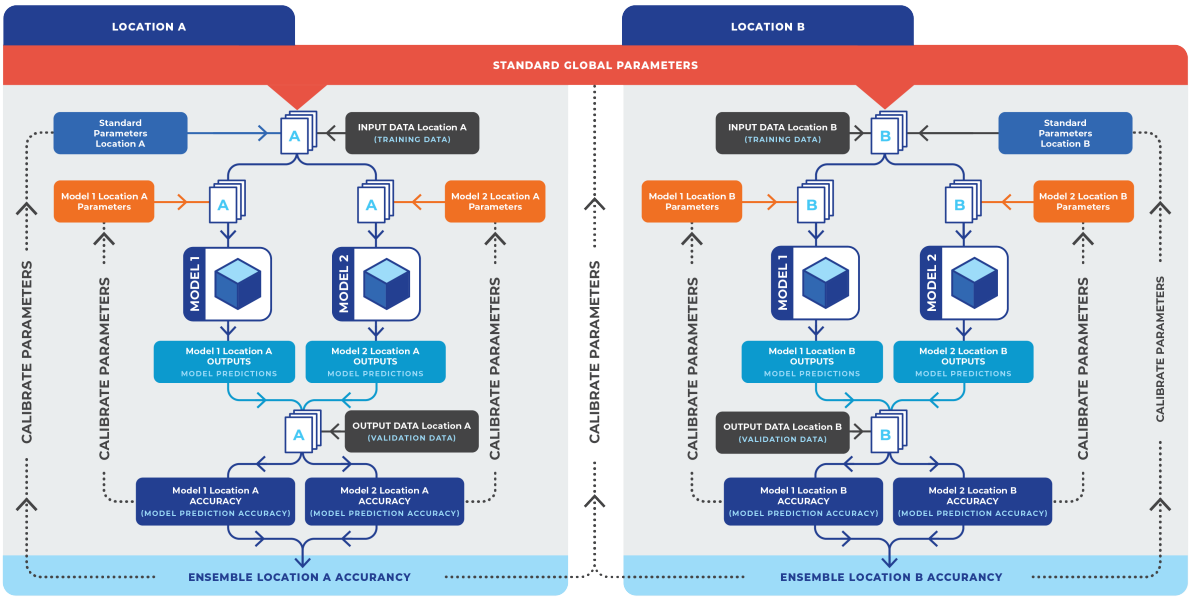


Figure 3.2: Models developed and calibrated as part of an ensemble

to make input parameters consistently labeled across the models and as a precursor for the input parameters standard coming in §3.3.1.

3.2.1 Effective Reproduction Number

Both the Cyber-Physical model – see §3.2.2 – and the CTMC model – see §3.2.3 – use the concept of "effective" reproduction number, i.e., the reproduction number for Covid-19 given: 1) the base reproduction numbers for the Covid-19 variants; 2) the proportion of each of the variants present in the population; 3) the level that each NPI is being used at; and 4) a function/model for calculating how the NPIs affect reproduction of each of the variants. In this initial version of ensemble model calibration, we use a simple model where:

$$R^v_{\text{eff}} = R^v_0 - \sum_{n \in \text{NPIs}} \tau(v, n, l(n, d))$$

where:

$$\begin{aligned} R^v_{\text{eff},d} &= \text{effective reproduction number of variant } v \text{ on day } d; \\ R^v_0 &= \text{base reproduction number of variant } v; \\ \tau(v, n, l) &= \text{effective coefficient on transmission for variant } v \text{ when NPI } n \text{ is at level } l; \\ l(n, d) &= \text{level of NPI } n \text{ on day } d. \end{aligned}$$

3.2.2 Cyber-Physical Model

Ro et al. (2020) used the plant-controller concept from Cyber-Physical Systems (Lee, 2008; Alur, 2015) with the plant representing the epidemiological model and the controller representing the NPIs being deployed to mitigate spread. The parameters for the plant and controller components of the model are described next.

3.2.2.1 Plant

The global parameters used by the (revised) plant are as follows:

- The reproduction numbers for each COVID-19 variant that are used within an estimated effective transmission model (which is location-specific):
  - R_0^α – the reproduction number of the α variant;
  - R_0^δ – the reproduction number of the δ variant;
  - R_0^o – the reproduction number of the o variant.
- The stages of Covid-19: S – Susceptible, E – Exposed, P – Presymptomatic; I_1 – Infected and contagious; I_2 – Infected and not contagious; R – Recovered; and D – Deceased.
- The transition rates are the reciprocal of the expected time in the "from" stages
  - From Exposed to Pre-Symptomatic;
  - From Pre-Symptomatic to Infected;
  - From Infected and contagious to Recovered – note that the Cyber-Physical method does not include the Infected and not contagious stage.
- The relative infectiousness when Pre-Symptomatic (ϵ).

The location (specific) parameters for the plant are as follows:

- The proportion of Covid-19 variants present in the population, p_α, p_δ, p_o and are assumed to be constant for the time horizon being modelled.
- The effective transmission function/model, see §3.2.1. This model converts the reproduction numbers for each variant into effective production numbers given the NPI levels that are in place:
  - $R^\alpha_{\text{eff},d}$ for α;
  - $R^\delta_{\text{eff},d}$ for δ; and
  - $R^o_{\text{eff},d}$ for o.

- The case fatality ratios:
  - When there are available Intensive Care Unit (ICU) beds;
  - When there aren't available ICU beds.
- The population size and the initial numbers of the population in each stage.
- The horizon of the model in terms of start and finish dates and step size (a.k.a. interval) in days.
- The proportion of cases that require ICU care.
- The number of available ICU beds.

The location and model (specific) parameters are:

- The logging rate – how often the plant (epidemiological model) logs results such as the number of cases and number of deaths.

3.2.2.2 Controller

In the (revised, standardised) Cyber-Physical model the decisions that are controlled for the Plant – by the Controller – have been replaced by the NPI schedule. The NPI schedule informs:

1. The effective reproduction number;
2. The reproduction number of isolated cases; and
3. The testing rate.

All 3 values are determined by model-based parameters – see 1.3.1.2 – and a model that converts these parameters into values from the NPI schedule – see §3.2.1.

3.2.3 Continuous Time Markov Chain

The (revised, standardised) Continuous Time Markov Chain (CTMC) model uses similar states to the Cyber-Physical model (GPAI, 2021, Appendix A.3). However, the states from §3.2.2.1 are extended to include the ward stay in hospital as shown in Figure 3.3, i.e., the extra stages of Covid-19: W_1 – Hospital ward before any ICU stay; W_2 – Hospital ward after an ICU stay; and V – ICU. Note that $R_{\text{eff},d}$ is determined by the model for the effect the NPI schedule has on the effective reproduction number of Covid-19 – see §3.2.1. The CTMC uses these extended stages and associated transition values to determine the number of individuals in each stage on each day in the modelling horizon.

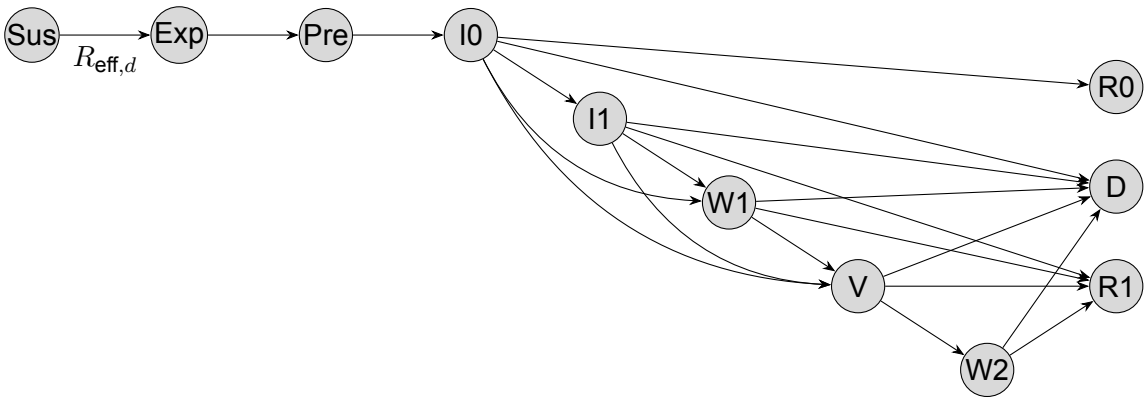


Figure 3.3: The states in the CTMC model for Covid-19 spread

Note that, in addition to most of the location (specific) parameters used by the Cyber-Physical model, the CTMC model uses extra location (specific) parameters, namely:

- Proportion of hospitalised (ward) patients that go to ICU;
- Relative infectiousness of patients in a ward;
- Relative infectiousness of patients in an ICU;
- Proportion of patients that go to hospital;
- Estimated length of stay in hospital;
- Estimated length of stay in ICU.

Unlike the Cyber-Physical model, the CTMC model assumes that everyone who is symptomatic will be tested. Note that this assumption was based on the freely available testing in Aotearoa | New Zealand and will likely not hold in all locations.

3.2.4 LSTM Models

The LSTM models consist of the v1 LSTM and the v2 Conditional LSTM. They are statistical models that learn disease spread patterns from historical data. Using machine learning, the models are trained to predict the time series of cases in previous data, provided with context data including the intervention plans and the recent history of cases. In the scenarios described in the report, the models are provided with similar context information and tasked to extend their predictions to the unseen scenario pandemic period. Thus the models apply knowledge gained from historical data to predict the likely progression of the pandemic in a real scenario.

The two models utilize LSTMs (Graves, 2012), which are a standard machine learning practice used to learn statistical associations on sequences. Internally, the LSTMs maintain a memory that is modified sequentially as the LSTM processes each day of the pandemic scenario. The LSTM has operations to reset its memory and to write new information to its

memory. It uses this memory to predict the number of cases as a statistical variable. During training, the prediction errors are utilized via backpropagation to modify what information is written to the memory and when it is written and reset, along with how the memory is associated with the model predictions. In this way, LSTMs learn salient statistical patterns from the input data, consisting of the context, and they learn to associate these patterns with the likely progression of the pandemic.

Two such statistical models were introduced in the previous report. The first model, the v1 LSTM, is a small statistical model designed to learn important patterns without overfitting, a common danger with larger models. Several innovations went into the design to produce the v2 Conditional LSTM model:

1. The architecture of the model was revised so that later layers became conditioned on the intervention plans so that these plans can have more complex nonlinear interaction with the predicted pandemic progression.
2. A new preprocessing component was added, so that the model became more robust to noise and variation in the baseline number of cases between countries.
3. The model was revised to be deeper and wider, allowing for more complex statistical associations to be learned.

To train the models, all available data prior to the scenario period is used. Thus the scenario period represents unseen data, consistent with the realistic scenario of needing to make predictions in an unseen context while utilizing all available data.

3.3 Standardising the COVID-19 models

In order for all models to be calibrated within the same framework, the interface, i.e., the input data and parameters and the output data, must be the same, i.e., standardised (even if some models choose not to use some data/parameters). One of the key contributions of this calibration framework is the standardisation of inputs/outputs across 3 different modelling approaches.

3.3.1 Input standard

The input standard is used by all 3 of the models included in this research, namely the LSTM model – see §3.2.4, the Cyber-Physical model – see §3.2.2, and the CTMC model – see §3.2.3. The LSTM model only uses the defined input data for each location – see the standard 1.2.2. The Cyber-Physical and CTMC models only use the most recent data from each location, i.e., the current number of cases and deaths. However, they both have model-specific parameters. Some of these parameters are shared, e.g., the reproduction numbers for the COVID-19 variants – see the standard 1.1.1.2.1, the number of ICU beds in each

location – see the standard 1.2.1.7.1, but some are only relevant for a particular models, e.g., extra stages for COVID-19 patients in hospital – see standard 1.2.1.3.

The full definition of the input standard is given in Table 3.1. These input standards are provided to the calibration framework using a JSON file. For example, the NPI list and the max NPI level and variant list are shown in Figure 3.4. Some of the more complex input standards are described in greater detail in §3.3.2 following Table 3.1.

Table 3.1: Input Standard Definitions

1 Inputs	**1.1 Global Inputs**	**1.1.1 Global Input Parameters**	1.1.1.1 List of NPIs	*1.1.1.1.1 NPI name/code* *1.1.1.1.2 Max NPI level (integer, currently 4)*
			1.1.1.2 List of variants (e.g., alpha, delta and omicron)	*1.1.1.2.1 Estimated "true" reproductive numbers for variant*
			1.1.1.3 Hierarchy of stages of Covid infection	*1.1.1.3.1 Stage name/symbol* *1.1.1.3.2 Transition rates between stage of Covid infection (dependent on the ICU being under or over capacity, IFR0 and IFR1 respectively)* *1.1.1.3.3 Relative infectiousness of Presymptomatic with respect to Infectious*
			1.1.1.4 Countries modeled	*List of countries to produce model outputs for*
	1.2 Location Inputs	**1.2.1 Location Input Parameters**	1.2.1.1 Proportion of variants	*The proportion of each of the variants at the locations, listed by country*
			1.2.1.2 Effective transmission rates	
			1.2.1.3 Extra stages of Covid infection for location's health system	

Table 3.1: Input Standard Definitions (continued)

			1.2.1.4 Estimates of total population size and number of population in each stage (including extra stages from 1.2.1.3)	
			1.2.1.5 Time periods for model horizon	
			1.2.1.6 NPI schedule - level of NPI for each time period in the model horizon	
			1.2.1.7 Health system parameters (lists for each country)	*1.2.1.7.1 Number of ICU beds* *1.2.1.7.2 Proportion of hospitalised (ward) patients that go to ICU* *1.2.1.7.3 Relative infectiousness of patients in a ward* *1.2.1.7.4 Relative infectiousness of patients in an ICU* *1.2.1.7.5 Proportion of patients that go to hospital* *1.2.1.7.6 Estimated length of stay in hospital* *1.2.1.7.7 Estimated length of stay in ICU*
		1.2.2 Location Input Data	1.2.2.1 Time periods for historical data	
			1.2.2.2 Case numbers for each time period for the previous 3 months-1 year depending on availability	

Table 3.1: Input Standard Definitions (continued)

			1.2.2.3 Deaths for each time period for the previous 3 months-1 year depending on availability	
			1.2.2.4 Level of NPI for each historical time period	
	1.3 Model Specific (a.k.a. Model Inputs)	**1.3.1 Model Input Parameters**	1.3.1.1 List of models being used	
			1.3.1.2 Table of values related to NPIs	• *Coefficients of effective transmission* • *Coefficients for transmission of a confirmed case* • *Coefficients for testing rates*
			1.3.1.3 Output locations for the models	

3.3.2 Detailed descriptions of complex input standards

This subsection contains a more detailed description the more complex input standards, the numbering is consistent with the standard presented in Table 3.1.

1.1.1.1.2 Max NPI level (integer, currently 4)

See Oxford Covid-19 Government Response Tracker (OxCGRT) codebook. In the GitHub repository under data/time series, there are CSV files for each NPI. Instead of Economic effects, the tracker monitored countries that used selected economic policies which include:

```
E1_income support                E3_fiscal measures
E2_debt/contract relief          E4_international support
```

```
{
  "global": {
    "parameters": {
      "npi_list": {
        "code": "1.1.1.1",
        "list": [
          "C1_School closing",
          "C2_Workplace closing",
```

⋮

```
          "V3_Vaccine Financial Support (summary)",
          "V4_Mandatory Vaccination (summary)"
        ],
        "max_npi_level": {
          "code": "1.1.1.1.2",
          "value": 4
        }
      },
      "variant_list": {
        "code": "1.1.1.2",
        "list": [
          "alpha",
          "delta",
          "omicron"
        ]
      },
      "reproductive_numbers": {
        "code": "1.1.1.2.1",
        "rates": [{
            "variant": "alpha",
            "R0": 2.79,
```

⋮

Figure 3.4: Snippet from example JSON in §A.3

Countries that did not implement economic policies are assigned 0. Flags are used to determine if a policy applied to a proportion of a population or not. However, economic policies and flags were not used because it was assumed that it did not have a significant impact on confirmed cases and deaths in the countries that were being studied. Predicting and controlling confirmed cases and deaths were the priority at this time. However, using the flags and economic effects may have provided a more accurate view. It is helpful to take note of how some NPIs have cross impacts and as a result of combining NPIs may lead to double changes to predictions.

1.2.1.2 Effective transmission rates This would be a model or function that estimates the effective transmission rates between Covid stages at a particular location taking into account the transmission rate between stages and the levels of NPIs in place. It needs to be consis-

tent across models.

3.3.3 Output standard

The output standard is also used by all 3 of the models included in this research, namely the LSTM model – see §3.2.4, the Cyber-Physical model – see §3.2.2, and the CTMC model – see §3.2.3. It defines what outputs each model should produce, e.g., case number estimates, and also the output data that is available for calibrating the models.

As with the input standard – see §3.3.1, the full definition of the output standard is given in Table 3.2. These output standards are also provided to the calibration framework using a JSON file.

Table 3.2: Output Standard Definitions

2 Outputs	**2.1 Location Outputs**	**2.1.1 Location Output Estimates**	2.1.1.1 Location information
			2.1.1.2 Time periods for model horizon
			2.1.1.3 Case number estimates for each time period in the model horizon
			2.1.1.4 Recovered number estimates for each time period in the model horizon
			2.1.1.5 Death estimates for each time period in the model horizon
		2.1.2 Location Output Data	2.1.2.1 Time periods for historical data
			2.1.2.2 Case numbers for each time period for the previous 3 months-1 year depending on availability
			2.1.2.3 Deaths for each time period for the previous 3 months-1 year depending on availability
			2.1.2.4 Level of NPI for each historical time period

3.4 Setting up the Multi-Objective Genetic Algorithm

Genetic Algorithms (GAs) maintain a set of solutions and look to improve the objective function values of these solutions. Each solution is referred to an individual or chromosome, made up of distinct pieces called genes. Holland's first version of GAs assumes genes to

be binary digits (Holland, 1992), but later versions included a wider range of gene types. Mapping links solutions to their corresponding chromosome and GA operators create new chromosomes, hence solutions, and the GA evolves the set of solutions, a.k.a. population, using their objective functions values, a.k.a. fitness. GA selection techniques vary based on the fitness values used with the most commonly used selection methods including proportional, ranking, and tournament.

For MOGAs, each solution provides a vector of objective values that are not dominated by any other solution, i.e., that are non-dominated (Konak, Coit, and Smith, 2006).GAs are customised to address multiple objectives by utilizing specialized fitness functions that enable a population of solutions to evolve in a way that maintains the non-dominated property of a population. In this research we use the pymoo package (Blank and K. Deb, 2020) which is based on the NSGA-II algorithm (Kalyanmoy Deb et al., 2002).

3.4.1 Pymoo

This section is summarised from the pymoo documentation (Blank and K. Deb, 2020).

The general formulation of a multi-objective optimization problem is given as:

$$\begin{aligned} \min \quad & f_m(x) && m = 1, \ldots, M \\ & g_j(x) \leq 0 && j = 1, \ldots, J \\ & h_k(x) = 0 && k = 1, \ldots, K \\ & x_x^L \leq x_i \leq x_i^U && i = 1, \ldots, N \\ & x \in \Omega \end{aligned}$$

where Ω is the set of valid values for x.

The example shown by Blank and K. Deb (2020) is:

$$\begin{aligned} \min \quad & f_1(x) = 100(x_1^2 + x_2^2) \\ \max \quad & f_2(x) = 1(x_1 - 1)^2 - x_2^2) \\ & g_1(x) = 2(x_1 - 0.1)(x_1 - 0.9) && \leq 0 \\ & g_2(x) = 20(x_1 - 0.4)(x_1 - 0.6) && \geq 0 \\ & -2 \leq x_1 \leq 2 \\ & -2 \leq x_2 \leq 2 \\ & x \in \mathbb{R}^2 \end{aligned}$$

For the calibration framework, each solution x consists of values for a subset of the input parameters from the input standard – see §3.3.1 – as described by a calibration standard next in §3.4.2. No explicit constraints are used, but both proportions and population parameters are adjusted to sum to 1 and the total population respectively before being used within the model ensemble for the objective function, a.k.a. fitness, calculations. Within the NSGA-II algorithm, the objective function values are calculated from the solution set, the ensemble of models, and the model performance metrics as follows. First, each solution is mapped

to its corresponding input parameters. For those models in which the model performance will change with the input parameters, the models are run with these new input parameters. The new performance metrics are calculated for each model and the ensemble metrics are then calculated. The final ensemble metrics are returned to the NSGA-II algorithm as the objective function values for the given solution, i.e., input parameters.

Next is a summary of the calibration framework problem solved using pymoo with representative bounds provided for the given parameters (the reproduction number for the α and δ variants respectively).

$$
\begin{aligned}
\min \quad & f_1(x) = \text{ensemble metric from daily cases prediction ranking} \\
\min \quad & f_2(x) = \text{ensemble metric from daily deaths prediction ranking} \\
& 2.29 \leq R_0^{\alpha} \leq 3.29 \\
& 4.58 \leq R_0^{\delta} \leq 5.58 \\
& \vdots \\
& x \in \text{set of feasible input parameters}
\end{aligned}
$$

Note that feasibility of the input parameters includes considerations such as proportions must add to 1, initial estimates of the amount of the population in each disease stage, e.g., that are presymptomatic, must sum to match the total population.

In this research, the LSTM models are not affected by changes in input parameters because they use historical data only. Both the Cyber-Physical and CTMC models are affected by changes in input parameters and must be run for each solution identified by the NSGA-II algorithm. The performance metrics considered are the ranking of each model for predicting new daily cases and predicting new daily deaths. The average across all the models for each metric is used as the ensemble metric for given input parameters. If no prediction – hence ranking – is available, e.g., the ensemble does not include that model in the average calculation. This is the case for the LSTM models which don't predict new daily deaths.

Hence, the calibration framework searches the input parameters space to find the set of parameters that best predict new daily cases and new daily deaths, where best is defined in a multi-objective way, i.e., the prediction of one metric does not improve without a corresponding deterioration in the other metric.

As mentioned previously, the implementation of the calibration framework uses a standard, similar to the implementation of the ensemble of models, so that it can be reasonably easily extended in the future. This standard is presented next.

3.4.2 Calibration standard

The calibration standard is used so that input parameters being calibrated, within the calibration framework using the ensemble of models, are well defined and can be easily expanded in the future. The full standard is defined in table 3.3 and is a subset of the input parameters along with some extra information on how the parameters are allowed to change during

calibration, i.e., how the input parameter space is defined.

1 Inputs	**1.1 Global Inputs**	**1.1.1 Global Input Parameters**	1.1.1.2 List of variants (e.g., alpha, delta and omicron)	*1.1.1.2.1 Estimated "true" reproductive numbers for variant* *Bounded, so value calibration will search between [lower bound, upper bound], e.g.,* α *[2.29, 3.29]* δ *[4.58, 5.58]* o *[9, 10]*
			1.1.1.3.3 Relative infectiousness of Presymptomatic with respect to Infectious	*Bounded, usually [0, 1]*
	1.2 Location Inputs	**1.2.1 Location Input Parameters**	1.2.1.1 Proportion of variants	*The proportion of each of the variants at the locations* *Bounded,usually [0, 1], but must sum to 1*
			1.2.1.2 Effective transmission rates	*The function won't change during calibration, but parameters that define it – see 1.3.1.2 – will change*
			1.2.1.4 Estimates of total population size and number of population in each stage (including extra stages from 1.2.1.3)	*The total population won't change, but the initial numbers in each stage are bounded, but must sum to = the total population. Note that the bounds are set using the change proportion parameter, a calibration run parameter, but all initial number values may be scaled down to ensure the total is = the total population*

			1.2.1.7 Health system parameters (lists for each country)	*1.2.1.7.1 Number of ICU beds* *Bounded, e.g., [510, 530] for Aotearoa | New Zealand* *1.2.1.7.2 Proportion of hospitalised (ward) patients that go to ICU* *Bounded, within [0, 1]* *1.2.1.7.3 Relative infectiousness of patients in a ward* *Bounded, within [0, 1]* *1.2.1.7.4 Relative infectiousness of patients in an ICU* *Bounded, within [0, 1]* *1.2.1.7.5 Proportion of patients that go to hospital* *Bounded, within [0, 1]* *1.2.1.7.6 Estimated length of stay in hospital* *Bounded, e.g., [18, 25] for Aotearoa | New Zealand* *1.2.1.7.7 Estimated length of stay in ICU* *Bounded, e.g., [15, 21] for Aotearoa | New Zealand*
	1.3 Model Specific (a.k.a. Model Inputs)	**1.3.1 Model Input Parameters**	1.3.1.2 Table of values related to NPIs	• *Coefficients of effective transmission* *Minimum and maximum values are bounded and linear interpolation used for values in between*

			• *Coefficients for transmission of a confirmed case* *Bounded, with a "plus/minus" value to ensure they don't change too much from initial estimates* • *Coefficients for testing rates* *Bounded, also with a "plus/minus" value to ensure they don't change too much from initial estimates*

Table 3.3: Calibration Standard Definitions

Figure 3.5 shows a snippet of the JSON file that implements the calibration standard within the calibration framework, including a typical bounded parameter (R_0 for the α variant) and a more customised parameter (the effective transmission coefficients which are allowed to vary by $\pm 10\%$ during calibration).

```
"global": {
  "parameters": {
    "reproductive_numbers": {
      "code": "1.1.1.2.1",
      "rates": [
        {
          "variant": "alpha",
          "R0": { "type": "bounds", "lb" : 2.29, "ub": 3.29 }
        },
```

⋮

```
"model": {
  "effective_transmission_coeffs": {
      "code": "1.3.1.2",
      "file": { "type": "custom", "pm": 0.1 },
```

⋮

Figure 3.5: Snippet from example JSON in §A.5

The code snippet in figure 3.6 shows how the calibration standard information is used to set up the solution space for the calibration framework. The setup for both typical bounded

parameters (in the first snippet) and customised ± parameters (in the second snippet) are shown.

A code snippet that shows how a calibration framework MOGA problem is defined from the input and calibration standard JSON files and then the NSGA-II algorithm is configured to solve the problem is given in figure 3.7. The size of the solution set (`NPOP`), the number of new parameter sets (`NOFF`), and number of generations (`NGEN`) can be defined. One other key configuration value is the change proportion (of the total population) allowed for the initial number of the population in each disease stage. The closer this is to 1.0, the more flexibility for these parameters, but also the larger the overall solution space (and hence more generations required to search the space).

Using the calibration framework code, the pymoo package and the JSON files, an ensemble of models can be calibrated across multiple locations simultaneously. The rest of this section discusses the efficacy of a calibrated ensemble of models for making predictions before two case studies of the calibration framework in use are given in §5.

3.5 Evaluating solutions

The models produce a predicted number of cases for each day in the scenario period. These numbers are compared against the true number of cases for the same day. To quantify the quality of the predictions, we use the standard mean average error (MAE) metric, which measures the deviation of the model predictions from the truth. There is a Jupyter notebook available, produced with the last report, which does this error computation and also produces plots of the model predictions and ground truth number of cases and deaths.

4 Ensembling the models

This report has described diverse modeling approaches incorporated in an ensemble. But how is the multitude of produced predictions converted into usable insights? To help summarize the information we introduce an ensemble prediction, thus integrating the various sources of predictions into a summary prediction.

This report introduces the RIO ensemble method. A description of ensembles was provided in a previous GPAI report (GPAI, 2023), and in this report we improve upon the previous work using Residual Estimation with I/O Kernels (Qiu, Meyerson, and Miikkulainen, 2020). These models offer two capabilities in service of ensemble integration: (1) error correction and (2) uncertainty estimation.

This section begins by reviewing related work in uncertainty estimation and then details the implementation of the RIO-based ensembles. In Section 5.3 a description of the results of ensembling is provided, documenting how error correction and uncertainty estimation affect the results.

```
          # 5a. If the variable is changing using bounds
          if key[-1] in ["bounds", "integer"]:
            # Get the "reduced" key (for use in adding to the temp JSONs)
            reducedKey = key[0:len(key) - 2]

            # Update the number of parameters
            numParms += 1
            # Get the lower and upper bound values from the calibration JSON file
            lb = get_from_dict(calibrateJSON, reducedKey + ['lb'])
            if key[-1] == "bounds":
              ub = get_from_dict(calibrateJSON, reducedKey + ['ub'])
            else:
              # Increase the interval by 1 for integer bounds
              ub = get_from_dict(calibrateJSON, reducedKey + ['ub']) + 1
            # Add the key to the list
            if key[-1] == "integer":
               reducedKey = ["integer"] + reducedKey
            keys.append(reducedKey)
            # Add the lower and upper bound values
            xll.append(lb)
            xul.append(ub)
```

⋮

```
            # Save a list of confirmed transmission coefficients which are
            # ("conf_coeff", country, npi, variant, level) pairs
            self.confcoeffs = []
            # Loop over the countries
            for country in countries:
              for npi in npis:
                for variant in variants:
                  for l in range(maxLevels + 1):
                    df = coeffsDf.loc[(coeffsDf['CountryName'] == country) &
                                      (coeffsDf['NPI_Oxford'] == npi) &
                                      (coeffsDf['Variant'] == variant),
                                      "conf_coeff_%d" % l]
                    if df.shape[0] == 1:
                      val = df.item()
                      if not np.isnan(val):
                        # There is some effect for this NPI
                        # Set the bounds for the maximum coeff country & NPI
                        lb = max(0.0, val * (1 - pmValue))
                        ub = min(1.0, val * (1 + pmValue))
                        # Create a reduced key for use during evaluation
                        reducedKey = ["conf_coeff", country, npi, variant, l]
                        # Add this key to the list of initial count keys
                        self.confcoeffs.append((country, npi, variant, l))
                        # Increase the number of parameters being calibrated
                        numParms += 1
                        # Add the key
                        keys.append(reducedKey)
                        # Add the bounds
                        xll.append(lb)
                        xul.append(ub)
```

⋮

Figure 3.6: Code snippet for setting up variable bounds used within MOGA

⋮

```
if __name__ == "__main__":
    LOGGER.info(f"=====> STARTED CALIBRATION RUN")
    LOGGER.info("Calibration hyper parameters: pop = {NPOP}, off = {NOFF}, gen = {NGEN}, chg = {
    POP_PROP_INC}")

    # Create a CalibationProblem with input and calibration JSON files
    problem = CalibrationProblem("examples/instance%d/input%d.json" % (INSTANCE, INSTANCE),
                                 "examples/instance%d/calibrate%d.json" % (INSTANCE, INSTANCE))

    # Set up the NSGA-II algorithm to solve the CalibrationProblem
    algorithm = NSGA2(
        # How many parameter sets are being considered in each iteration
        pop_size=NPOP,
        # How many new parameter sets to create each iteration
        n_offsprings=NOFF,
        # Other NSGA-II parameters affected how new solutions are created (sampling),
        # how much existing solutions interact to create new solutions (crossover),
        # how much existing solutions change (mutation) and whether duplicates are
        # allowed
        sampling=FloatRandomSampling(),
        crossover=SBX(prob=0.9, eta=15),
        mutation=PM(eta=20),
        eliminate_duplicates=True
    )


    # Set how many iterations to run for
    termination = get_termination("n_gen", NGEN)

    # Solve the CalibrationProblem using NSGA-II
    res = minimize(problem,
                   algorithm,
                   termination,
                   seed=1,
                   save_history=True,
                   verbose=True)
```

⋮

Figure 3.7: Code snippet for creating MOGA instance

4.1 Overview of uncertainty estimation

Uncertainty estimation refers to evaluating how much trust should be placed in the produced predictions of a model given what is known about the model and the context it is making predictions. Many strategies have been proposed to address the question of knowing what a model knows, including ensemble methods (Gawlikowski et al., 2023). Indeed, ensembles of networks are able to detect out-of-distribution data points, which are more prone to error, by reporting a higher predictive uncertainty (Lakshminarayanan, Pritzel, and Blundell, 2017).

Crucial to the success of these combination strategies is the diversity between models, thus motivating our work to apply a variety of modeling techniques. It is well-documented in machine learning that combining multiple models improves performance (Mohammed and Kora, 2023). In our work, some models (the LSTM and Conditional LSTM) model the statistical progression of pandemics, while other models (the CTMC and Cyber-Physical models) incorporate domain-specific knowledge. Therefore, the kinds of biases these models will make when applied to the data are very diverse, leading to a strong final ensemble.

4.2 The residual estimation approach

The panel of expert systems represented in the ensemble produces a variety of different predictions, so how do we know who to trust and when? To integrate the predictions from each model in the ensemble, first an uncertainty estimate is derived for each model prediction so that models that are more prone to error produce a more diffuse contribution to the final ensemble result. The process is illustrated in Figure 4.1.

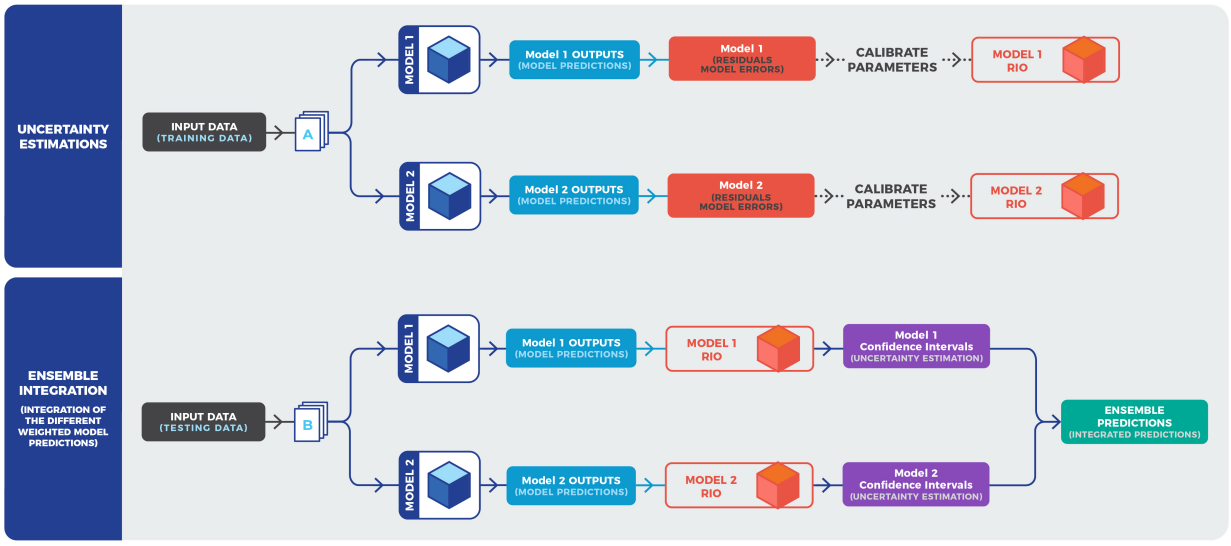


Figure 4.1: The ensemble is formed using a collection of predictive models along with corresponding RIO models. In the top flow, the RIO models are calibrated on training data to learn the error patterns of their respective models. In the bottom flow, the uncertainty estimates are used to construct confidence intervals for each model, which are then integrated into a final ensemble prediction.

To form the uncertainty estimates, residual estimation via Gaussian processes is applied (Qiu, Meyerson, and Miikkulainen, 2020). These estimators learn in what situations a model performs well or poorly and assign a confidence estimate appropriately. They do this by forming a statistical model of the error residuals, constituting not only a corrective estimate of the residual, which can be used to adjust the model prediction and reduce error, but also an estimate of the uncertainty in that residual estimate, indicating the confidence of the prediction given the current context.

To train the RIO models, several rollout predictions are produced on the training data prior to the scenario. Given these rollouts, the RIO model is trained to produce uncertainty estimates for each model. These uncertainty estimates take the form of Gaussian distributions. The final ensemble prediction is then represented as a mixture of Gaussians distribution, taking into account the contributions and uncertainty of each model.

5 Experiments and results

To test the calibration framework described in §3, two case studies were considered. For each case study, the standard input, output, and calibration files were defined. Further parameters for the GA used for calibration were also defined across both case studies, namely:

1. for a short calibration run:
    - Five individuals in the solution population, i.e., five sets of input parameters were tested at each generation of the GA;
    - Two offspring were created, i.e., two new sets of input parameters were created and added to the five sets in the population before the selection of the fittest individuals reduced the population back to five;
    - Three generations were run, so the creation of offspring and selection of the fittest was performed three times; and
    - 10% of the total population was used as the maximum amount of change. i.e., defined bounds for the GA, from initial values in each disease stage;

    and
2. for a long calibration run: 20 individuals, 15 offspring, 500 generations, 50% of total population allowed for change from initial values, hence more solutions and change within each generation and more generations for the solutions to evolve and improve.

For each case study, there are four sets of plots:

1. The seven-day average of predicted new cases;
2. The rankings of the models (i.e., given models with specific parameters) for predicting new cases;
3. The seven-day average of predicted new deaths; and
4. The rankings of the models for predicting new deaths.

For the rankings, lower is better as it uses mean average error to rank the models. For the predictions, getting as close as possible to the actual data, labeled “Ground Truth” is desirable.

Case study 1 considers Aotearoa | New Zealand, Kenya, Sweden, and the United Kingdom from March to June of 2020 – see §5.1. Case study 2 considers Aotearoa | New Zealand and

Sudan from June to September 2020. These case studies were chosen by the authors as they were deemed interesting in terms of the diversity of the countries and their COVID-19 experience as well as being familiar to some of the authors.

5.1 Case study 1: Aotearoa | New Zealand, Kenya, Sweden and the United Kingdom

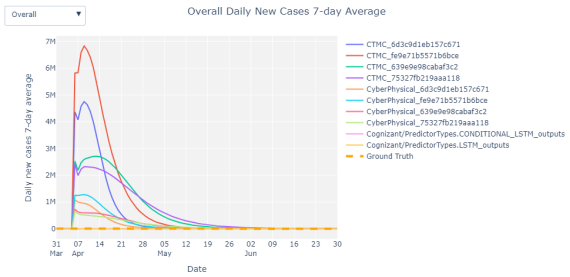
Figures 5.1 and 5.2 compare the prediction of new cases for the (Aotearoa | New Zealand, Kenya, Sweden, United Kingdom) case study between a preliminary, short calibration run (5 individuals, 2 offspring, 3 generations, 10% of the total population allowed for change from initial values) and a more thorough, long calibration run (20 individuals, 15 offspring, 500 generations, 50% of the total population allowed for change from initial values). Note that the Cyber-Physical and CTMC algorithm labels have a unique code appended that links to the input parameters being used. Figure 5.2 zooms in to a smaller time period from Figure 5.1. Figure 5.3 gives rankings for the case predictions for each of the models over time. In Appendix A.1, Figures A.1 and A.2 show similar metrics, i.e., predictions and rankings respectively, for predicting new deaths overall and in each country.

For overall new cases, Cognizant/PredictorType.LSTM shows the best performance for both short and long calibrations. In April, all the models (except LSTM) with short calibration overestimated the daily new cases. The performance improves from the middle of May – see Figure 5.1a. However, the models for long calibration underestimate the overall new cases – see Figure 5.1b.

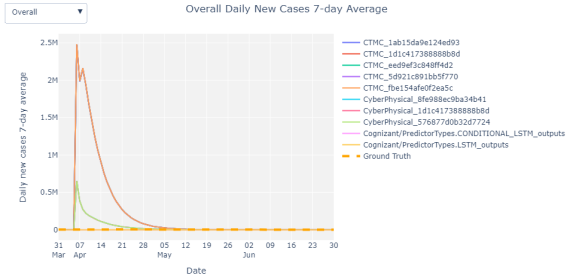
For each of the countries considered, the pattern is similar to the overall pattern, namely that the models other than the LSTM models over-predict new cases at the start of the horizon before dropping close to the Ground Truth. The key difference between the short and long runs is the magnitude of the over-prediction. Figure 5.2 shows there is a significant decrease in the over-prediction by all the Cyber-Physical and CTMC models with more calibration. In all plots, the Cyber-Physical models perform better than the CTMC models.

Figure 5.3 confirms these observations with the LSTM models outperforming the Cyber-Physical models which outperform the CTMC models although that ranking stabilises which shows the mean average errors are dropping to 0.

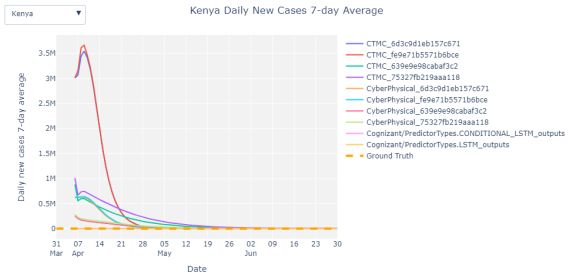
As can be seen in Appendix A.1, when predicting new deaths, the results are quite different. First, the LSTM models have not been configured to predict new deaths so they drop out of consideration. However, the model performance – see Figure A.1, hence rankings – see Figure A.2, are reversed with the CTMC models providing more accurate predictions than the Cyber-Physical models. As with the new cases, the long calibration run produces better results, i.e., more accurate predictions. This is expected as these models have been better calibrated.



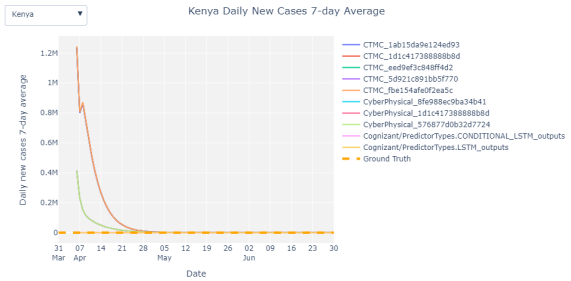
(a) Overall new cases – short



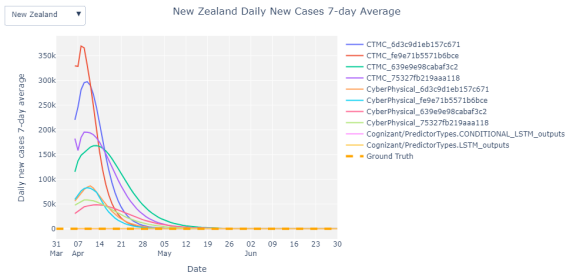
(b) Overall new cases – long



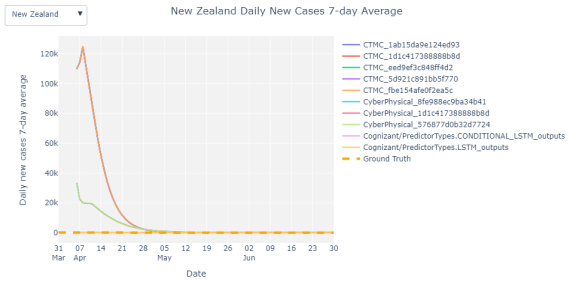
(c) Kenya new cases – short



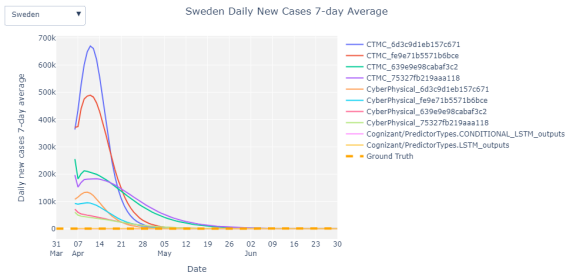
(d) Kenya new cases – long



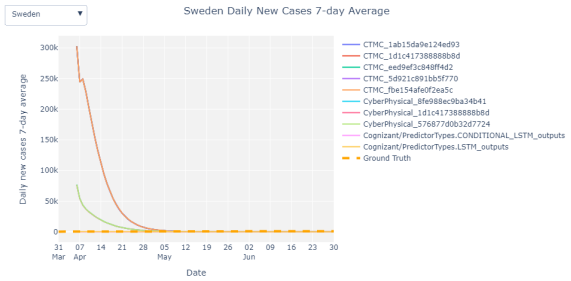
(e) Aotearoa | NZ new cases – short



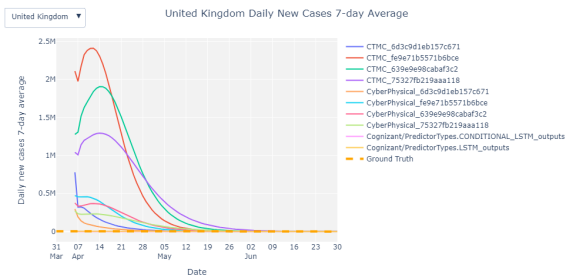
(f) Aotearoa | NZ new cases – long



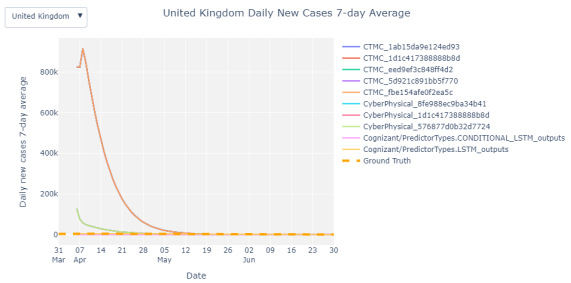
(g) Sweden new cases – short



(h) Sweden new cases – long

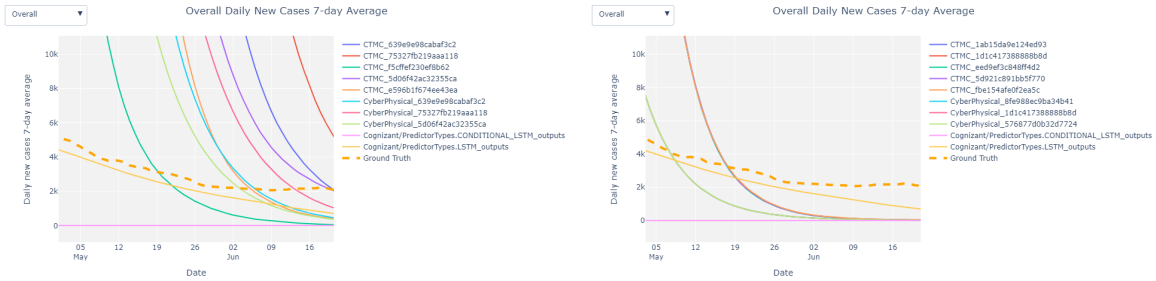


(i) United Kingdom new cases – short



(j) United Kingdom new cases – long

Figure 5.1: Comparing new cases metric: short calibration run (5 individuals, 2 offspring, 3 generations, change proportion = 0.1) on the left, long calibration run (20 individuals, 15 offspring, 500 generations, change proportion = 0.5) on the right. Note that the Cyber-Physical and CTMC algorithm labels have a unique code appended that links to the input parameters being used



(a) Overall new cases – short, May 5 to June 16

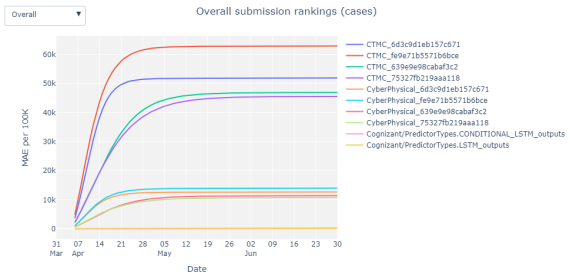
(b) Overall new cases – long, May 5 to June 16

Figure 5.2: Comparing new cases metric: short calibration run for May 5 to June 16 on the left, long calibration run for the same time period on the right.

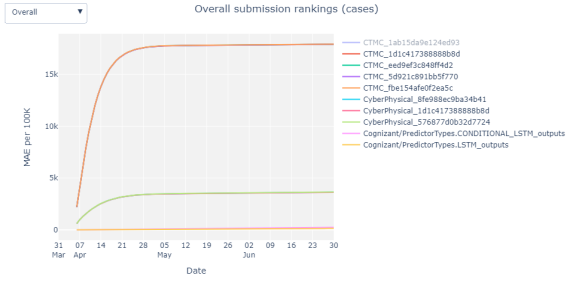
5.1.1 Manual calibration

Figure 5.4 shows the metrics for the Aotearoa | New Zealand case study, with Cyber-Physical and CTMC models resulting from a long calibration shown alongside LSTM models and a manually calibrated Cyber-Physical model.

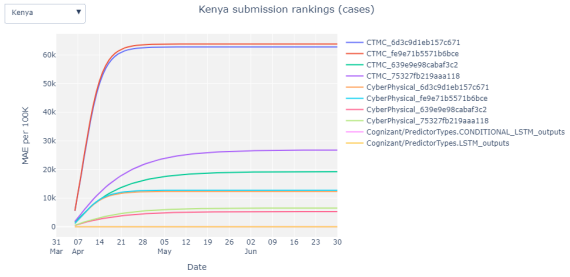
Note that the manually calibrated model performs better initially than the other Cyber-Physical models and the CTMC models. However, its accuracy deteriorates later in the time horizon for both Aotearoa | New Zealand – see Figure 5.4b – and for the other locations – this can be observed from the second peak in errors in Figure 5.4a that is not from the Aotearoa | New Zealand errors (which contributes to the first peak). Zooming in to the beginning of the modelling time horizon for Aotearoa | New Zealand, both the errors and ranking of the manually calibrated Cyber-Physical model are very good, but deteriorating quickly, for just over a week (from April 7-14) until it is no longer the best ranked model. Figure 5.4 shows the difficulty of calibrating models manually, hence the value of automated calibration. While the manual calibration by experts is superior for the initial period of modelling (which was likely their focus), by extending the model errors (and associated ranking) over a longer time horizon and calibrating the model to reduce those errors, the calibration framework has produced models are better predictors of disease spread in the long term, (although worse over the first week of the time horizon). Further manual calibration considering the entire time horizon may result in a better model, but that is both difficult due the the large number of input parameters and, hence, time intensive.



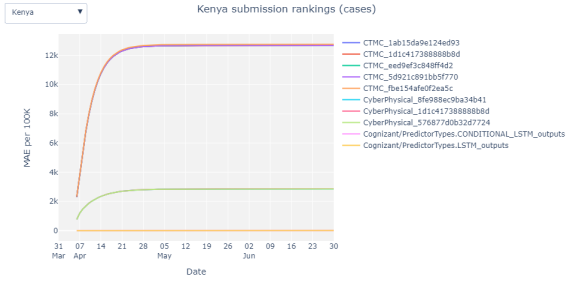
(a) Overall cases ranking – short



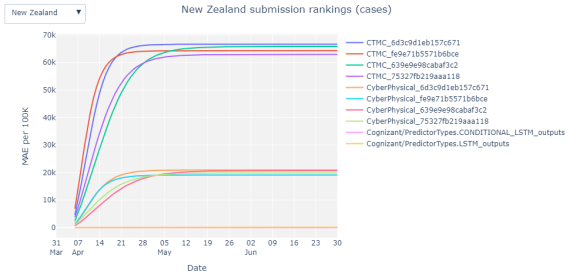
(b) Overall cases ranking – long



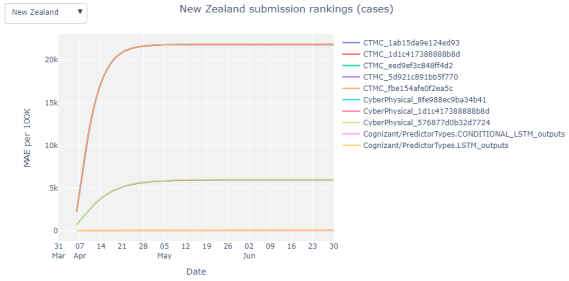
(c) Kenya cases ranking – short



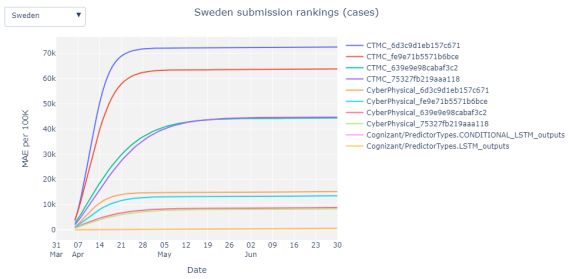
(d) Kenya cases ranking – long



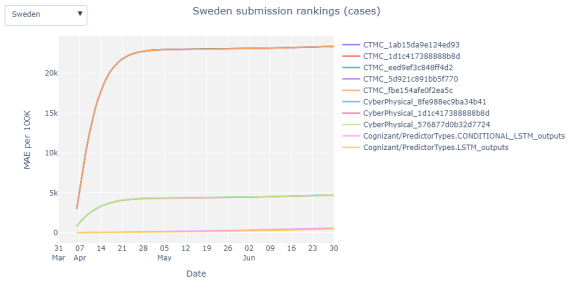
(e) Aotearoa | NZ cases ranking – short



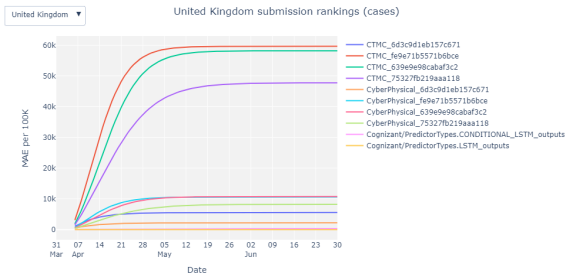
(f) Aotearoa | NZ cases ranking – long



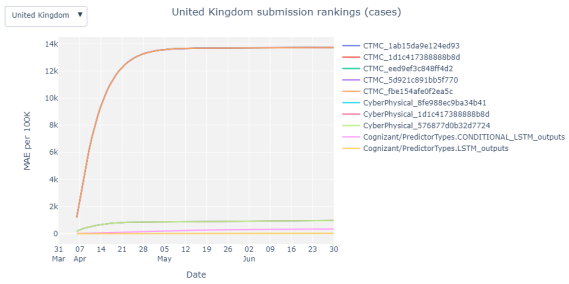
(g) Sweden cases ranking – short



(h) Sweden cases ranking – long

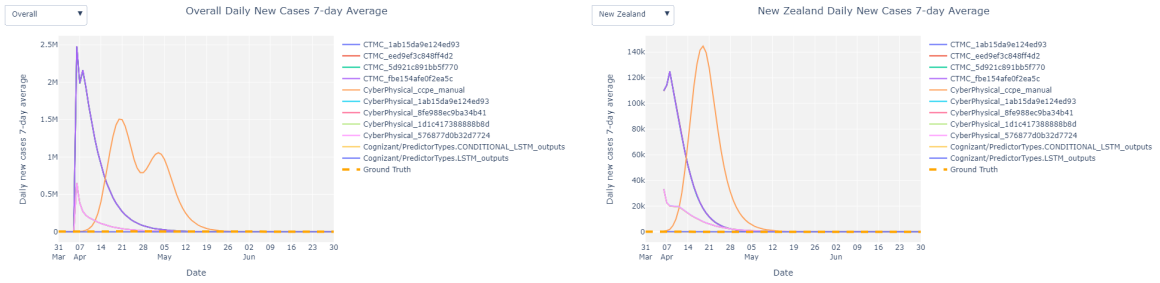


(i) United Kingdom cases ranking – short



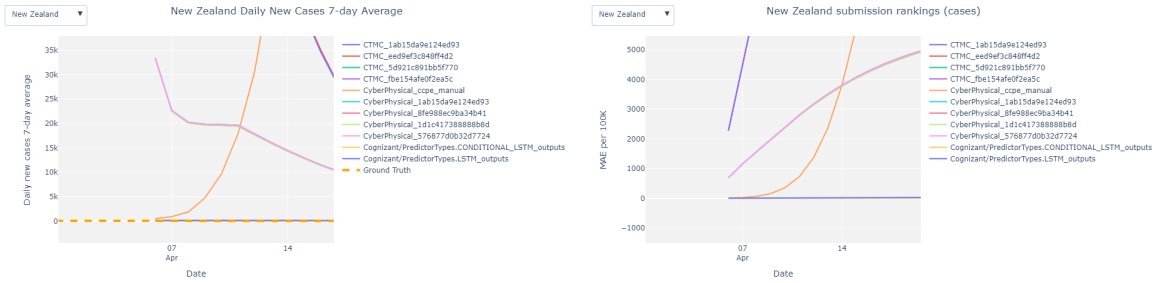
(j) United Kingdom cases ranking – long

Figure 5.3: Comparing cases ranking metric: short on the left, long on the right – same definitions as figure 5.1



(a) Overall new cases – long & manual calibration

(b) Aotearoa | NZ new cases – long & manual calibration



(c) Aotearoa | NZ new cases, early dates – long & manual calibration

(d) Aotearoa | NZ cases ranking, early dates – long & manual calibration

Figure 5.4: Comparing manual calibration of the Cyber-Physical model

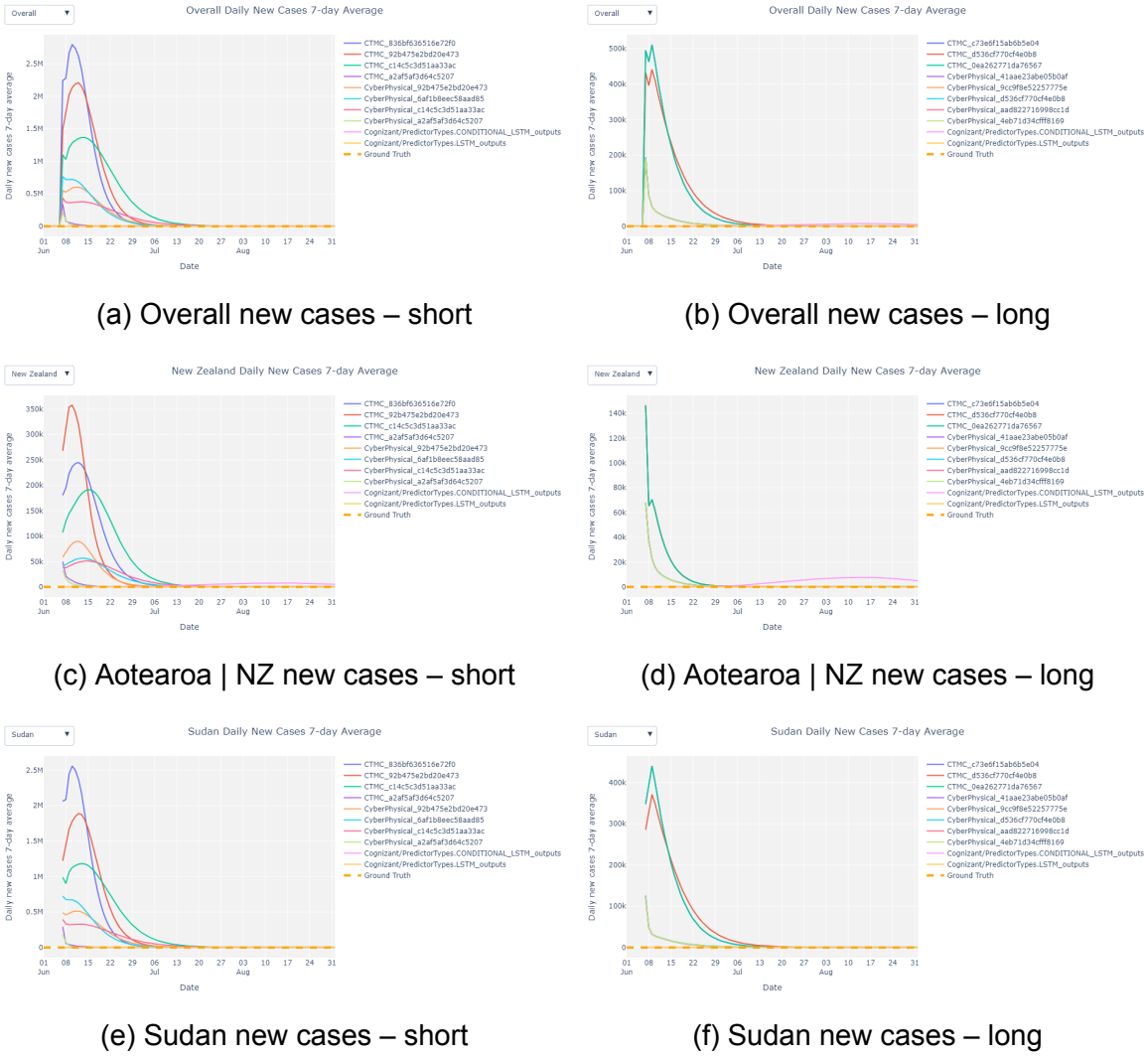
5.2 Case study 2: Aotearoa | New Zealand and Sudan

Similar to Case study 1 – see §5.1, figures 5.5 and5.6 compare the new cases metrics for the (Aotearoa | New Zealand, Sudan) case study between a preliminary, short calibration run (5 individuals, 2 offspring, 3 generations, 10% allowed change from initial values) and a more thorough, long calibration run (20 individuals, 15 offspring, 500 generations, 50% change allowed from initial values). Appendix A.2 contains plots of the equivalent metrics for new deaths in Figures A.3 and A.4. Note that – as in Case Study 1, §5.1 – the Cyber-Physical and CTMC algorithm labels have a unique code appended that links to the input parameters being used. Given new calibration runs for this case study, the codes are different from Case Study 1.

5.2.1 Comparison of case study results

One location, Aotearoa | New Zealand, was common to both case studies although the case studies had different modelling time horizons. In Figure 5.7 we compare the errors in predicting case numbers from both case studies and from short and long calibration runs.

In both case studies, despite the difference in time horizon, there are significant errors in case number prediction at the beginning of the time horizon. This error also drops significantly (by approximately 66% and 60% respectively) with more calibration time. This indicates that the errors in the models being calibrated are due to interactions between the models and



(a) Overall new cases – short

(b) Overall new cases – long

(c) Aotearoa | NZ new cases – short

(d) Aotearoa | NZ new cases – long

(e) Sudan new cases – short

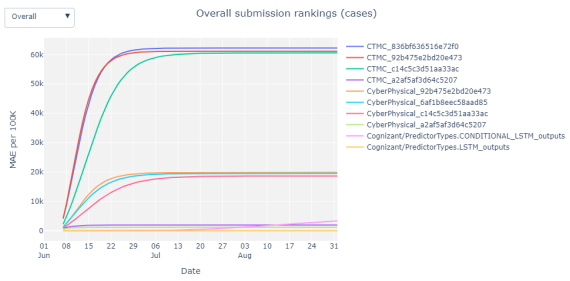
(f) Sudan new cases – long

Figure 5.5: Comparing new cases metric: – same definitions as figure 5.1

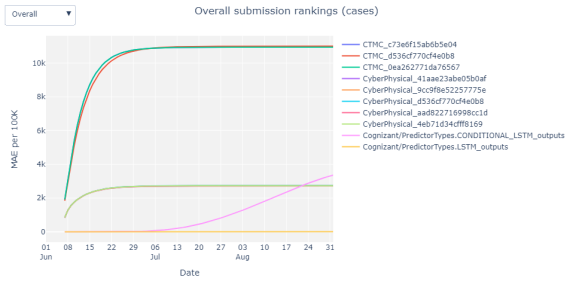
the input parameters, e.g., initial numbers of people in each disease state, rather than an inherent difficulty predicting case numbers at a particular point in time. For example, the errors in June are very low for case study 1 – see Figures 5.7a and 5.7b – but high for case study 2 – see Figures 5.7c and 5.7d, due to June being at the beginning of case study 2's time horizon. More research, such as refining the models within the calibration framework, is needed, but the results demonstrate that it is possible to get the models to give good predictions of case numbers - just not (yet) at the beginning of the time horizon.

5.3 Results from ensemble integration

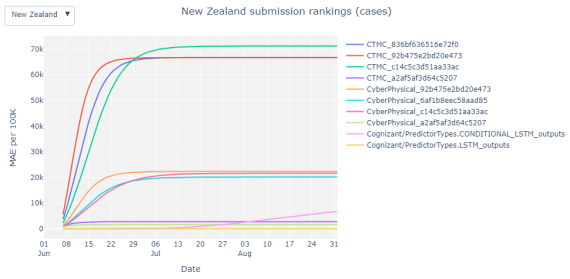
The ensemble creates an integrated prediction that is informed by each constituent model. The results indicate that the final prediction benefits from the contribution of each model while staying robust by ignoring contributions from errant models. The ability to pay attention to accurate predictions while ignoring systemic errors is a demonstration of the RIO-based uncertainty estimation introduced in this report. This section covers the capabilities of the RIO-based ensemble in (1) error correction and (2) uncertainty estimation.



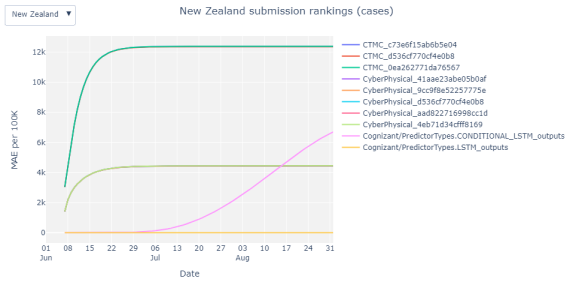
(a) Overall cases ranking – short



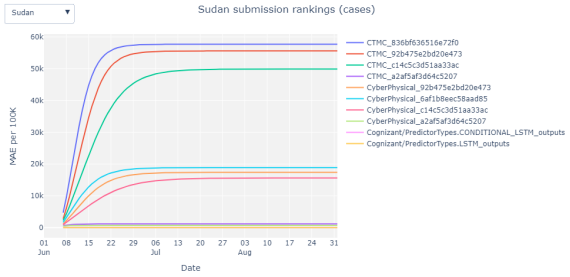
(b) Overall cases ranking – long



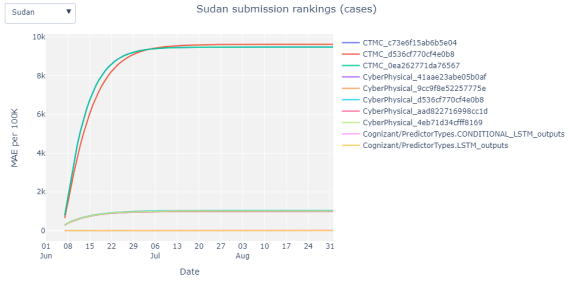
(c) Aotearoa | NZ cases ranking – short



(d) Aotearoa | NZ cases ranking – long



(e) Sudan cases ranking – short



(f) Sudan cases ranking – long

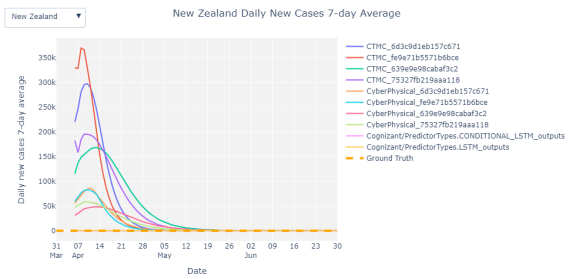
Figure 5.6: Comparing cases ranking metric: short on the left, long on the right – same definitions as figure 5.1

5.3.1 Error correction

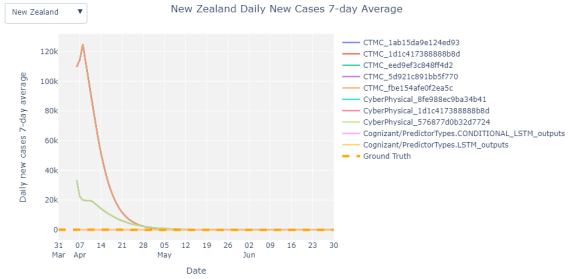
The ensemble system benefits from two sources of error correction, illustrated in the second and third columns, respectively, of Figure 5.8, which illustrates the results of case study 1 for Aotearoa | NZ:

1. The error correction models, implemented using RIO, provide a source of error correction based on previously seen training data. These models estimate the level of uncertainty and provide corrections based on the kinds of errors the model has made in the past on known data.

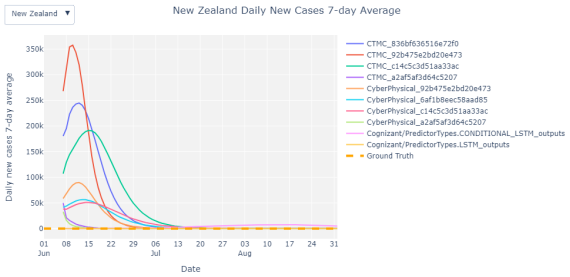
2. When the constituent models are integrated into the full ensemble, the biases of each model cancel out. The averaging out of various biases is the core idea behind ensembling and the reason our approach uses a variety of diverse modeling techniques.



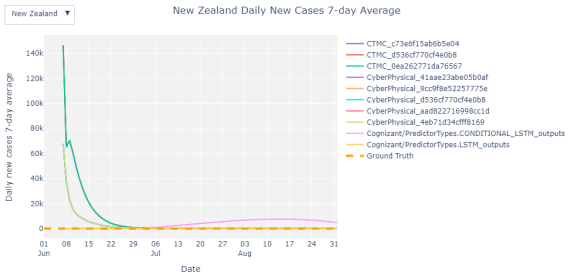
(a) Case study 1, Aotearoa | NZ new cases – short



(b) Case study 1, Aotearoa | NZ new cases – long



(c) Case study 2, Aotearoa | NZ new cases – short



(d) Case study 2, Aotearoa | NZ new cases – long

Figure 5.7: Comparing case studies, new cases metric for Aotearoa | New Zealand – same definitions as figure 5.1

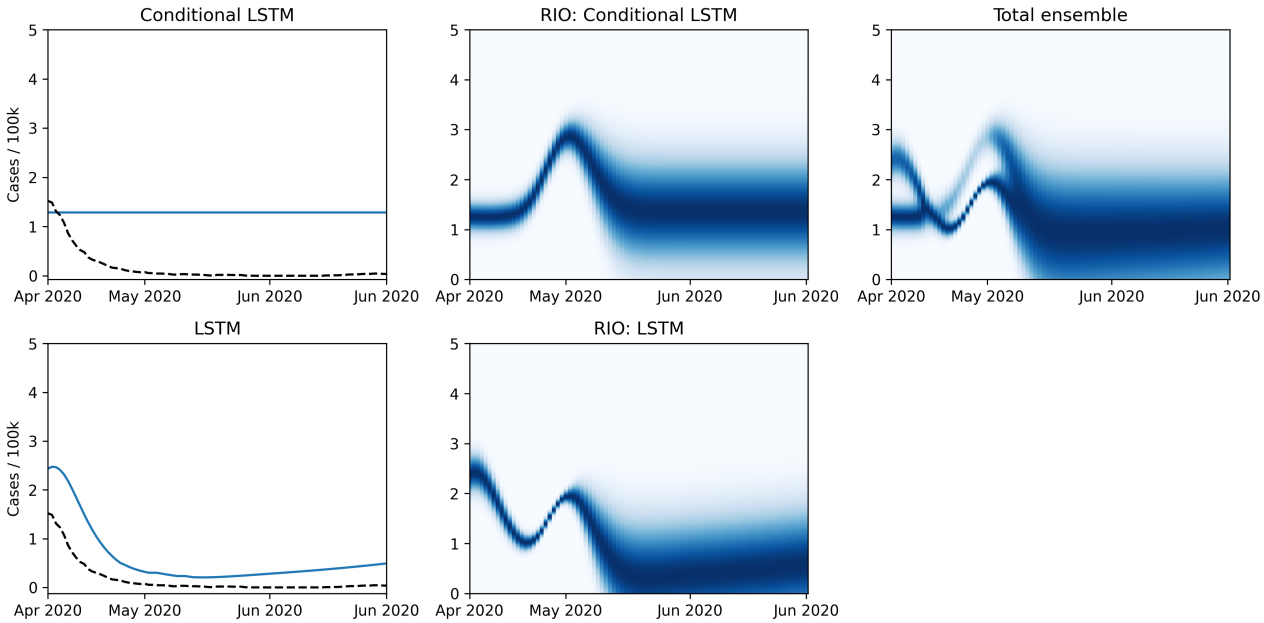


Figure 5.8: (Left Column) The predictions of the v1 LSTM and v2 Conditional LSTM are shown in blue, alongside the ground truth during this scenario shown with a black dashed line. (Middle Column) RIO has been applied to each model, producing a corrective adjustment along with an estimate of the uncertainty of the model. (Right Column) The two uncertainty distributions are integrated into a final prediction with uncertainty.

We evaluate the integrated ensemble prediction compared with the individual models alone (Figure 5.9). To do this, we compare the most likely prediction of the ensemble and computed

error as the average absolute difference between its prediction and the true number of cases per one hundred thousand population. The results indicate that the ensemble system is an improvement in the accuracy of the various inputs while maintaining robustness to member predictions that had high errors. In some cases, as in Sweden and the United Kingdom, the ensemble greatly outperforms any constituent model prediction by producing more accurate predictions.

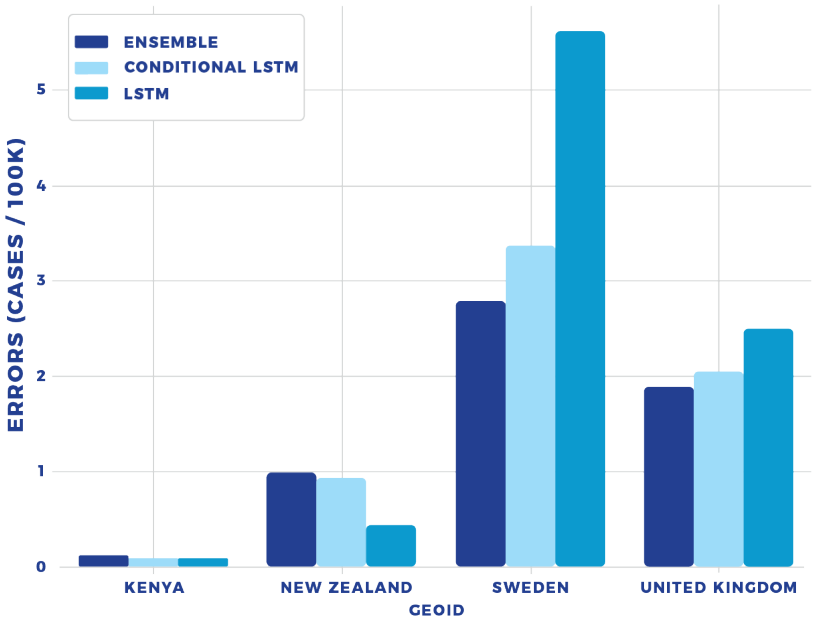


Figure 5.9: The prediction errors for the ensemble members are shown, along with the prediction error for the integrated full ensemble.

5.3.2 Uncertainty estimation

The ensemble serves to produce a final estimate that incorporates all the member predictions while being robust to errors. The RIO models capture the characteristic patterns of error of each model and thus output a high level of estimated uncertainty corresponding to regimes in which models perform poorly.

We observed that the ensemble system successfully filtered out noisy predictions so that they did not interfere with the final ensemble (Figure 5.10). The system correctly estimated a high level of uncertainty in contexts where models were more prone to errors. In the final results, the models assigned greater uncertainty had less impact on the final forecast.

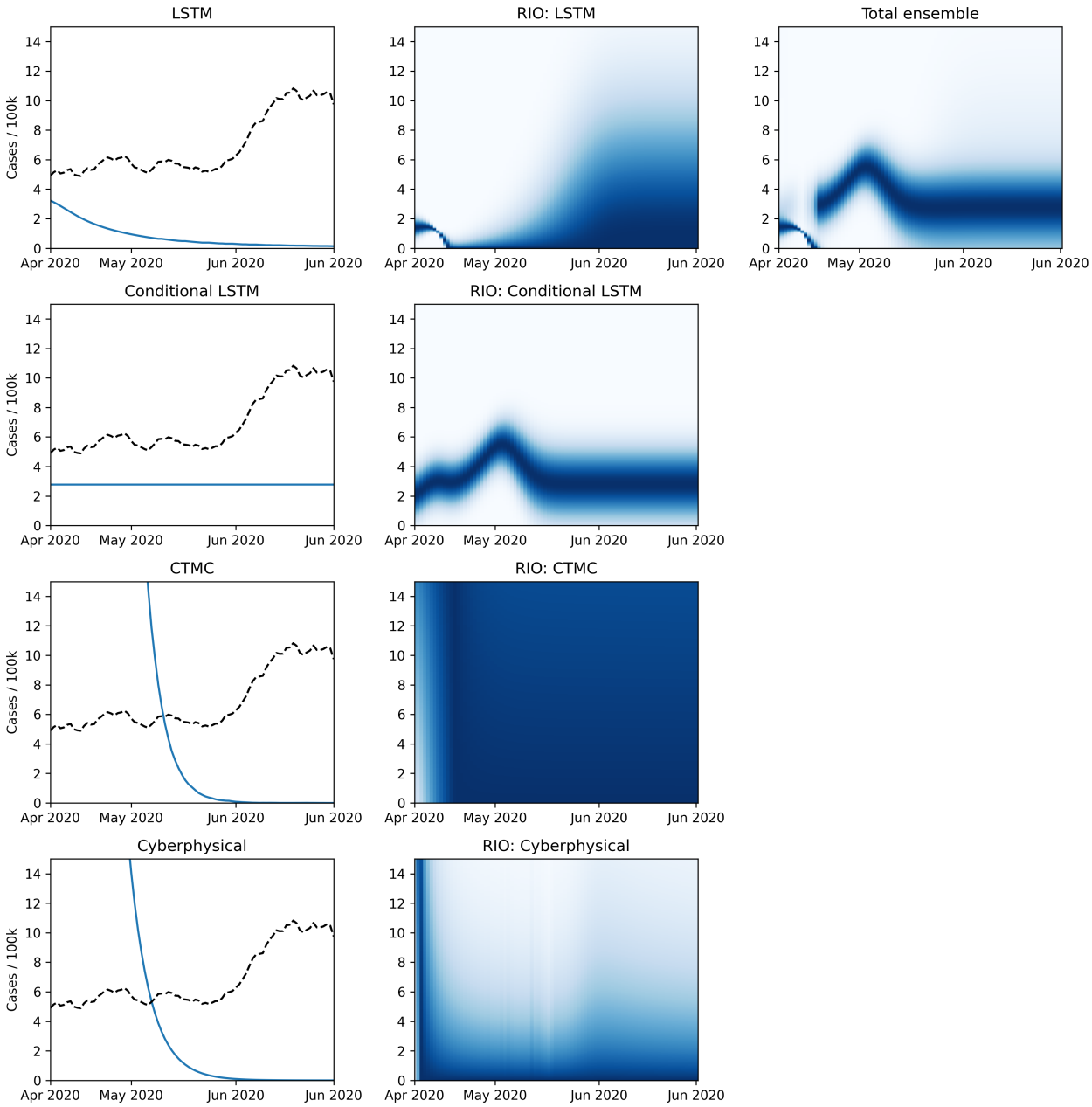


Figure 5.10: The results of the RIO-ensemble on Scenario 7 for Sweden demonstrate the uncertainty estimation capability of the ensemble integration. The first column indicates the predicted number of cases produced by each constituent model in the ensemble. The second column indicates the RIO-adjusted forecasts, where darker colors indicate the likelihood of the forecast. Note that the CTMC model is assigned a wide window of likelihood, corresponding to a high estimated uncertainty in the predictions. The third column indicates the integrated ensemble. Accordingly, the contributions from models with high uncertainty do not significantly impact the results.

6 Conclusion

This report is the second of two progress reports from the GPAI Pandemic Resilience project. The previous report (GPAI, 2023) presented recommendations on various approaches that

aligned pandemic modelling with responsible AI. To start the conclusion of this report we revisit several of those recommendations and discuss how the research described in this report is aligned with the previous recommendations. We then discuss the limitations of the research to date and look to the future in terms of how this research could be used and further research and development possibilities.

6.1 Contributions

This section discusses the contributions from the Pandemic Resilience project's research including standardisation, ensemble models and principles of responsible AI.

6.1.1 Standardisation is important

For models to be consistent, shared, compared and/or combined, there needs to be a standard interface, e.g., an API, for the inputs, data and outputs used by the models. We used the OxCGRT standard dataset to provide standardised input data as well as standardised data to compare our model outputs against. We defined standards for both inputs and outputs, as well as how the inputs are enabled to change during calibration. This standardisation meant that all 3 of the models considered could be used for any defined problem instance, e.g., the two case studies in §5.1 and §5.2 respectively. The models can be calibrated consistently and produce synchronised outputs for evaluation and combined into an ensemble – as demonstrated via the case studies. Although some assumptions were required to create the standardised API for the selected models, this standardised approach has meant that the Cyber-Physical and CTMC models can now be used for locations and time periods beyond their initial development. The standards also enable new models to be added to the calibration framework and ensemble model with relative ease.

6.1.2 Ensemble models provide robustness and perspective

One of the previous report's recommendations was the use of ensemble modelling due to its robustness and its ability to incorporate diverse perspectives by bringing models together. The preliminary research into ensemble models presented by GPAI (2023) has been extended in this report where a RIO-based ensemble method has been presented in §4. The RIO-based ensemble model not only combines multiple models when making predictions, but also adapts how the models are combined. This adaptive approach ensures that as model performance changes, their influence on the overall ensemble is adjusted to match.

The use of ensemble modelling in this research is key for both building trust in the predictions from the models as well as embedding diverse perspectives, via different models in the ensemble. The RIO-based ensemble provides both predictions and estimates of uncertainties in its predictions, so decision-makers also get a sense of the variability in potential outcomes and, hence, can include robustness in their planning, i.e., plan for uncertainty.

6.1.3 Enables informed, data-driven decision making

Policymakers require actionable insights and recommendations that are based on evidence. Faced with a constantly evolving complex world, models offer a tool to understand, discuss, participate and eventually find a workable agreement which takes account of all available information provided by experts and practitioners. The Pandemic Resilience project enables calibrated models to be easily created across multiple locations and time periods. It also streamlines the creation of models for new locations, given data for new locations is available. This means that decision-makers have access to high-quality models for their desired location and time period, which the assurance that the models are consistent across other locations and times. Hence, decision-makers use data and models to inform their decision-making with confidence in the validity of both the data and the models. These models can also be used to evaluate the effect of decisions in terms of disease spread and mortality and, in future iterations of this research, the models can also be used to recommend decisions to achieve specified aims such as, for example, minimising the number of deaths.

Although the models produced by the Pandemic Resilience project can be prescriptive, this research intends to keep human values at the core of decision-making, but support that decision-making with sophisticated modelling and comprehensive data. By making both the modelling and data more accessible, even to those who may not have the requisite resources and/or skills to develop them independently, the Pandemic Resilience project aims to empower decision-makers to use data-driven approaches to inform their decision-making. In addition, new data and/or models may be reasonably easily added to the framework, so diversity of geography and perspective is included in the Pandemic Resilience project.

6.1.4 Preparedness

Over the last two decades, there have been numerous outbreaks of viral infections, including Chikungunya, Ebola, Zika, Nipah, H7N9 avian flu, H1N1, SARS, MERS, and COVID-19. While COVID-19 continues, an exceptional number of monkeypox (a.k.a. mpox) cases have recently been documented in non-endemic areas. The calibration framework and ensemble modelling presented in this report have been developed to be part of a pandemic preparedness suite of tools. In addition, the approach taken in the development of the calibration framework, e.g., standardisation, using an ensemble of models, was deliberately chosen so the developed tools could be useful for future pandemics. For example, although there are currently no specific treatments for monkeypox, several antiviral medications developed to treat smallpox (Tecovirimat, Cidofovir, Brincidofovir, and Vaccinia Immune Globulin Intravenous – VIGIV) are being used to treat monkeypox disease. Combining approaches, including AI, for diagnosing, tracking, and monitoring monkeypox infections can produce high-quality standardised data (cf. OxCGRT) which can then be combined with either more general ML models (e.g., the LSTM model) or disease-specific parameterised models (e.g., the CTMC model) within an ensemble for evaluating monkeypox metrics (e.g., spread, mortality). The ensemble model can then be calibrated using a customisation of the framework

described in this report to provide accessible, calibrated monkeypox predictions – based on any NPIs that are being used – for multiple locations.

6.2 Limitations of the calibration framework and ensemble models

One clear limitation of the research in this report is the efficacy of the pandemic spread models used within the ensemble. While the LSTM models are generic and can be used with any pandemic data, they require a reasonable volume of data to achieve good accuracy. The Cyber-Physical and CTMC models were developed specifically for COVID-19 spread and also come from the start of the pandemic, so don't (yet) have reinfection as part of the models. However, the ability of the calibration model to be quickly customised to new models and new diseases mitigates the limitations of the underlying models to some extent.

The reliance of both the Cyber-Physical and CTMC models on good input parameters is clear from both the manual calibration results – see §5.1.1 – and the artefact of their low accuracy at the start of the modelling horizon, regardless of the actual data values at that time – see §5.2.1. The automated calibration is improving this accuracy, but it is still significant after a reasonably long calibration. Further research involving longer calibration runs is needed to observe the trade-off between calibration time and accuracy. In addition, some deeper analysis of how the performance of the Cyber-Physical and CTMC models is affected by their input parameters may help provide more accurate estimates.

Similar to the calibration framework, the ensemble model is limited by the performance of the models that underpin the ensemble. However, the RIO-based approach also adapts the ensemble so that the best-performing models inform predictions. Hence, shortcomings of individual model/input configurations can be overcome within the ensemble, as much as possible within the RIO-based approach.

6.3 Use cases

Given the successful prototyping of the calibration framework and ensemble modelling described in this report, we next turn our attention to how these digital tools might be used in practice by decision-makers. In this section, we summarise three use cases – originally specified in the previous Pandemic Resilience report (GPAI, 2023) – and describe how the calibration framework/ensemble model could support those use cases.

6.3.1 Forecasting

Forecasting requires an ability to accurately forecast the future. When managing pandemics the focus is on short-term to medium-term forecasting given decisions on, e.g., the use of NPIs. Forecasting with ensemble models enables the robustness of forecasts and the inclu-

sion of levels of uncertainty to support decision-making. Incorporating a forecasting ensemble model in public health policy assists decision-making processes for resource allocation and response planning.

The use case demonstrated in the case studies – see §5 – is forecasting. Parameters for the models, including initial numbers of people in various disease stages, are estimated and the calibration framework uses all available data to improve the forecasting from the ensemble model. Automated adjustment of (uncertain) parameter estimates results in an improved forecasting model for supporting decision-makers.

6.3.2 Scenarios

Longer forecast horizons result in greater uncertainty and larger forecast errors. The use of scenarios can help to guide a participatory conversation about what is most likely to happen in the future. Combining scenarios with ensemble modelling can help to manage risk and avoid disastrous consequences. Mitigating against forecast error within ensemble models that use scenarios involves incorporating scenario-based analysis, using model averaging or weighted averaging, applying data assimilation techniques, and employing post-processing techniques (Kuhl et al., 2007; Cawood and Zyl, 2021; Rayner and Bolhuis, 2020). The use of a variety of scenarios alongside ensemble modelling when planning, e.g., a pandemic response, adds robustness to the decision-makers' plans.

As discussed in §6.3.1, accurate forecasts can be determined using the calibration framework and embedded ensemble model. However, scenarios are also supported as input parameters – including initial numbers of people in various disease stages and/or the NPI schedule – can be specified and fixed, i.e., the calibration framework will not automatically adjust them. Then, the calibration framework will determine the most accurate ensemble model *given the fixed input parameters*, so the effect of fixing particular parameters to specific values on forecast accuracy can be observed. In addition, calibrated input parameters can be blended with input parameters for given scenarios so scenario outcomes can be measured with as little "noise" from uncertain parameters as possible, i.e., the calibration framework can provide valuable inputs to scenario analysis. Further research and model Development Operations (DevOps) are required to streamline scenario analysis alongside the calibration framework presented here.

6.3.3 Policy laboratories

A policy laboratory (a.k.a. policy lab) refers to a space or platform or controlled environment where interdisciplinary teams, including policymakers, researchers, and stakeholders, come together to test and evaluate different policy interventions, strategies, and approaches before large-scale implementation (Lunn and Choisdealbha, 2018). Policy labs are used to identify risks, vulnerabilities, and unintended consequences, and evaluate the effectiveness and feasibility of different policy measures(Saam and Kerber, 2008). As an example of the utility of integrating diverse datasets, Agyapon-Ntra and McSharry (2022) combined OxCGRT with

Google mobility data to quantify compliance at the country level and evaluate the efficacy of different policies in reducing cases. The platform offers a means of safely exploring different policies and better understanding the costs and benefits (economic and health).

The calibration framework is designed for use in a policy laboratory environment. Decision makers can set up experiments, e.g., locations and time horizons of interest along with estimates of input parameters, and automatically calibrate an ensemble model to get accurate forecasts of pandemic spread. Future research and model DevOps will enable the resulting ensemble model to be available for experimentation in terms of the input parameters, e.g., estimated numbers of presymptomatic people in the population, the NPIs being used. Hence, exploration of how pandemic responses relate to a variety of well-calibrated modelled outcomes is possible.

6.4 Future work

6.4.1 Ensuring robust forecasting via diversity

After the COVID-19 pandemic and the current monkeypox (a.k.a. mpox) epidemic, the world is becoming more conscious of the need for pandemic preparation. While there was a burst of modelling activity during the COVID-19 pandemic, there is still a scarcity of modelling environments that are suitable for guiding pandemic response. This work developed a research prototype that uses AI to calibrate and deploy multiple pandemic spread models in an ensemble over numerous geographies. While these models are COVID-specific, the AI calibration framework is disease-independent, and the ensemble model method allows more models to be added as needed.

This research started to explore the use of multiple models in an ensemble as well as consistent parameter values across different locations so that the model outcomes are diverse both in their perspective (i.e., combining different models) and geography (i.e., using data from different locations). This diversity promotes robustness both in the way the RIO-based ensemble estimates uncertainty and how the calibration framework ensures good model performance for multiple locations.

The next level of Pandemic Resilience could expand the AI-calibration study prototype to include Pandemic Preparedness and further models. This can automate the usage of the calibration and models to support decision-making. The next phase of the Pandemic Resilience project could look to add other models (ensuring diversity of perspective) and/or consider different ensemble methods (ensuring diverse perspectives are appropriately valued). More testing of the calibration framework is also needed to make sure models are fit-for-purpose across multiple locations. Finally, embedding the calibration framework within a pandemic preparedness approach would ensure robust forecasting informs good decision-making – see §6.4.4 for more.

6.4.2 Extension to other diseases

As mentioned in §6.4.1, the calibration framework was developed using COVID-specific models, but the calibration framework is model agnostic. Hence, the framework could be used to support responses to various diseases with similar methods but different data and models. For example, given available data for influenza, respiratory syncytial virus (RSV), monkeypox (a.k.a. mpox), and adenovirus, predictive models may be rapidly calibrated across multiple locations. It is conceivable that the calibration framework could be part of a platform with a library of disease models and used as part of a comprehensive continuous monitoring and intervention strategy for public health officials.

6.4.3 Adding economic performance

Decision-making usually involves considering the costs and benefits of different actions. The pandemic presented policymakers with substantial challenges in that neither the adverse economic impacts nor the benefits in terms of protecting society were readily available. To make matters worse, different variants of the virus presented a range of mortality rates and the efficacy of the available vaccines were not fully understood or quantified initially.

The calibration framework can easily include new models and ensure that their inputs are consistent with all other models. Hence, an economic model could be added and it would use the same NPIs as the COVID-19 spread models. After calibration the ensemble model could then be used to explore how to use NPIs to find the best balance between public health and economic outcomes.

6.4.4 Policy recommendations

Currently, the calibration framework aims to provide robust forecasting, scenario analysis and policy lab functionality. However, since NPIs are part of the input parameters, NPI scenarios can be optimized based on several objectives such as the number of cases and economic impact. Prescriptive models can then help decision-makers make informed choices by suggesting NPI timelines based on desired tradeoffs between objectives. Decision makers can also limit the prescriptive model's search to certain scenarios, such as schools closing for a maximum of two weeks, "recommendation for full lockdown", and so on. Additionally, decision-makers can predict the effectiveness of proposed improvements.

Effective use of the technology presented in this report requires close coordination between governments, decision-makers, and computational experts. Although the prospect of prescriptive models is appealing, decision-makers must maintain accountability for model-based decisions. Responsible AI requires integrating ethics, transparency, and responsibility into decision support systems.

As we've seen with the COVID-19 pandemic and climate change, the next global disaster might be a complex web of interconnected problems. Models will become more important

in guiding decision-makers through these complex difficulties. Human-centred AI and modelling are essential for responsible and successful decision support systems. The use of such technology will have impactful implications for understanding and responding to future pandemics or global emergencies. Governments, with their ability to make consequential decisions, should look for effective ways to embed this technology within policy recommendations.

6.4.5 Scientific publication

In addition to future work that would enable the calibration framework and ensemble models to be part of pandemic preparedness initiatives, the next phase of the Pandemic Resilience project will also involve more comprehensive testing and analysis for publication in academic journals. The outcomes of the research presented in this report are of interest not only to public health and government more widely, but for researchers in modelling and AI. Transforming and extending the work in this report and the previous Pandemic resilience report will provide material for one or more transdisciplinary journal articles on modelling and AI for pandemic response and/or public health policy.

6.5 Final remarks – Responsible AI

The Responsible AI Working Group within GPAI describes responsible AI as “human-centred, fair, equitable, inclusive and respectful of human rights and democracy, and that aims at contributing positively to the public good” (GPAI, 2024). The research in this report was designed to align with these responsible AI principles. This research ultimately aims to contribute to the public good by developing a suite of digital tools to enhance pandemic preparedness and assist decision makers to best utilise NPIs and keep people safe from disease spread.

The ensemble approach is inherently inclusive as different models can be easily added to the calibration, so multiple perspectives from diverse groups can be considered and these perspectives can be used to inform decision making and support robustness. Safeguarding inclusion and accessibility means being mindful of the composition of contributors to the ensemble model. This approach that incorporates diverse perspectives into modelling is also a natural way to enhance key characteristics of responsible AI such as interoperability, explainability and robustness.

The ensemble approach is equitable in that the calibration framework and standardised approach enable people from any location to rapidly create models that can accurately predict the effect of NPIs on disease spreading in their location. Moreover, the standardization of the ensemble modelling approach aims to provide easy access to models and data for those jurisdictions that do not currently have the resources to develop such models themselves. This equity of access aims to provide digital modelling to those who may not otherwise be able to access it, hence the research in this report aims to, in a small way, enhance the democratisation of digital technology and AI thus reducing the digital divide between developed and

developing nations.

However, it is crucial to be mindful of the fact that the ensemble model developed is intended to support a human decision maker, i.e., it evaluates decisions on NPIs and, in future iterations, could suggest NPIs schedules. Thus far, the Pandemic Resilience project outputs are limited to prediction and so the decision making and hence, accountability should remain with the human decision makers using the Pandemic Resilience ensemble model. It is still essential to provide AI and modeling tools as decision making support rather than direct decision making systems for maintaining human-centred values in decision making. Such decision support systems, in the hands of public health communities who understand how they work and how to use them responsibly, has the potential to significantly enhance their decision making relative to pandemic preparedness and response.

Keeping the decision makers responsible and accountable for the decisions they take does not mean that modelers and developers are free from responsibility and accountability. As it is true for humans, the many existing and potential forms of models and artificial intelligences are also subject to biases and errors, among other risks. Thus, the project advocates for transparency and explainability of the models algorithms and the types of data they use so the public and expert communities can interact and question the decisions and how they are being made.

Questioning the data used and the way it is being collected, stored and shared is also crucial. Governments could play a central role in making the required data available for pandemic modeling, and in ensuring data rights are being protected. While data sanitization and privacy enhancing technologies are promising avenues to protect the privacy rights of individuals and communities, thoughtful reflections and actions need to go into ensuring responsible data governance for such systems.

Some argue that Responsible AI can be understood as a destination, an ideal to keep aiming for. The hard collaborative work done the pandemic Resilience Project demonstrates sustained honest efforts in that direction.

References

Abdulaal, Ahmed et al. (2020). “Prognostic modeling of COVID-19 using artificial intelligence in the United Kingdom: model development and validation”. In: *Journal of Medical Internet Research* 22.8, e20259.

Aguas, Ricardo et al. (2020). “Modelling the COVID-19 pandemic in context: an international participatory approach”. In: *BMJ global health* 5.12, e003126.

Agyapon-Ntra, Kwadwo and Patrick E McSharry (2022). “A global analysis of the effectiveness of policy responses to COVID-19”. In: *Scientific Reports* 13, p. 5629.

Alur, Rajeev (2015). *Principles of cyber-physical systems*. MIT press.

Annan, James D and Julia C Hargreaves (2020). “Model calibration, nowcasting, and operational prediction of the COVID-19 pandemic”. In: *medrxiv*, pp. 2020–04.

Banholzer, Nicolas et al. (2021). “Estimating the effects of non-pharmaceutical interventions on the number of new infections with COVID-19 during the first epidemic wave”. In: *PLoS one* 16.6, e0252827.

Bertozzi, Andrea L. et al. (2020). “The challenges of modeling and forecasting the spread of COVID-19”. In: *Proceedings of the National Academy of Sciences* 117.29, pp. 16732–16738. DOI: 10.1073/pnas.2006520117. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2006520117. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2006520117.

Beyerer, Jürgen, Christian Kühnert, and Oliver Niggemann (2019). *Machine Learning for Cyber Physical Systems: Selected Papers from the International Conference ML4CPS 2018*. Springer Nature.

Blank, J. and K. Deb (2020). “pymoo: Multi-Objective Optimization in Python”. In: *IEEE Access* 8, pp. 89497–89509.

Brauner, Jan M et al. (2021). “Inferring the effectiveness of government interventions against COVID-19”. In: *Science* 371.6531, eabd9338.

Cawood, Pieter and Terence L van Zyl (2021). “Feature-weighted stacking for nonseasonal time series forecasts: A case study of the covid-19 epidemic curves”. In: *2021 8th International Conference on Soft Computing & Machine Intelligence (ISCMI)*. IEEE, pp. 53–59.

Centers for Disease Control and Prevention (CDC) (2022). *Nonpharmaceutical Interventions (NPIs)*. https://www.cdc.gov/nonpharmaceutical-interventions/index.html [Last Reviewed: 3rd October 2022, Accessed: 31st October 2023].

Chen, Mu-Fa and Yong-Hua Mao (2021). *Introduction to stochastic processes*. Vol. 2. World Scientific.

Chowdhury, Rajiv et al. (2020). “Dynamic interventions to control COVID-19 pandemic: a multivariate prediction modelling study comparing 16 worldwide countries”. In: *European journal of epidemiology* 35, pp. 389–399.

Cognizant (2021). *AI for good: XPRIZE Pandemic Response Challenge winners*. https://www.cognizant.com/us/en/services/ai/pandemic-response. [Online; accessed 16-May-2024].

Cooper, Ian, Argha Mondal, and Chris G Antonopoulos (2020). "A SIR model assumption for the spread of COVID-19 in different communities". In: *Chaos, Solitons & Fractals* 139, p. 110057.

Cosgriff, Christopher V, Daniel K Ebner, and Leo Anthony Celi (2020). "Data sharing in the era of COVID-19". In: *The Lancet Digital Health* 2.5, e224.

Cramer, Estee Y et al. (2022). "Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States". In: *Proceedings of the National Academy of Sciences* 119.15, e2113561119.

Cristianini, Nello and Elisa Ricci (2008). "Support Vector Machines". In: *Encyclopedia of Algorithms*. Ed. by Ming-Yang Kao. Boston, MA: Springer US, pp. 928–932. ISBN: 978-0-387-30162-4. DOI: 10.1007/978-0-387-30162-4_415. URL: https://doi.org/10.1007/978-0-387-30162-4_415.

Deb, Kalyanmoy et al. (2002). "A fast and elitist multiobjective genetic algorithm: NSGA-II". In: *IEEE transactions on evolutionary computation* 6.2, pp. 182–197.

Ebert, Christof et al. (2016). "DevOps". In: *IEEE software* 33.3, pp. 94–100.

Elsheikh, S et al. (2020). "A mathematical model for the transmission of corona virus disease (COVID-19) in Sudan". In: *Preprint*.

Fonseca, Carlos M and Peter J Fleming (1993). "Multiobjective genetic algorithms". In: *IEE colloquium on genetic algorithms for control systems engineering*. Iet, pp. 6–1.

Gallo Marin, Benjamin et al. (2021). "Predictors of COVID-19 severity: a literature review". In: *Reviews in medical virology* 31.1, pp. 1–10.

Gao, Frank et al. (2020). "Management and data sharing of COVID-19 pandemic information". In: *Biopreservation and biobanking* 18.6, pp. 570–580.

Gawlikowski, Jakob et al. (Oct. 1, 2023). "A survey of uncertainty in deep neural networks". In: *Artificial Intelligence Review* 56.1, pp. 1513–1589. ISSN: 1573-7462. DOI: 10.1007/s10462-023-10562-9. URL: https://doi.org/10.1007/s10462-023-10562-9 (visited on 07/18/2024).

Giabbanelli, Philippe J et al. (2021). "Opportunities and challenges in developing covid-19 simulation models: Lessons from six funded projects". In: *2021 Annual Modeling and Simulation Conference (ANNSIM)*. IEEE, pp. 1–12.

GPAI (2021). *Responsible AI for Social Media Governance: A proposed collaborative method for studying the effects of social media recommender systems on users*. Global Partnership on Artificial Intelligence report. Authors: Knott, A., Hannah, K., Pedreschi, D., Chakraborti, T., Hattotuwa, S., Trotman, A., Baeza-Yates, R., Roy, R., Eyers, D., Morini, V. and Pansanella, V.

— (2023). *Pandemic Resilience: Developing an AI-calibrated Ensemble of Models to Inform Decision Making*. GlobalPartnership on ArtificialIntelligence report. Authors: Michael O'Sullivan, M., Hupert, N., McSharry, P., Miikkulainen, R., Francon, O., Quenneville-Langis, A., Warner, J., Agyepong, V., Allen, N., Chatterjee, S., Bahrami, S., and Farid, A.

GPAI (2024). *Responsible AI*. https://gpai.ai/projects/responsible-ai/. Accessed: 2nd October 2024.

Graves, Alex (2012). "Long Short-Term Memory". In: *Supervised Sequence Labelling with Recurrent Neural Networks*. Vol. 385. Series Title: Studies in Computational Intelligence. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 37–45. DOI: 10.1007/978-3-642-24797-2_4. URL: http://link.springer.com/10.1007/978-3-642-24797-2_4 (visited on 09/15/2023).

Hale, Thomas et al. (2021). "A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker)". In: *Nature human behaviour* 5.4, pp. 529–538.

Hayman, David T.S. et al. (2023). "Developing One Health surveillance systems". In: *One Health* 17, p. 100617. ISSN: 2352-7714. DOI: https://doi.org/10.1016/j.onehlt.2023.100617. URL: https://www.sciencedirect.com/science/article/pii/S2352771423001374.

Hazelbag, C Marijn et al. (2020). "Calibration of individual-based models to epidemiological data: A systematic review". In: *PLoS computational biology* 16.5, e1007893.

Hendy, Shaun et al. (2021). "Mathematical modelling to inform New Zealand's COVID-19 response". In: *Journal of the Royal Society of New Zealand* 51.sup1, S86–S106. DOI: 10.1080/03036758.2021.1876111. eprint: https://doi.org/10.1080/03036758.2021.1876111. URL: https://doi.org/10.1080/03036758.2021.1876111.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory". In: *Neural computation* 9.8, pp. 1735–1780.

Holland, John H (1992). "Genetic algorithms". In: *Scientific american* 267.1, pp. 66–73.

Iboi, Enahoro et al. (2020). "Mathematical modeling and analysis of COVID-19 pandemic in Nigeria". In: *MedRxiv*, pp. 2020–05.

Ibrahim, Zurki, Pinar Tulay, and Jazuli Abdullahi (2023). "Multi-region machine learning-based novel ensemble approaches for predicting COVID-19 pandemic in Africa". In: *Environmental Science and Pollution Research* 30.2, pp. 3621–3643.

Jin, Weiqiu et al. (2022). "A data-driven hybrid ensemble AI model for COVID-19 infection forecast using multiple neural networks and reinforced learning". In: *Computers in Biology and Medicine* 146, p. 105560.

Johnson, Alistair, Tom Pollard, and Mark Roger (Sept. 2016). *MIMIC-III Clinical Database*. Version v1.4. PhysioNet. DOI: 10.13026/C2XW26. URL: https://doi.org/10.13026/C2XW26.

Konak, Abdullah, David W Coit, and Alice E Smith (2006). "Multi-objective optimization using genetic algorithms: A tutorial". In: *Reliability engineering & system safety* 91.9, pp. 992–1007.

Kong, Chung Yin, Pamela M McMahon, and G Scott Gazelle (2009). "Calibration of disease simulation model using an engineering approach". In: *Value in health* 12.4, pp. 521–529.

Kuhl, David D. et al. (2007). "Assessing Predictability With a Local Ensemble Kalman Filter". In: *Journal of the Atmospheric Sciences*. DOI: 10.1175/jas3885.1.

Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell (Nov. 3, 2017). *Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles*. DOI: 10.48550/

arXiv.1612.01474. arXiv: 1612.01474[cs,stat]. URL: http://arxiv.org/abs/1612.01474 (visited on 07/18/2024).

Lee, Edward A. (2008). “Cyber Physical Systems: Design Challenges”. In: *2008 11th IEEE International Symposium on Object and Component-Oriented Real-Time Distributed Computing (ISORC)*, pp. 363–369. DOI: 10.1109/ISORC.2008.25.

Liu, Meng, Raphael Thomadsen, and Song Yao (2020). “Forecasting the spread of COVID-19 under different reopening strategies”. In: *Scientific reports* 10.1, p. 20367.

Lunn, Pete and Áine Ní Choisdealbha (2018). “The Case for Laboratory Experiments in Behavioural Public Policy”. In: *Behavioural Public Policy*. DOI: 10.1017/bpp.2016.6.

Maaliw, Renato R et al. (2021). “An ensemble machine learning approach for time series forecasting of COVID-19 cases”. In: *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE, pp. 0633–0640.

Mitchell, Melanie (1996). *An introduction to genetic algorithms*. eng. Cambridge, Mass: MIT Press. ISBN: 9780585030944.

Mohammed, Ammar and Rania Kora (Feb. 1, 2023). “A comprehensive review on ensemble deep learning: Opportunities and challenges”. In: *Journal of King Saud University - Computer and Information Sciences* 35.2, pp. 757–774. ISSN: 1319-1578. DOI: 10.1016/j.jksuci.2023.01.014. URL: https://www.sciencedirect.com/science/article/pii/S1319157823000228 (visited on 07/18/2024).

Nastasi, Giovanni et al. (2022). “A time-delayed deterministic model for the spread of COVID-19 with calibration on a real dataset”. In: *Mathematics* 10.4, p. 661.

Oh, Phil Seok and Sung Jin Oh (2011). “What teachers of science need to know about models: An overview”. In: *International Journal of Science Education* 33.8, pp. 1109–1130.

Organisation for Economic Co-operation and Development (OECD) (2019). *Recommendation of the Council on Artificial Intelligence*. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 [Adopted on: 22nd May 05 2019, Accessed: 1st November 2023].

Paireau, Juliette et al. (2022). “An ensemble model based on early predictors to forecast COVID-19 health care demand in France”. In: *Proceedings of the National Academy of Sciences* 119.18, e2103302119.

Qiu, Xin, Elliot Meyerson, and Risto Miikkulainen (June 2020). *Quantifying Point-Prediction Uncertainty in Neural Networks via Residual Estimation with an I/O Kernel*. arXiv:1906.00588 [cs, stat]. URL: http://arxiv.org/abs/1906.00588 (visited on 04/24/2023).

Rahman, Mohammad Marufur et al. (2021). “Machine learning approaches for tackling novel coronavirus (COVID-19) pandemic”. In: *SN computer Science* 2, pp. 1–10.

Rayner, Brett and Marijn Bolhuis (2020). “Deus Ex Machina? A Framework for Macro Forecasting With Machine Learning”. In: *Imf Working Paper*. DOI: 10.5089/9781513531724.001.

Ro, Jin Woo et al. (2020). “Compositional cyber-physical epidemiology of COVID-19”. In: *Scientific Reports* 10.1, p. 19537.

Saam, Nicole J. and Wolfgang Kerber (2008). “Policy Innovation, Decentralised Experimentation, and Laboratory Federalism”. In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.1301502.

Sarkar, Kankan, Subhas Khajanchi, and Juan J Nieto (2020). “Modeling and forecasting the COVID-19 pandemic in India”. In: *Chaos, Solitons & Fractals* 139, p. 110049.

Sass, Julian et al. (2020). “The German Corona Consensus Dataset (GECCO): a standardized dataset for COVID-19 research in university medicine and beyond”. In: *BMC Medical Informatics and Decision Making* 20, pp. 1–7.

Shastri, Sourabh et al. (2021). “Deep-LSTM ensemble framework to forecast Covid-19: an insight to the global pandemic”. In: *International Journal of Information Technology* 13.4, pp. 1291–1301.

Sherratt, Katharine et al. (2023). “Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations”. In: *Elife* 12, e81916.

Tang, Yuanji and Shixia Wang (2020). “Mathematic modeling of COVID-19 in the United States”. In: *Emerging microbes & infections* 9.1, pp. 827–829.

Tayarani-Najaran, Mohammad-Hassan (2022). “A novel ensemble machine learning and an evolutionary algorithm in modeling the COVID-19 epidemic and optimizing government policies”. In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 52.10, pp. 6362–6372.

Wang, QJ (1997). “Using genetic algorithms to optimise model parameters”. In: *Environmental Modelling & Software* 12.1, pp. 27–34.

Wangari, Isaac Mwangi et al. (2021). “Mathematical modelling of COVID-19 transmission in Kenya: a model with reinfection transmission mechanism”. In: *Computational and Mathematical Methods in Medicine* 2021, pp. 1–18.

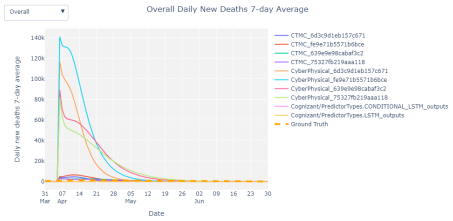
Youn, Seokjun, H. Neil Geismar, and Michael Pinedo (2022). “Planning and Scheduling in Healthcare for Better Care Coordination: Current Understanding, Trending Topics, and Future Opportunities”. In: *Production and Operations Management*. DOI: 10.1111/poms.13867.

Ziedins, Ilze, Cameron Walker, and Michael O’Sullivan (N.D.). “Modelling Covid-19 in the New Zealand Healthcare System”. unpublished.

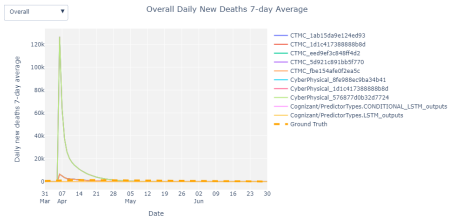
Appendices

The appendices contain plots of the new deaths metrics for the two case studies – in §A.1 and §A.2 respectively – and the full example JSON files for input, output and calibration parameters – in §A.3, §A.4 and §A.5 respectively. Note that these appendices start on the next page due to the size of the figures.

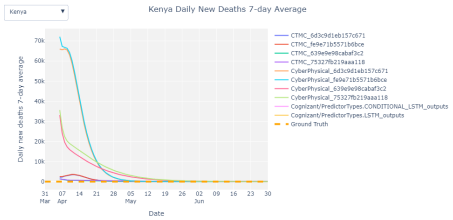
A.1 Case study 1: Aotearoa | New Zealand, Kenya, Sweden and the United Kingdom – New deaths metrics



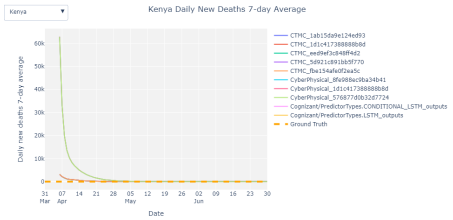
(a) Overall new deaths – short



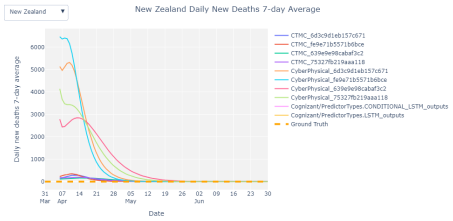
(b) Overall new deaths – long



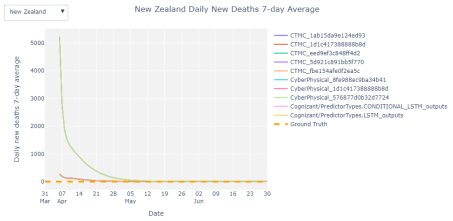
(c) Kenya new deaths – short



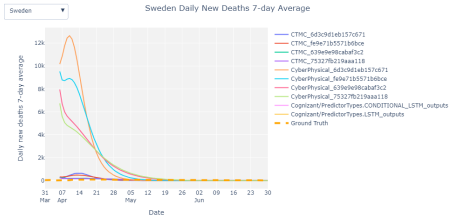
(d) Kenya new deaths – long



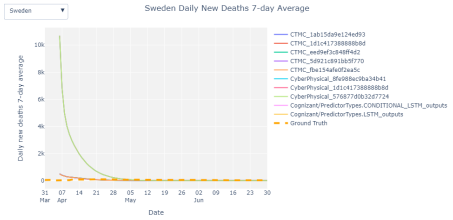
(e) Aotearoa | NZ new deaths – short



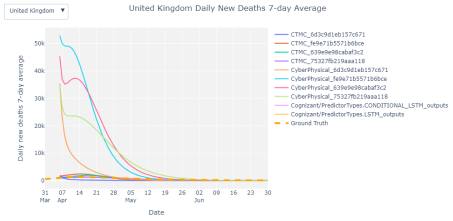
(f) Aotearoa | NZ new deaths – long



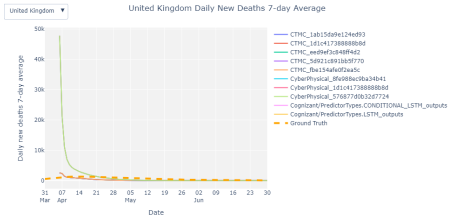
(g) Sweden new deaths – short



(h) Sweden new deaths – long

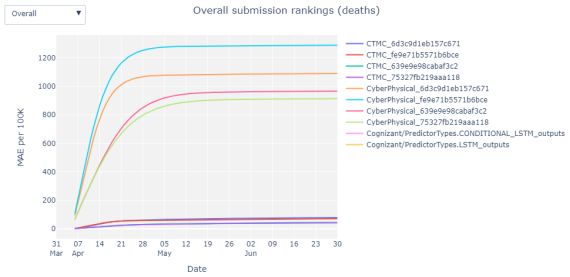


(i) United Kingdom new deaths – short

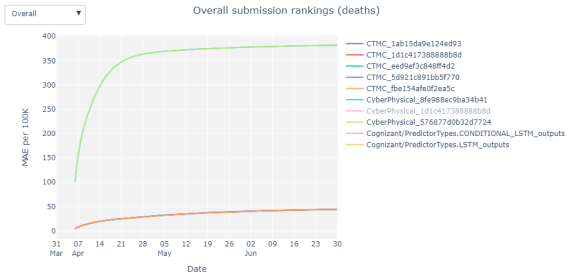


(j) United Kingdom new deaths – long

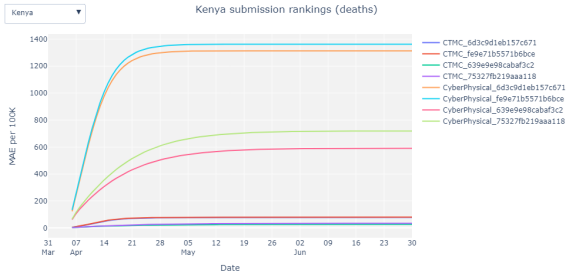
Figure A.1: Comparing new deaths metric: short on the left, long on the right – same definitions as figure 5.1



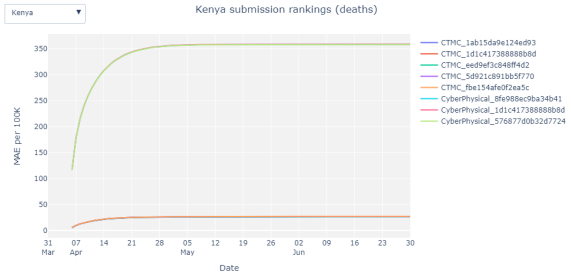
(a) Overall deaths ranking – short



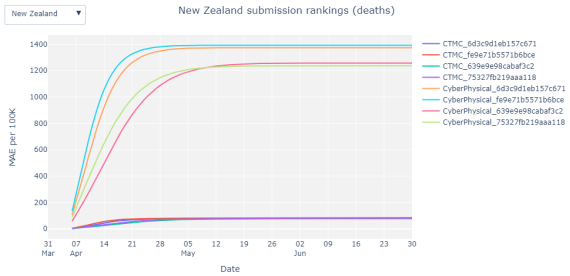
(b) Overall deaths ranking – long



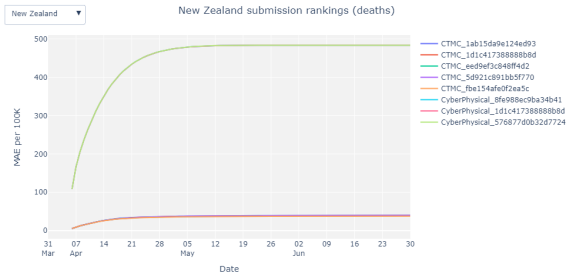
(c) Kenya deaths ranking – short



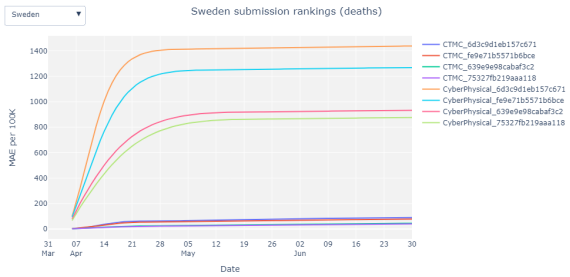
(d) Kenya deaths ranking – long



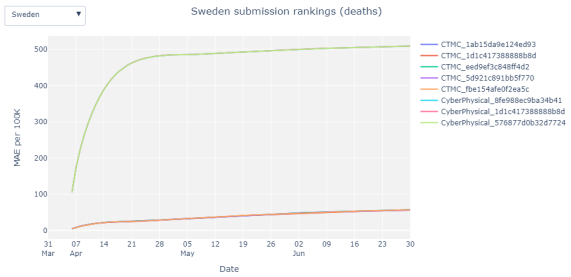
(e) Aotearoa | NZ deaths ranking – short



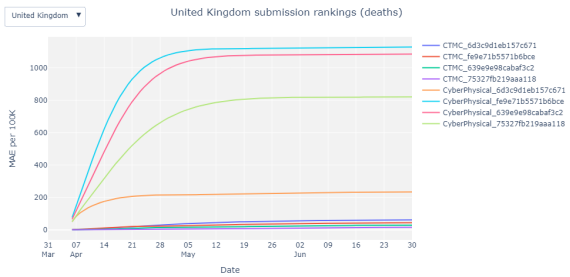
(f) Aotearoa | NZ deaths ranking – long



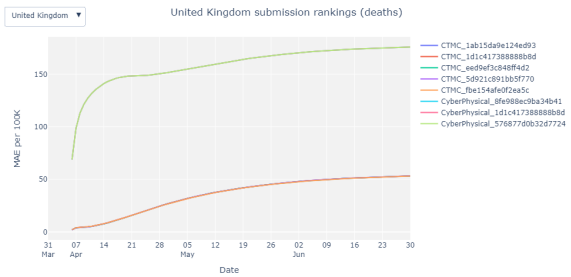
(g) Sweden deaths ranking – short



(h) Sweden deaths ranking – long



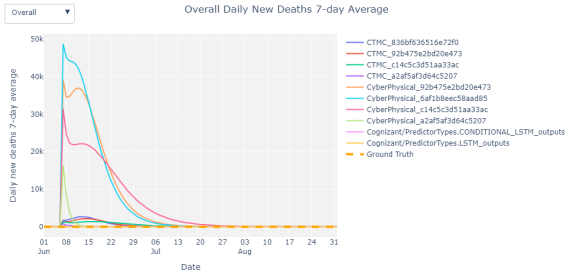
(i) United Kingdom deaths ranking – short



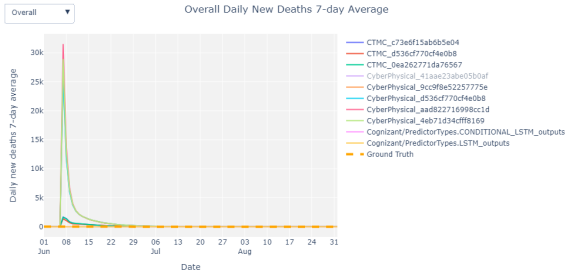
(j) United Kingdom deaths ranking – long

Figure A.2: Comparing deaths ranking metric: short on the left, long on the right – same definitions as figure 5.1

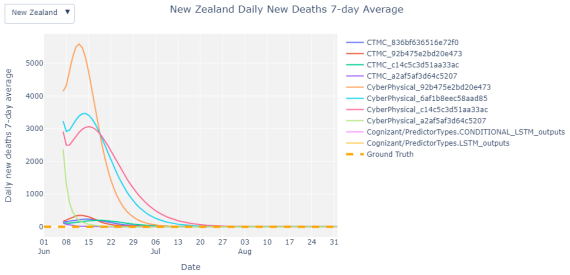
A.2 Case study 2: Aotearoa | New Zealand and Sudan – New Deaths Metrics



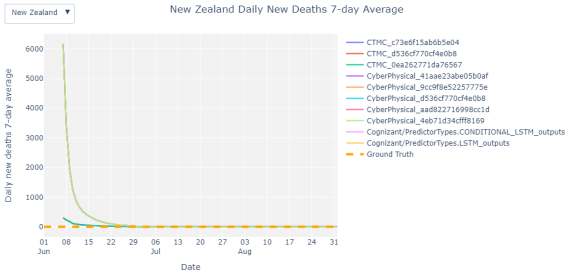
(a) Overall new deaths – short



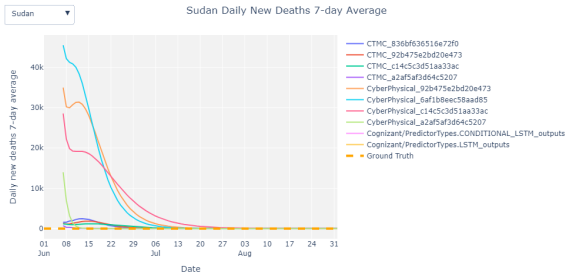
(b) Overall new deaths – long



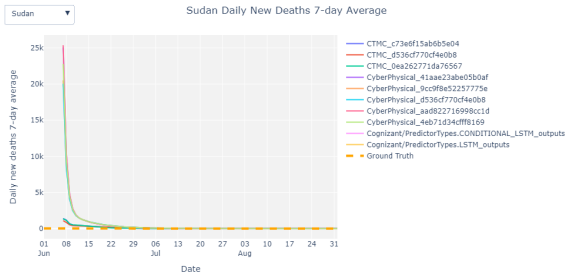
(c) Aotearoa | NZ new deaths – short



(d) Aotearoa | NZ new deaths – long

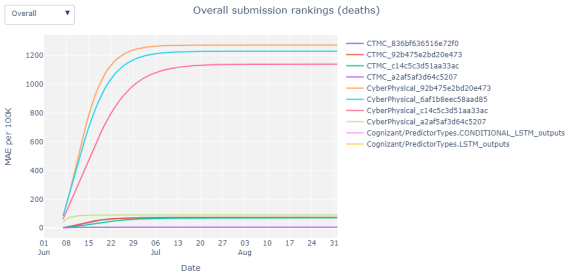


(e) Sudan new deaths – short

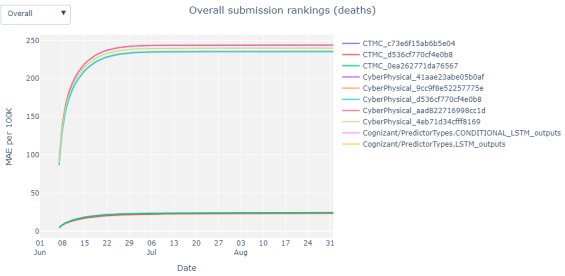


(f) Sudan new deaths – long

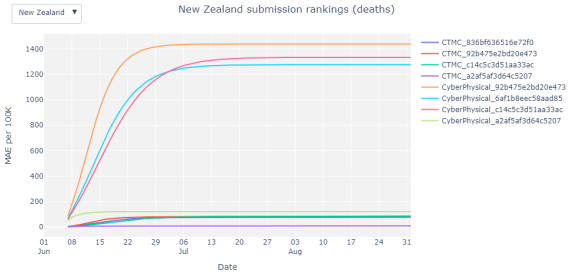
Figure A.3: Comparing new deaths metric: short on the left, long on the right – same definitions as figure 5.1



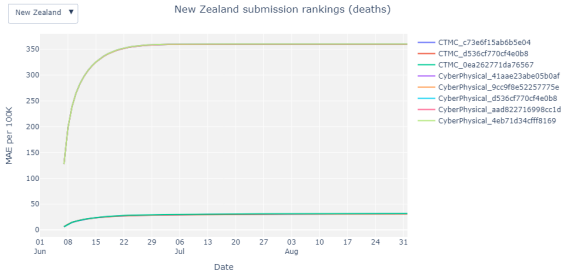
(a) Overall deaths ranking – short



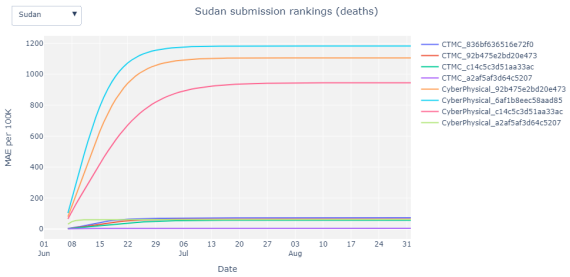
(b) Overall deaths ranking – long



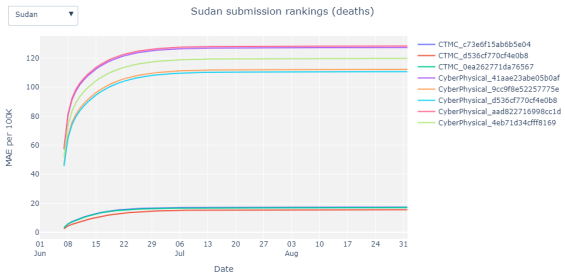
(c) Aotearoa | NZ deaths ranking – short



(d) Aotearoa | NZ deaths ranking – long



(e) Sudan deaths ranking – short



(f) Sudan deaths ranking – long

Figure A.4: Comparing deaths ranking metric: short on the left, long on the right – same definitions as figure 5.1

A.3 Example input JSON file

The following JSON file is the input JSON file for Case Study 2: Aotearoa | New Zealand and Sudan – see §5.2.

```
{
  "global": {
    "parameters": {
      "npi_list": {
        "code": "1.1.1.1",
        "list": [
          "C1_School closing",
          "C2_Workplace closing",
          "C3_Cancel public events",
          "C4_Restrictions on gatherings",
          "C5_Close public transport",
```

```
          "C6_Stay at home requirements",
          "C7_Restrictions on internal movement",
          "C8_International travel controls",
          "E1_Income support",
          "E2_Debt/contract relief",
          "E3_Fiscal measures",
          "E4_International support",
          "H1_Public information campaigns",
          "H2_Testing policy",
          "H3_Contact tracing",
          "H4_Emergency investment in healthcare",
          "H5_Investment in vaccines",
          "H6_Facial Coverings",
          "H7_Vaccination policy",
          "H8_Protection of elderly people",
          "M1_Wildcard",
          "V1_Vaccine Prioritisation (summary)",
          "V2A_Vaccine Availability (summary)",
          "V2B_Vaccine age eligibility/availability age floor (
             general population summary)",
          "V2C_Vaccine age eligibility/availability age floor (at
             risk summary)",
          "V2D_Medically/ clinically vulnerable (Non-elderly)",
          "V2E_Education",
          "V2F_Frontline workers  (non healthcare)",
          "V2G_Frontline workers  (healthcare)",
          "V3_Vaccine Financial Support (summary)",
          "V4_Mandatory Vaccination (summary)"
        ],
        "max_npi_level": {
          "code": "1.1.1.1.2",
          "value": 4
        }
      },
      "variant_list": {
        "code": "1.1.1.2",
        "list": [
          "alpha",
          "delta",
          "omicron"
        ]
      },
      "reproductive_numbers": {
        "code": "1.1.1.2.1",
```

```
        "rates": [{
            "variant": "alpha",
            "R0": 2.79,
            "comment": "Estimation of basic reproductive number aka
                R0 of the Alpha  variant of SARS-CoV-2 was sourced
               from Farmaz, S., Yousefian, N., Tehranipoor, P. &
               Kowsari, Z. (2022) URL: https://journals.plos.org/
               plosone/article?id=10.1371/journal.pone.0265489.
               These can be adjusted to suit different scenarios. "
          },
          {
            "variant": "delta",
            "R0": 5.08,
            "comment": "Estimation of basic reproductive number aka
                R0 of the Delta variant of SARS-CoV-2 was sourced
               from Liu, Y.& Roclov, J. (2022) URL: https://www.
               southernhealth.nz/sites/default/files/2021-10/The%20
               reproductive%20number%20of%20the%20Delta%20variant%2
               0of%20SARS-CoV-2%20is%20far%20higher%20compared%20to
               %20the%20ancestral%20SARS-CoV-2%20virus.pdf.These
               can be adjusted to suit different scenarios."
          },
          {
            "variant": "omicron",
            "R0": 9.5,
            "comment": "Estimation of basic reproductive number aka
                R0 of the Omicron variant of SARS-CoV-2 was sourced
                from Lui, Y. & Rocklove, J. (2022) URL: https://www
               .ncbi.nlm.nih.gov/pmc/articles/PMC8992231/ . These
               can be adjusted to suit different scenarios. "
          }
        ],
        "comment": "Transmission rates, sources provided"
      },
      "stages": {
        "code": "1.1.1.3",
        "list": ["S", "E", "P", "I1", "I2", "R", "D"],
        "list_alt": ["Sus", "Exp", "Pre", "I1", "I2", "Rec", "Dec"]
      },
      "transition_rates": {
        "code": "1.1.1.3.2",
        "file": "transition_rates.csv",
        "comment": "CSV file with transition rate matrix, in same
           directory as this JSON. Columns are Variant, CountryName
```

```
          , CountryCode, RegionName, RegionCode, Jurisdiction,
          StageFrom, StageTo, Rate. The formula used to derive
          rates can be found in formula_extra_transition_rates.csv
          . Note that S (Sus) to E (Exp) is not needed as it will
          be calculated from the effective reproductive number.
          The source of the data is the Covid modelling by Mike O'
          Sullivan, Cameron Walker and Ilze Ziedins - see Overleaf
           report and covid-19-parameters_current.xlsx"
      },
      "relative_infectiousness_presymptomatic": {
        "code": "1.1.1.3.3",
        "value": 0.5,
        "comment": "The source of the value is the Covid modelling
           by Mike O'Sullivan, Cameron Walker and Ilze Ziedins -
           see covid-19-parameters_current.xlsx"
      },
      "countries_modelled": {
        "code": "1.1.1.4",
        "list": [
          "New Zealand",
          "Sudan"
          ]
      },
      "IFR": {
        "code": "1.1.1.5",
        "list": [
          "IFR0", "IFR1"
          ],
        "comment": "IFR0 is IFR when ICU is under capacity and IFR1
            is IFR when ICU is over capacity"
      }
    }
  },
  "location": {
    "parameters": {
      "variant_proportions": {
        "code": "1.2.1.1",
        "list" : [
          {
            "country": "New Zealand",
            "list": [
              {
                "variant": "alpha",
                "proportion": 1.0,
```

```
                "comment": "Where did this estimate come from"
              },
              {
                "variant": "delta",
                "proportion": 0.0,
                "comment": "Where did this estimate come from"
              },
              {
                "variant": "omicron",
                "proportion": 0.0,
                "comment": "Where did this estimate come from"
              }
            ]
          },
          {
            "country": "Sudan",
            "list": [
              {
                "variant": "alpha",
                "proportion": 1.0,
                "comment": "Where did this estimate come from"
              },
              {
                "variant": "delta",
                "proportion": 0.0,
                "comment": "Where did this estimate come from"
              },
              {
                "variant": "omicron",
                "proportion": 0.0,
                "comment": "Where did this estimate come from"
              }
            ]
          }
        ]
      },
      "effective_transmission_rates": {
        "code": "1.2.1.2",
        "function": "basicTransmission"
      },
      "extra_stages": {
        "code": "1.2.1.3",
        "list": ["W1", "ICU"]
      },
```

```
      "extra_transition_rates": {
        "code": "1.2.1.3.2",
        "file": "extra_transition_rates.csv",
        "comment": "CSV file with transition rate matrix including
           extra stages, in same directory as this JSON. Columns
           are Variant, IFR4ICUCapacity, CountryName, CountryCode,
           RegionName, RegionCode, Jurisdiction, StageFrom, StageTo
           , Rate. The formula used to derive rates can be found in
            formula_extra_transition_rates.csv. The source of the
           data is the Covid modelling by Mike O'Sullivan, Cameron
           Walker and Ilze Ziedins - see Overleaf report and covid-
           19-parameters_current.xlsx"
      },
      "initial_numbers": {
        "code": "1.2.1.4",
        "file": "initial_numbers.csv",
        "comment": "CSV file with initial numbers, in same
           directory as this JSON"
      },
      "model_horizon": {
        "code": "1.2.1.5",
        "start": "1-06-2020",
        "finish": "1-09-2020",
        "interval_in_days": 1
      },
      "npi_schedule": {
        "code": "1.2.1.6",
        "file": "npi_schedule.csv",
        "comment": "CSV file with the planned schedule for NPIs in
           the given location"
      },
      "health_system_parameters": {
        "code": "1.2.1.7",
        "list": [
          {
            "country": "Sudan",
            "list": [
              {
                "parameter": "number_icu_beds",
                "code": "1.2.1.7.1",
                "value": 184,
                "comment": "Estimate from National Library for
                   Ministries Healthcare Africa"
              },
```

```
            {
              "parameter": "prop_to_icu",
              "code": "1.2.1.7.2",
              "value": 0.11,
              "comment": "The source of the value is the Covid
                 modelling by Mike O'Sullivan, Cameron Walker and
                  Ilze Ziedins - see covid-19-parameters_current.
                 xlsx"
            },
            {
              "parameter": "relative_infectiousness_ward",
              "code": "1.2.1.7.3",
              "value": 0.05,
              "comment": "The source of the value is the Covid
                 modelling by Mike O'Sullivan, Cameron Walker and
                  Ilze Ziedins - see covid-19-parameters_current.
                 xlsx"
            },
            {
              "parameter": "relative_infectiousness_icu",
              "code": "1.2.1.7.4",
              "value": 0.01,
              "comment": "The source of the value is the Covid
                 modelling by Mike O'Sullivan, Cameron Walker and
                  Ilze Ziedins - see covid-19-parameters_current.
                 xlsx"
            }
          ]
        },
        {
          "country": "New Zealand",
          "list": [
            {
              "parameter": "number_icu_beds",
              "code": "1.2.1.7.1",
              "value": 233,
              "comment": "The source of the value is the Covid
                 modelling by Mike O'Sullivan, Cameron Walker and
                  Ilze Ziedins - see covid-19-parameters_current.
                 xlsx"
            },
            {
              "parameter": "prop_to_icu",
              "code": "1.2.1.7.2",
```

```
                "value": 0.14,
                "comment": "The source of the value is the Covid
                   modelling by Mike O'Sullivan, Cameron Walker and
                    Ilze Ziedins - see covid-19-parameters_current.
                   xlsx"
              },
              {
                "parameter": "relative_infectiousness_ward",
                "code": "1.2.1.7.3",
                "value": 0.05,
                "comment": "The source of the value is the Covid
                   modelling by Mike O'Sullivan, Cameron Walker and
                    Ilze Ziedins - see covid-19-parameters_current.
                   xlsx"
              },
              {
                "parameter": "relative_infectiousness_icu",
                "code": "1.2.1.7.4",
                "value": 0.01,
                "comment": "The source of the value is the Covid
                   modelling by Mike O'Sullivan, Cameron Walker and
                    Ilze Ziedins - see covid-19-parameters_current.
                   xlsx"
              }
            ]
          }
        ]
      }
    },
    "data": {
      "code": "1.2.2",
      "file": "../OxCGRT_latest.csv",
      "format": " CountryName,CountryCode,RegionName,RegionCode,
         Jurisdiction,Date,C1_School closing,C1_Flag,C2_Workplace
         closing,C2_Flag,C3_Cancel public events,C3_Flag,C4
         _Restrictions on gatherings,C4_Flag,C5_Close public
         transport,C5_Flag,C6_Stay at home requirements,C6_Flag,C7
         _Restrictions on internal movement,C7_Flag,C8
         _International travel controls,E1_Income support,E1_Flag,E
         2_Debt/contract relief,E3_Fiscal measures,E4_International
          support,H1_Public information campaigns,H1_Flag,H2
         _Testing policy,H3_Contact tracing,H4_Emergency investment
          in healthcare,H5_Investment in vaccines,H6_Facial
         Coverings,H6_Flag,H7_Vaccination policy,H7_Flag,H8
```

```
        _Protection of elderly people,H8_Flag,M1_Wildcard,V1
        _Vaccine Prioritisation (summary),V2A_Vaccine Availability
         (summary),V2B_Vaccine age eligibility/availability age
        floor (general population summary),V2C_Vaccine age
        eligibility/availability age floor (at risk summary),V2
        D_Medically/ clinically vulnerable (Non-elderly),V2
        E_Education,V2F_Frontline workers  (non healthcare),V2
        G_Frontline workers  (healthcare),V3_Vaccine Financial
        Support (summary),V4_Mandatory Vaccination (summary),
        ConfirmedCases,ConfirmedDeaths,StringencyIndex,
        StringencyIndexForDisplay,StringencyLegacyIndex,
        StringencyLegacyIndexForDisplay,GovernmentResponseIndex,
        GovernmentResponseIndexForDisplay,ContainmentHealthIndex,
        ContainmentHealthIndexForDisplay,EconomicSupportIndex,
        EconomicSupportIndexForDisplay",
247     "comment": "CSV file with specified format, from https://
        github.com/OxCGRT/covid-policy-tracker-legacy/blob/main/
        legacy_data_202207/OxCGRT_latest.csv. Each row contains
        data as described in the format field and definitions of
        the NPI codes are given in the codebook documentation
        https://github.com/OxCGRT/covid-policy-tracker/blob/master
        /documentation/codebook.md"
248   },
249   "extra_data": {
250     "code": "1.2.3",
251     "file": "extra8.csv",
252     "format": " CountryName,CountryCode,RegionName,RegionCode,
        Jurisdiction,Date,ExternalInfectedArrivals,
        ConfirmedHospitalised,ConfirmedICU",
253     "comment": "CSV file with extra (optional) with extra data"
254   }
255 },
256 "model": {
257   "effective_transmission_coeffs": {
258       "code": "1.3.1.1",
259       "file": "transmission_coeffs.csv",
260       "comment": "CSV file with table, rows = NPIs, columns =
          level of NPI, variant, entries = coeff used in effective
           transmission function, i.e., 1.2.1.1. Initial
          assumption is that these are the same across alpha,
          delta and omicron, but these will be adjusted during
          calibration. The initial assumption is that these are
          the same across countries, but these will be adjusted
          during calibration"
```

```
    },
    "models_used": {
      "list": [
        "Cognizant",
        "CyberPhysical",
        "CTMC"
      ]
    },
    "output_location" : {
      "code": "1.3.1.2",
      "location" : {
        "CTMC": "CTMC_outputs",
        "CyberPhysical": "CyberPhysical_outputs",
        "Cognizant": "Cognizant_outputs"
      }
    }
  }
}
```

A.4 Example output JSON file

The following JSON file is the output JSON file for Case Study 2: Aotearoa | New Zealand and Sudan – see §5.2.

```
{
  "estimates": {
    "code": "2.1.1",
    "file": "model_outputs.csv",
    "format": " CountryName,CountryCode,RegionName,RegionCode,
       Jurisdiction,Date,ConfirmedCases,ConfirmedDeaths",
    "comment": "CSV file with specified format, adapted from from
       https://github.com/OxCGRT/covid-policy-tracker-legacy/blob/
       main/legacy_data_202207/OxCGRT_latest.csv. Each row contains
        data as described in the format field. From this file we
       can get 2.1.1.1 Location information, 2.1.1.2 Time periods
       for model horizon, 2.1.1.3 Case number estimates, 2.1.1.4
       Recovered number estimates, and 2.1.1.5 Death estimates"
  },
  "data": {
    "code": "2.1.2",
    "file": "OxCGRT_latest.csv",
    "format": " CountryName,CountryCode,RegionName,RegionCode,
       Jurisdiction,Date,C1_School closing,C1_Flag,C2_Workplace
```

```
        closing,C2_Flag,C3_Cancel public events,C3_Flag,C4
        _Restrictions on gatherings,C4_Flag,C5_Close public
        transport,C5_Flag,C6_Stay at home requirements,C6_Flag,C7
        _Restrictions on internal movement,C7_Flag,C8_International
        travel controls,E1_Income support,E1_Flag,E2_Debt/contract
        relief,E3_Fiscal measures,E4_International support,H1_Public
         information campaigns,H1_Flag,H2_Testing policy,H3_Contact
        tracing,H4_Emergency investment in healthcare,H5_Investment
        in vaccines,H6_Facial Coverings,H6_Flag,H7_Vaccination
        policy,H7_Flag,H8_Protection of elderly people,H8_Flag,M1
        _Wildcard,V1_Vaccine Prioritisation (summary),V2A_Vaccine
        Availability (summary),V2B_Vaccine age eligibility/
        availability age floor (general population summary),V2
        C_Vaccine age eligibility/availability age floor (at risk
        summary),V2D_Medically/ clinically vulnerable (Non-elderly),
        V2E_Education,V2F_Frontline workers  (non healthcare),V2
        G_Frontline workers  (healthcare),V3_Vaccine Financial
        Support (summary),V4_Mandatory Vaccination (summary),
        ConfirmedCases,ConfirmedDeaths,StringencyIndex,
        StringencyIndexForDisplay,StringencyLegacyIndex,
        StringencyLegacyIndexForDisplay,GovernmentResponseIndex,
        GovernmentResponseIndexForDisplay,ContainmentHealthIndex,
        ContainmentHealthIndexForDisplay,EconomicSupportIndex,
        EconomicSupportIndexForDisplay",
    "comment": "CSV file with specified format, from https://github
        .com/OxCGRT/covid-policy-tracker-legacy/blob/main/
        legacy_data_202207/OxCGRT_latest.csv. Each row contains data
         as described in the format field and definitions of the NPI
         codes are given in the codebook documentation https://
        github.com/OxCGRT/covid-policy-tracker/blob/master/
        documentation/codebook.md"
  }
}
```

A.5 Example calibration JSON file

The following JSON file is the calibration JSON file for Case Study 2: Aotearoa | New Zealand and Sudan – see §5.2.

```
{
  "global": {
    "parameters": {
      "reproductive_numbers": {
```

```
        "code": "1.1.1.2.1",
        "rates": [
          {
            "variant": "alpha",
            "R0": { "type": "bounds", "lb" : 2.29, "ub": 3.29 }
          },
          {
            "variant": "delta",
            "R0": { "type": "bounds", "lb" : 4.58, "ub": 5.58 }
          },
          {
            "variant": "omicron",
            "R0": { "type": "bounds", "lb": 9, "ub": 10 }
          }
        ]
      },
      "relative_infectiousness_presymptomatic": {
        "code": "1.1.1.3.3",
        "value": { "type": "bounds", "lb": 0, "ub": 1}
      }
    }
  },
  "location": {
    "parameters": {
      "variant_proportions": {
        "code": "1.2.1.1",
        "list" : [
          {
            "country": "New Zealand",
            "list": [
              {
                "variant": "alpha",
                "proportion": { "type": "bounds", "lb": 0, "ub": 1}
                  ,
                "comment": "Where did this estimate come from"
              },
              {
                "variant": "delta",
                "proportion": { "type": "bounds", "lb": 0, "ub": 1}
                  ,
                "comment": "Where did this estimate come from"
              },
              {
                "variant": "omicron",
```

```
                "proportion": { "type": "bounds", "lb": 0, "ub": 1}
                  ,
                "comment": "Where did this estimate come from"
              }
            ]
          },
          {
            "country": "Sudan",
            "list": [
              {
                "variant": "alpha",
                "proportion": { "type": "bounds", "lb": 0, "ub": 1}
                  ,
                "comment": "Where did this estimate come from"
              },
              {
                "variant": "delta",
                "proportion": { "type": "bounds", "lb": 0, "ub": 1}
                  ,
                "comment": "Where did this estimate come from"
              },
              {
                "variant": "omicron",
                "proportion": { "type": "bounds", "lb": 0, "ub": 1}
                  ,
                "comment": "Where did this estimate come from"
              }
            ]
          }
        ]
      },
      "initial_numbers": {
        "code": "1.2.1.3",
        "file": { "type": "custom" },
        "comment": "CSV file with initial numbers, in same
           directory as this JSON"
      },
      "npi_schedule": {
        "code": "1.2.1.5",
        "file": { "type": "custom" },
        "comment": "CSV file with the planned schedule for NPIs in
           the given location"
      },
      "health_system_parameters": {
```

```
        "code": "1.2.1.7",
        "list": [
          {
            "country": "Sudan",
            "list": [
              {
                "parameter": "number_icu_beds",
                "code": "1.2.1.7.1",
                "value": { "type": "integer", "lb" : 150, "ub": 200
                    },
                "comment": "The source of the value is the Covid
                   modelling by Mike O'Sullivan, Cameron Walker and
                    Ilze Ziedins - see covid-19-parameters_current.
                   xlsx"
              },
              {
                "parameter": "prop_to_icu",
                "code": "1.2.1.7.2",
                "value": { "type": "bounds", "lb" : 0.5, "ub": 0.8
                   },
                "comment": "The source of the value is the Covid
                   modelling by Mike O'Sullivan, Cameron Walker and
                    Ilze Ziedins - see covid-19-parameters_current.
                   xlsx"
              },
              {
                "parameter": "relative_infectiousness_ward",
                "code": "1.2.1.7.3",
                "value": { "type": "bounds", "lb" : 0.03, "ub": 0.0
                   7 },
                "comment": "The source of the value is the Covid
                   modelling by Mike O'Sullivan, Cameron Walker and
                    Ilze Ziedins - see covid-19-parameters_current.
                   xlsx"
              },
              {
                "parameter": "relative_infectiousness_icu",
                "code": "1.2.1.7.4",
                "value": { "type": "bounds", "lb" : 0.0005, "ub": 0
                   .0015 },
                "comment": "The source of the value is the Covid
                   modelling by Mike O'Sullivan, Cameron Walker and
                    Ilze Ziedins - see covid-19-parameters_current.
                   xlsx"
```

```
113                 },
114                 {
115                   "parameter": "prop_to_hospital",
116                   "code": "1.2.1.7.5",
117                   "value": { "type": "bounds", "lb" : 0.01, "ub": 0.0
                        5 },
118                   "comment": "The source of the value is the Covid
                        modelling by Mike O'Sullivan, Cameron Walker and
                         Ilze Ziedins - see covid-19-parameters_current.
                        xlsx"
119                 },
120                 {
121                   "parameter": "hospLoS",
122                   "code": "1.2.1.7.6",
123                   "value": { "type": "bounds", "lb" : 18, "ub": 25 },
124                   "comment": "The source of the value is the Covid
                        modelling by Mike O'Sullivan, Cameron Walker and
                         Ilze Ziedins - see covid-19-parameters_current.
                        xlsx"
125                 },
126                 {
127                   "parameter": "ICULoS",
128                   "code": "1.2.1.7.7",
129                   "value": { "type": "bounds", "lb" : 15, "ub": 21 },
130                   "comment": "The source of the value is the Covid
                        modelling by Mike O'Sullivan, Cameron Walker and
                         Ilze Ziedins - see covid-19-parameters_current.
                        xlsx"
131                 }
132               ]
133             },
134             {
135               "country": "New Zealand",
136               "list": [
137                 {
138                   "parameter": "number_icu_beds",
139                   "code": "1.2.1.7.1",
140                   "value": { "type": "integer", "lb" : 200, "ub": 250
                          },
141                   "comment": "The source of the value is the Covid
                        modelling by Mike O'Sullivan, Cameron Walker and
                         Ilze Ziedins - see covid-19-parameters_current.
                        xlsx"
142                 },
```

```
143                 {
144                   "parameter": "prop_to_icu",
145                   "code": "1.2.1.7.2",
146                   "value": { "type": "bounds", "lb" : 0.5, "ub": 0.8
                        },
147                   "comment": "The source of the value is the Covid
                        modelling by Mike O'Sullivan, Cameron Walker and
                         Ilze Ziedins - see covid-19-parameters_current.
                        xlsx"
148                 },
149                 {
150                   "parameter": "relative_infectiousness_ward",
151                   "code": "1.2.1.7.3",
152                   "value": { "type": "bounds", "lb" : 0.03, "ub": 0.0
                        7 },
153                   "comment": "The source of the value is the Covid
                        modelling by Mike O'Sullivan, Cameron Walker and
                         Ilze Ziedins - see covid-19-parameters_current.
                        xlsx"
154                 },
155                 {
156                   "parameter": "relative_infectiousness_icu",
157                   "code": "1.2.1.7.4",
158                   "value": { "type": "bounds", "lb" : 0.0005, "ub": 0
                        .0015 },
159                   "comment": "The source of the value is the Covid
                        modelling by Mike O'Sullivan, Cameron Walker and
                         Ilze Ziedins - see covid-19-parameters_current.
                        xlsx"
160                 },
161                 {
162                   "parameter": "prop_to_hospital",
163                   "code": "1.2.1.7.5",
164                   "value": { "type": "bounds", "lb" : 0.01, "ub": 0.0
                        5 },
165                   "comment": "The source of the value is the Covid
                        modelling by Mike O'Sullivan, Cameron Walker and
                         Ilze Ziedins - see covid-19-parameters_current.
                        xlsx"
166                 },
167                 {
168                   "parameter": "hospLoS",
169                   "code": "1.2.1.7.6",
170                   "value": { "type": "bounds", "lb" : 18, "ub": 25 },
```

```
                "comment": "The source of the value is the Covid
                   modelling by Mike O'Sullivan, Cameron Walker and
                    Ilze Ziedins - see covid-19-parameters_current.
                   xlsx"
              },
              {
                "parameter": "ICULoS",
                "code": "1.2.1.7.7",
                "value": { "type": "bounds", "lb" : 15, "ub": 21 },
                "comment": "The source of the value is the Covid
                   modelling by Mike O'Sullivan, Cameron Walker and
                    Ilze Ziedins - see covid-19-parameters_current.
                   xlsx"
              }
            ]
          }
        ]
      }
    }
  },
  "model": {
    "effective_transmission_coeffs": {
        "code": "1.3.1.2",
        "file": { "type": "custom", "pm": 0.1 },
        "comment": "CSV file with table, rows = NPIs, columns =
           level of NPI, variant, entries = coeff used in effective
            transmission function, i.e., 1.2.1.2. Initial
           assumption is that these are the same across alpha,
           delta and omicron, but these will be adjusted during
           calibration. The initial assumption is that these are
           the same across countries, but these will be adjusted
           during calibration"
    }
  }
}
```