

# Hierarchical Emulation of Stochastic Agent-Based Models

SIAM UQ 2024

Andrew Iskauskas

Durham University

28th February, 2024

LONDON  
SCHOOL of  
HYGIENE  
& TROPICAL  
MEDICINE



Durham  
University



# Talk Structure

- 1 Introduction and Motivation
- 2 Emulation and History Matching
- 3 Application: HPVsim
- 4 Summary

# Complex Models of Real-World Phenomena

Complex computer models (or *simulators*) are used in a variety of fields, including

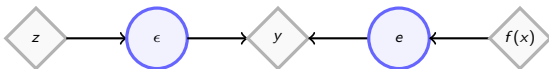
- Oil Industry (oil reservoir and geology models) [3]
- Climate Science (climate models of global warming) [11]
- Systems Biology (genetic and metabolic network models) [10]
- Cosmology (galaxy formation simulations) [9]
- Nuclear Physics (quantum many-body models of nuclei) [4]
- **Epidemiology** HIV, TB, Covid, ... [1, 7]

Simulators are often computationally expensive: a full exploration of the parameter space using only the simulator is infeasible.

# Uncertainty Structure for Models

Consider a simulator  $f(x)$  that represents a physical process  $y$ , from which we may obtain observed quantities  $z$ . Two main sources of uncertainty are

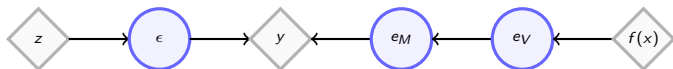
- **Observational error.** Our observations  $z$  of  $y$  are made imperfectly:  $z = y + \epsilon$ ;
- **Model discrepancy.** Our simulator  $f(x)$  cannot faithfully represent the process  $y$ :  $y = f(x) + e$ .



# Uncertainty Structure for Models

Consider a **stochastic** simulator  $f(x)$  that represents a physical process  $y$ , from which we may obtain observed quantities  $z$ . Two main sources of uncertainty are

- **Observational error.** Our observations  $z$  of  $y$  are made imperfectly:  $z = y + \epsilon$ ;
- **Model discrepancy.** Our simulator  $f(x)$  cannot faithfully represent the process  $y$ :  $y = f(x) + e$ . *Moreover, repeated evaluations of  $f(x)$  at the same point  $x$  give different values.*



# The Emulator

An *emulator* is a statistical approximation of a complex computer simulator [2].

Let  $f(x)$  be an output from the simulator at a given parameter set  $x \in \mathbb{R}^d$ , corresponding to some real physical process  $y$ . Then we define a emulator for output  $f(x)$  as

$$g(x) = \sum_i \beta_i h_i(x_A) + u(x_A) + w(x)$$

The  $h_i(x_A)$  are a collection of basis functions in the *active variables*  $x_A$ ,  $\beta_i$  the coefficients,  $u(x_A)$  a weakly stationary process in the active variables, and  $w(x)$  a 'nugget term'.

Pragmatic choice: consider Bayes Linear emulators, so only need prior beliefs for expectations, variances, and covariances.

# The Bayes Linear Update Equations

Let  $D = \{f(x_1), f(x_2), \dots, f(x_n)\}$  be runs from the simulator at points  $x_1, \dots, x_n$ . The Bayes linear update equations give the emulator's posterior prediction for the model output at an unseen point  $x$ , given  $D$ :

$$\begin{aligned}E_D[g(x)] &= E[g(x)] + \text{Cov}[g(x), D]\text{Var}[D]^{-1}(D - E[D]), \\ \text{Var}_D[g(x)] &= \text{Var}[g(x)] - \text{Cov}[g(x), D]\text{Var}[D]^{-1}\text{Cov}[D, g(x)].\end{aligned}$$

# Stochastic Emulation

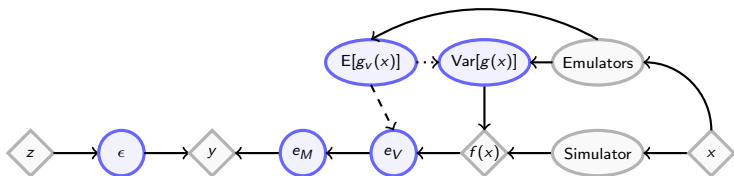
The quantity  $\text{Var}[g(x)]$  encodes the uncertainty of the emulator prediction. For stochastic models, we apply a *hierarchical* approach to accurately account for model variability.

- Train emulator  $g_V(x) = \sum_i \beta_{Vi} h_{Vi}(x_{VA}) + u_V(x_{VA}) + w_V(x)$  to the *stochasticity* of the model output;
- Use  $E[g_V(x)]$  as an informed prior for  $\text{Var}[g(x)]$ , and create output emulators  $g(x)$ .
- $g_V(x)$  does not just contribute to the prior for  $g(x)$ ; it also helps us encode uncertainty due to stochasticity,  $e_V$ .



# Uncertainty Structure: Emulation

Add in the emulators into our schematic for the model structure:



# Variance Emulation

There are complications in emulating stochastic systems:

- Our model output is not consistent with repeated evaluations: output  $f(x)$  gives  $(f_1(x), f_2(x), \dots, f_n(x))$  which are not the same in general;

# Variance Emulation

There are complications in emulating stochastic systems:

- Our model output is not consistent with repeated evaluations: output  $f(x)$  gives  $(f_1(x), f_2(x), \dots, f_n(x))$  which are not the same in general;
- We can only observe *sample* quantities at  $x$  – for example, the mean  $\bar{f}_n(x)$  and standard deviation  $s_n^2(x)$ ;

# Variance Emulation

There are complications in emulating stochastic systems:

- Our model output is not consistent with repeated evaluations: output  $f(x)$  gives  $(f_1(x), f_2(x), \dots, f_n(x))$  which are not the same in general;
- We can only observe *sample* quantities at  $x$  – for example, the mean  $\bar{f}_n(x)$  and standard deviation  $s_n^2(x)$ ;
- The emulator  $g(x)$  is designed to predict the ‘true’ mean response,  $\mathcal{M}(f(x)) \neq \bar{f}_n(x)$ ...

# Second-Order Exchangeability

Link the 'true' mean of the system to sample quantities using second-order exchangeability:

$$f_k(x) = \mathcal{M}(f(x)) + \mathcal{R}_k(f(x)),$$

where  $\mathcal{R}_k(f(x))$ ,  $k = 1, \dots, n$  are uncorrelated, zero-mean residuals. Then  $E[\bar{f}_n(x)] = E[\mathcal{M}(f(x))]$ , with variability related via  $\text{Var}[\bar{f}_n(x)] = \text{Var}[\mathcal{M}(f(x))] + \text{Var}[\mathcal{R}_k(f(x))]$ . Similar structure for stochasticity via

$$[\mathcal{R}_k(f(x))]^2 = \mathcal{M}(V(x)) + \mathcal{R}_k(V(x)).$$

A similar argument holds if we also emulate covariance, by considering  $\mathcal{R}_k(f_i(x))\mathcal{R}_k(f_j(x))$ .

# History Matching

Given observed data corresponding to a simulator output, what combinations of input parameters could give rise to output consistent with this observation?

*History matching* works on the principle of complementarity: a point  $x$  is considered unsuitable if **even accounting for the uncertainties in the system**, the prediction  $E_D[g(x)]$  cannot be 'close' to the observed value  $z$ . Closeness is defined via an *implausibility* measure

$$I^2(x) = E_D[g(x) - z]^T \text{Var}_D[g(x) - z]^{-1} (E_D[g(x) - z])$$

# Implausibility Structure for Stochastic Systems

Suppose we can only observe sample quantities in reality – i.e. inferring reproduction rate from a (relatively) small sample of the population. Then  $\text{Var}_D[g(x) - z]$  has the form

$$\begin{aligned} \text{Var}_D[g(x)] + \text{Var}[e_M] + \text{Var}[\epsilon] \\ + \frac{1}{m} \left( \text{Exp}_D[g_v(x)] + 2\rho\sqrt{\text{Exp}_D[g_v(x)]}V_{R_\epsilon} + V_{R_\epsilon} + V_{R_e} \right) \end{aligned}$$

# Implausibility Structure for Stochastic Systems

Suppose we can only observe sample quantities in reality – i.e. inferring reproduction rate from a (relatively) small sample of the population. Then  $\text{Var}_D[g(x) - z]$  has the form

$$\text{Var}_D[g(x)] + \text{Var}[e_M] + \text{Var}[\epsilon] \\ + \frac{1}{m} \left( \text{Exp}_D[g_v(x)] + 2\rho\sqrt{\text{Exp}_D[g_v(x)]V_{R_\epsilon}} + V_{R_\epsilon} + V_{R_e} \right)$$

'Deterministic' uncertainty, albeit  $\text{Var}_D[g(x)]$  has informed prior variance based on hierarchical emulation



# Implausibility Structure for Stochastic Systems

Suppose we can only observe sample quantities in reality – i.e. inferring reproduction rate from a (relatively) small sample of the population. Then  $\text{Var}_D[g(x) - z]$  has the form

$$\begin{aligned} & \text{Var}_D[g(x)] + \text{Var}[e_M] + \text{Var}[\epsilon] \\ & + \frac{1}{m} \left( \text{Exp}_D[g_v(x)] + 2\rho\sqrt{\text{Exp}_D[g_v(x)]}V_{R_\epsilon} + V_{R_\epsilon} + V_{R_e} \right) \end{aligned}$$

Contribution to stochasticity from observed sample standard deviations

# Implausibility Structure for Stochastic Systems

Suppose we can only observe sample quantities in reality – i.e. inferring reproduction rate from a (relatively) small sample of the population. Then  $\text{Var}_D[g(x) - z]$  has the form

$$\begin{aligned} & \text{Var}_D[g(x)] + \text{Var}[e_M] + \text{Var}[\epsilon] \\ & + \frac{1}{m} \left( \text{Exp}_D[g_v(x)] + 2\rho\sqrt{\text{Exp}_D[g_v(x)]}V_{R_\epsilon} + V_{R_\epsilon} + V_{R_e} \right) \end{aligned}$$

Variance of the residual variation in measurement and model discrepancy (second-order exchangeability)

# Implausibility Structure for Stochastic Systems

Suppose we can only observe sample quantities in reality – i.e. inferring reproduction rate from a (relatively) small sample of the population. Then  $\text{Var}_D[g(x) - z]$  has the form

$$\begin{aligned} & \text{Var}_D[g(x)] + \text{Var}[e_M] + \text{Var}[\epsilon] \\ & + \frac{1}{m} \left( \text{Exp}_D[g_v(x)] + 2\rho\sqrt{\text{Exp}_D[g_v(x)]V_{R_\epsilon}} + V_{R_\epsilon} + V_{R_e} \right) \end{aligned}$$

Link between model structure and stochasticity: eg over-dispersion of model output vs observation gives  $\rho < 0$ .

# Emulation and HM: Summary

Emulators can **efficiently** and **robustly** predict simulator output at unseen points, given a small collection of known runs.

We can construct **hierarchical** emulators that allow for a nuanced prior determination of stochastic variability and account for **imperfect data** arising from computational constraints.

History matching allows us to leverage the uncertainty structure to find **all** acceptable matches to data arising from our model, accounting for any beliefs we have about observations, model inadequacies, and dispersion.

# Specificational Burden

Bayes linear framework reduces the specificational burden, but there are still some quantities to determine.



$E[\beta]$ ,  $h(x)$ , hyperparameters in  $u(x)$ ,  $\text{Var}[e_M]$ ,  $\text{Var}[\epsilon]$ , *fourth-order* quantities for (co)variance emulation, prior statements on  $V_{R_\epsilon}$ ,  $V_{R_e}$ ,  $\rho \dots$

Some of these must come from expert elicitation. For all the others...

# Obligatory Plug: hmer [5]

## hmer: History Matching and Emulation Package

A set of objects and functions for Bayes Linear emulation and history matching. Core functionality includes automated training of emulators to data, diagnostic functions to ensure suitability, and a variety of proposal methods for generating 'waves' of points. For details on the mathematical background, there are many papers available on the topic (see references attached to function help files); for details of the functions in this package, consult the manual or help files.

Version: 1.5.6  
Depends: R (≥ 4.1.0)  
Imports: [purrr](#), [stringr](#), [tidyr](#), [dplyr](#), [ggplot2](#), [lhs](#), [MASS](#), [R6](#), [viridis](#), [mvtnorm](#), [GGally](#), [rlang](#), [isoband](#), [cluster](#), [pdist](#), [ggbeeswarm](#)  
Suggests: [spelling](#), [knitr](#), [rmarkdown](#), [deSolve](#), [testthat](#) (≥ 3.0.0), [covr](#), [progressr](#)  
Published: 2023-08-30  
Author: Andrew Iskauskas  [aut, cre], TJ McKinley  [aut]  
Maintainer: Andrew Iskauskas <andrew.iskauskas@durham.ac.uk>  
BugReports: <https://github.com/andy-iskauskas/hmer/issues>  
License: [MIT](#) + file [LICENSE](#)  
URL: <https://github.com/andy-iskauskas/hmer>, <https://hmer-package.github.io/website/>

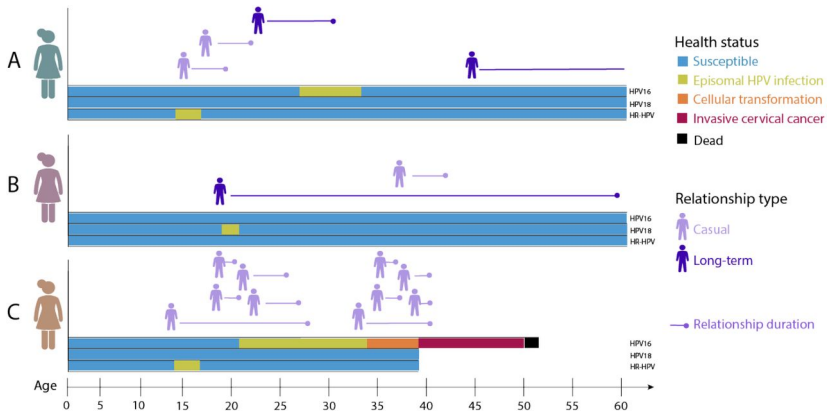


- Automated prior specifications for deterministic and stochastic models
- Diagnostics on emulators
- Robust point proposal schemes
- Visualisation
- Customisability via `Proto_emulator`

# The HPVsim model

- Developed by Institute for Disease Modelling: one of the 'Starsim' models [8, 6]
- Detailed contact structure, sexual networks, genotype-specific parametrisation, . . .
- Large populations handled using dynamic rescaling: computational efficiency without compromised accuracy of results
- <https://docs.idmod.org/projects/hpvsim/en/latest/>

# Natural History





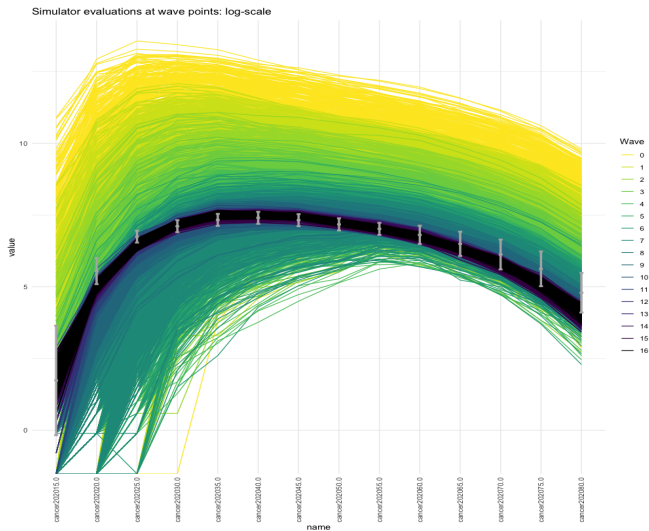
# Problem Statement

- Model run-time per parameter set: 2 minutes to 1 hour
- 22 observational targets: new cancer cases, aggregated by age, and proportion of four genotype classes in individuals with cancer and high-grade lesions (CIN3)
- 33 input parameters identified for calibration problem
- Epidemiological interest:
  - What age group should be targeted for HPV screening? What about for vaccination?
  - What would the impact of interventions be on future cancer incidence?

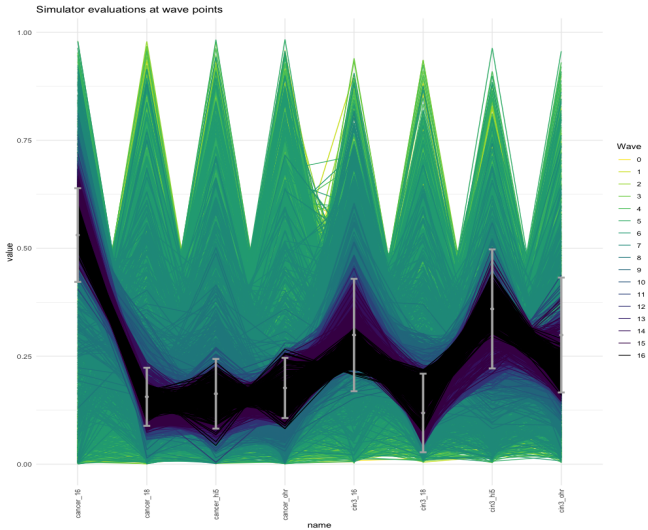
# Emulating HPVsim

- Sensitivity analysis on non-varying parameters to estimate internal model discrepancy; simulation studies on population size to motivate external model discrepancy
- 16 waves of emulation performed
- All waves use hierarchical emulation; final wave emulates covariances between outputs
- 16 repetitions per parameter set, 330 parameter sets per wave
- Final non-implausible region used to simulate future cancer cases with higher repetition number

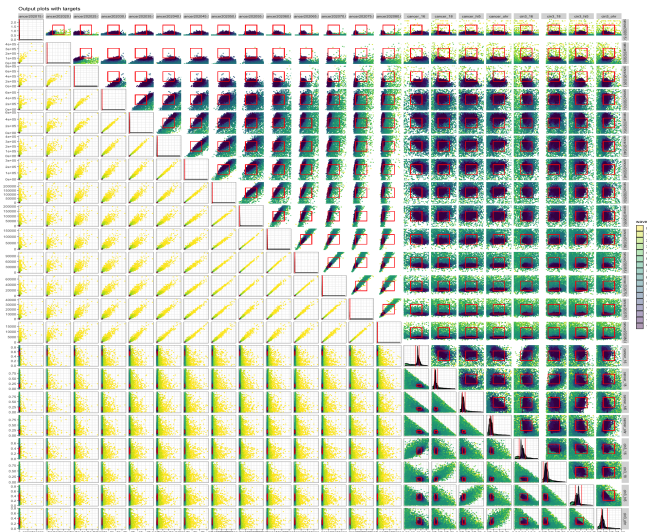
# Results of Emulation



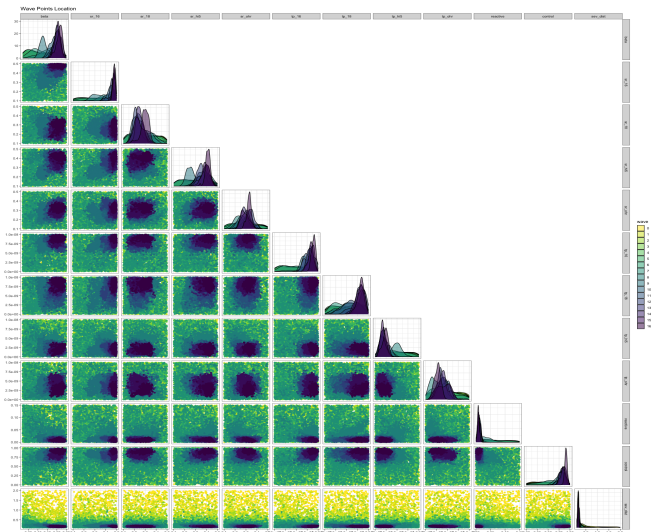
# Results of Emulation



# Results of Emulation



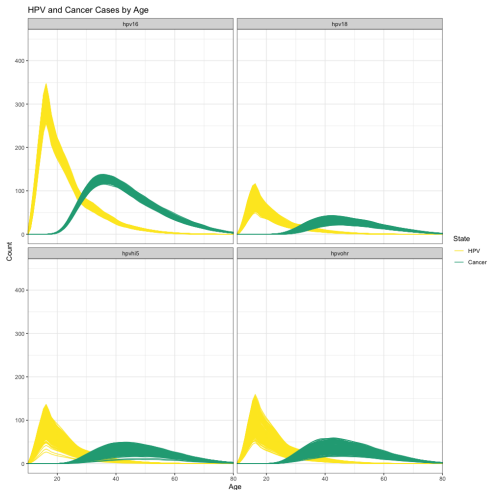
# Results of Emulation



# Analysis of the Non-Implausible Region

- Final non-implausible region:  $5 \times 10^{-17}$  of original parameter space
- Over all waves, 800 points proposed consistent with observational data and uncertainty ('yield'  $\sim 10\%$ )
- Last wave emulators' proposal: yield over 50%
- Non-identifiability of parameters encapsulated by correlations within final non-implausible region.

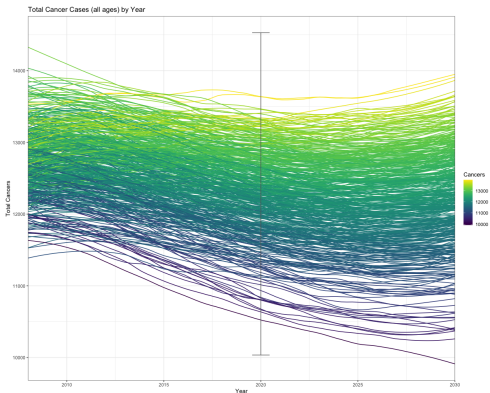
# HPV Demographics



Genotype peaks for HPV broadly similar, suggesting vaccination might be optimal around ages 16-17. Cancer most aggressive for HPV16: average 21 years from HPV acquisition vs ~ 30 years for other genotypes. Higher uncertainty in 'hi5' and 'ohr' genotypes.



# Future Cancer Cases



High variability in possible cancer cases in 2030 (no intervention).  
Consequence of limited historical data and large uncertainty in observations.  
Understanding of the variability is crucial for meaningful statements about predicting the efficacy of intervention strategies via modelling.

# Emulating HPVsim

HPVsim is a complex, high-dimensional, moderately expensive simulator of HPV and cervical cancer, with high sensitivity to stochastic effects.

Hierarchical emulation allows us to accurately quantify all sources of uncertainty and find a *complete* parameter space consistent with observational data.

The resulting space can be used to determine latent properties of the disease progression and aid in prediction for future scenario analysis/intervention modelling.

# Open Questions and Future Research

- Future data gathering: what observational data will be most effective in inferring disease properties?
- Decision support for intervention: propagating uncertainty in simulator predictions to ensure robust analysis and decision making.
- Similarities between countries: if we repeat the analysis for a different country, how similar are the results? Are there obvious geographical/demographic trends?

# Selected References I

- [1] Ioannis Andrianakis et al. “Bayesian History Matching of Complex Infectious Disease Models Using Emulation: A Tutorial and a Case Study on HIV in Uganda”. In: *PLoS computational biology* 11.1 (2015), e1003968.
- [2] Peter S Craig et al. “Constructing Partial Prior Specifications for Models of Complex Physical Systems”. In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 47.1 (1998), pp. 37–53.
- [3] Jonathan A Cumming and Michael Goldstein. “Bayes Linear Uncertainty Analysis for Oil Reservoirs Based on Multiscale Computer Experiments”. In: *The Oxford handbook of applied Bayesian analysis* (2010), pp. 241–270.
- [4] Baishan Hu et al. “Ab Initio Predictions Link the Neutron Skin and 208Pb to Nuclear Forces”. In: *Nature Physics* 18.10 (2022), pp. 1196–1200.
- [5] Andrew Iskauskas et al. “Emulation and History Matching using the hmer Package”. In: *arXiv preprint arXiv:2209.05265* (2022).
- [6] Cliff C Kerr et al. “Covasim: An Agent-Based Model of COVID-19 Dynamics and Interventions”. In: *PLOS Computational Biology* 17.7 (2021).
- [7] Danny Scarponi et al. “Demonstrating Multi-country Calibration of a Tuberculosis Model Using New History Matching and Emulation Package - hmer”. In: *Epidemics* 43 (2023), p. 100678.

## Selected References II

- [8] Robyn M Stuart et al. “HPVsim: An Agent-Based Model of HPV Transmission and Cervical Disease”. In: *medRxiv* (2023).
- [9] Ian Vernon, Michael Goldstein, and Richard Bower. “Galaxy Formation: A Bayesian Uncertainty Analysis”. In: *Bayesian analysis* 5.4 (2010), pp. 619–669.
- [10] Ian Vernon et al. “Bayesian Uncertainty Analysis for Complex Systems Biology Models: Emulation, Global Parameter Searches and Evaluation of Gene Functions”. In: *BMC systems biology* 12.1 (2018), pp. 1–29.
- [11] Daniel Williamson et al. “History Matching for Exploring and Reducing Climate Model Parameter Space using Observations and a Large Perturbed Physics Ensemble”. In: *Climate dynamics* 41.7 (2013), pp. 1703–1729.