XLVI CILAMCE
VITÓRIA - ES - BRAZIL - 2025
XLVI IBERO-LATIN AMERICAN CONGRESS ON COMPUTATIONAL METHODS IN ENGENEERING

# Ensemble methodology incorporating uncertainty quantification applied to modeling COVID-19 outbreaks

Douglas S. de Albuquerque[1], Renato S. Silva[1], Gilson A. Giraldi[1], Regina C. C. de Almeida[1]

[1]*Laboratório Nacional de Computação Científica, LNCC*
*Av. Getúlio Vargas - 333, CEP 25651-075, Petrópolis/RJ, Brasil*
*dougalbu@posgrad.lncc.br, rssr@lncc.br, gilson@lncc.br, rcca@lncc.br*

**Abstract.** Individual models often fail to capture the full dynamics of complex systems. In this paper, we propose an ensemble built from four distinct models calibrated with real COVID-19 data and with quantified uncertainty. Predictions were combined using a weighted average based on Root Mean Square Deviation (RMSD). By incorporating the uncertainty of each model, the ensemble produced robust estimates, highlighting its potential as a reliable forecasting approach for complex scenarios.

**Keywords:** Ensemble method, SIR Model, PINN, Uncertainty Quantification, Covid-19

## 1 Introduction

Mathematical models are essential for understanding physical phenomena and developing simulations to support decision-making. However, the observed phenomenon may be complex in nature, making its dynamics difficult to capture fully with a single model. In such cases, it is appropriate to apply a set of models capable of representing different aspects of the problem, identifying which ones perform better, and strategically combining their outputs. This approach describes ensemble methods [1], which provide a flexible framework that combines diverse perspectives, thereby leading to more robust and informative results [2–4] .

When mathematical models are calibrated with real-world data, uncertainty is an inherent characteristic that arises due to the variability or limitations of observations. Quantifying this uncertainty is essential, both for assessing the reliability of parameter estimates and for better understanding the confidence and robustness of model-based predictions. In this context, uncertainty also plays an essential function in evaluating the consistency and complementarity of models within an ensemble.

The COVID-19 pandemic highlighted the challenges of modeling the transmission of infectious diseases, especially due to behavioral practices, variable transmission rates, and data limitations [4]. These factors make it difficult for a single model to capture the full complexity of the outbreak dynamics consistently. In this work, we apply ensemble methods, complemented by uncertainty quantification, as a strategy to address these limitations.

## 2 Ensemble Methods

Ensemble methods aim to improve predictive capability by strategically combining multiple base models, each contributing according to its individual performance. Rather than selecting a single best model, this approach leverages the complementary strengths of diverse plausible models, preserving distinct dynamics and aggregating broader information about the underlying phenomenon [2]. The ensemble integrates models addressing the same problem to produce a unified, robust prediction [1].

For regression tasks, weighted averaging provides a fundamental combination mechanism, assigning importance to each model based on individual performance [1]. Consider a set of base models $T = \{h_1, \ldots, h_T\}$, where each $h_i(\boldsymbol{x})$ generates predictions for the sample $\boldsymbol{x}$, for all $1 \leq i \leq T$. Let $\boldsymbol{w} = [w_1, \ldots, w_T]$ denote the corresponding weight array. The combined output $H(\boldsymbol{x})$ is defined as:

$$H(\boldsymbol{x}) = \sum_{i=1}^{T} w_i h_i(\boldsymbol{x}); \tag{1}$$

subject to the constraints:

$$w_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{T} w_i = 1.$$

The performance of individual models is quantified by the Root Mean Square Deviation (RMSD), with lower values indicating higher predictive accuracy. Let $y_k$ for $k = 1, \ldots, n$ denote the experimental data and $n$ the number of samples; the RMSD for model $h_i$ is defined as:

$$RMSD_i = \sqrt{\frac{1}{n}\sum_{k=1}^{n} (y_k - h_i(\boldsymbol{x}_k))^2}. \tag{2}$$

Model weights are then computed based on the inverse of their RMSD (eq. (2)):

$$w_i = \frac{(RMSD_i)^{-1}}{\sum_{j=1}^{T} (RMSD_j)^{-1}}. \tag{3}$$

## 3 Models

This work employs an ensemble composed of four distinct models: a mechanistic SIR model capturing epidemiological dynamics; a phenomenological Generalized Richards Model (GRM), which empirically reproduces observed patterns without assuming underlying mechanisms; the purely data-driven Holt model; and a Physics-Informed Neural Network (PINN), a machine learning model incorporating physical constraints into the learning process. This heterogeneous ensemble leverages complementary modeling paradigms, each of which is briefly described below.

### 3.1 SIR Model

The SIR model categorizes the population into three different compartments [5]: Susceptible ($\boldsymbol{S}$), healthy individuals who are susceptible to infection upon contact; Infected ($\boldsymbol{I}$) individuals who are actively infectious and capable of transmitting the disease; and Recovered ($\boldsymbol{R}$) individuals who have recovered and acquired immunity. This model operates under the fundamental assumption that individuals remain infectious from the moment of infection until acquiring permanent immunity. All individuals belong to the same homogeneously mixed population without physical segregation, which enables interactions among them and transitions between compartments.

The SIR model is defined by the system of ordinary differential equations presented in eq. (4). The parameter $\beta$ denotes the contact rate, quantifying the daily probability of susceptible individuals becoming infected through effective contact. The recovery rate, $\gamma$, is the reciprocal of the average infectious period (in days), indicating the required time for infected individuals to recover and acquire immunity. The subpopulations are time-dependent functions whose sum equals the total population $N$ (eq. (5)), which is assumed to remain constant over time $t$ [5]:

$$\frac{dS}{dt} = -\beta\, S\, I; \quad \frac{dI}{dt} = \beta\, S\, I - \gamma\, I; \quad \frac{dR}{dt} = \gamma\, I. \tag{4}$$

$$S(t) + I(t) + R(t) = N. \tag{5}$$

The system requires initial conditions $S(0) = S_0$, $I(0) = I_0$, and $R(0) = R_0$. Typically, $I_0$ is obtained from early epidemic data, $R_0$ is assumed to be zero, and $S_0 = N - I_0 - R_0$.

### 3.2 Generalized Richards Model

The Generalized Richards Model (GRM) effectively simulates epidemic cumulative curves by capturing both early subexponential growth and late-stage saturation through its parameter structure [2, 6]. The following ordinary differential equation defines GRM:

$$\frac{dC(t)}{dt} = r\,(C(t))^q \left(1 - \left(\frac{C(t)}{K}\right)^{\alpha}\right), \tag{6}$$

where $C(t)$ denotes the cumulative number of cases at time $t$. The parameter $r$ represents the growth rate under unrestricted transmission conditions, reproducing the initial exponential phase of the epidemic curve. The parameter

$q$ controls the deceleration of growth, enabling the model to represent constant ($q = 0$), exponential ($q = 1$), and subexponential ($0 < q < 1$) growth regimes. The parameter $K$ is defined as the carrying capacity, corresponding to the final size of the epidemic or plateau level. Finally, the parameter $\alpha$ quantifies the deviation from the S-shaped dynamics, controlling how rapidly the curve approaches the plateau [2, 6].

### 3.3 Holt Model

Holt's linear exponential smoothing is a widely used forecasting method for modeling time series data that exhibits a trend [7]. This method applies double exponential smoothing to estimate both the level and the trend components, using two distinct smoothing parameters, $\lambda_1$ and $\lambda_2$, which control the rate of adaptation to evolutions in each component. Both the parameters take values in $[0, 1]$: $\lambda_1$ governs the smoothing of the observed data, while $\lambda_2$ regulates the smoothing of the trend component. The model is updated recursively at each time step $t$ through the smoothed level $\tilde{y}_t$, and the estimated trend $F_t$, defined by the equations eq. (7) and eq. (8):

$$\tilde{y}_t = \lambda_1 y_t + (1 - \lambda_1)(\tilde{y}_{t-1} + F_{t-1}), \tag{7}$$

$$F_t = \lambda_2(\tilde{y}_t - \tilde{y}_{t-1}) + (1 - \lambda_2)F_{t-1}. \tag{8}$$

The forecast for $h$ days ahead is computed using eq. (9), defined below:

$$\hat{y}_{t+h} = \tilde{y}_t + hF_t. \tag{9}$$

### 3.4 Physics-Informed Neural Network

Artificial Neural Networks (ANN) are machine learning models capable of learning and replicating complex patterns from observational data. The ANN maps an input dataset $\boldsymbol{X}$ to an output $\boldsymbol{y}$ through a learning task, applying an optimal search that fits the parameters of the model, minimizing a loss function [8], typically the Mean Squared Error (MSE):

$$MSE = \frac{1}{n}\sum_{k=1}^{n}(y_k - \hat{y}_k)^2. \tag{10}$$

where, $\hat{y}$ is the predicted values, $y$ is the ground truth, and $n$ is the sample size.

The ANN architecture, defined by layer and neuron configuration, processes information through successive transformations. Each layer applies a nonlinear activation function $\phi$, enabling the ANN to capture and model complex relationships.

Physics-Informed Neural Network (PINN) provides a significant advance in conventional ANN by incorporating *a priori* physical knowledge through governing differential equations that represent the physics of the problem [8]. This integration acts as a regularizer agent, constraining the search space for admissible solutions and improving generalization in data-scarce regimes. PINN minimizes the loss function by calculating the sum of the error from the data and the residual of the equations of differential equations (eq. (15)). The residual of the SIR Model (eq. (4)), adopting the MSE, is defined below:

$$MSE_S = \frac{1}{n}\sum_{k=1}^{n}\left(\frac{dS(t_k)}{dt_k} + \beta\, S(t_k)I(t_k)\right)^2, \tag{11}$$

$$MSE_I = \frac{1}{n}\sum_{k=1}^{n}\left(\frac{dI(t_k)}{dt_k} - \beta\, S(t_k)I(t_k) + \gamma\, I(t_k)\right)^2, \tag{12}$$

$$MSE_R = \frac{1}{n}\sum_{k=1}^{n}\left(\frac{dR(t_k)}{dt_k} - \gamma\, I(t_k)\right)^2, \tag{13}$$

$$\tag{14}$$

with the loss function:

$$\boldsymbol{L} = MSE_{Data} + MSE_{SIR}. \tag{15}$$

Where, $MSE_{Data}$ refers to eq. (10) and $MSE_{SIR}$ is the sum of eq. (11), eq. (12) and eq. (13).

# 4 Computational Experiments

The experimental dataset consists of daily reported COVID-19 infection cases in the city of Recife, Brazil, obtained from a public repository, and comprises data up to May 26, 2020 [6]. To address high variability in the raw observations, cubic spline smoothing was applied to the calibration data to reduce noise. Both raw and smoothed curves are presented in Fig. 1a. From the smoothed daily data, cumulative case counts were computed to support GRM calibration (Fig. 1b). The full dataset spans 79 days, of which the first 60 were used for model calibration, while the remaining data were reserved for out-of-sample forecasting evaluation. A dashed vertical line in Fig. 1a and Fig. 1b marks this division. The data for calibration and forecasting subsets were normalized by their own maximum value, constraining the scale to $[0, 1]$.




(a) Daily infection data

(b) Cumulative infection data

Figure 1. Experimental data

## 4.1 Model Configurations

Both the SIR and GRM models were calibrated using the PyMC library in Python, applying Markov Chain Monte Carlo (MCMC) with the Differential Evolution Metropolis algorithm [9] running four parallel chains. Uniform priors were assigned to all parameters, and a Gaussian likelihood was adopted, with its standard deviation treated as a calibration hyperparameter. Convergence was evaluated via the Gelman–Rubin diagnostic, confirming satisfactory results for all parameters. After convergence, the samples at the maximum *a posteriori* (MAP) value were selected as the optimal parameter estimates.

For the SIR model, the parameters $\beta$ and $\gamma$ followed uniform priors in the range $\mathcal{U}[0, 1]$. Each MCMC chain generated $50000$ iterations, with the first $15\%$ discarded as burn-in. For the GRM model, priors were defined as $\mathcal{U}[0, 2]$ for $r$ and $K$, and $\mathcal{U}[0, 1]$ for $q$ and $\alpha$. Each chain generated $250000$ samples, with $40\%$ discarded as burn-in.

The parameters of the Holt model, $\lambda_1$ and $\lambda_2$, were calibrated via a two-stage grid search. The first stage performed a wide search over the closed interval $[0.1, 1]$ with step size $0.1$. For each candidate parameter pair, the model was fitted and evaluated against the experimental data using RMSD. The pair yielding the lowest RMSD was selected for refinement. In the second stage, a refined grid search was conducted over 50 equally spaced points within a $\pm 25\%$ range around the initial optimum. This fine search is terminated early if any candidate pair achieves an RMSD value lower than $0.005$, in which case those parameters are adopted as optimal.

For the training of PINN, the experimental dataset was randomly partitioned into $80\%$ for training and $20\%$ for testing. The network architecture consisted of one input node (time $t$), three hidden layers with 64 neurons each, and three output nodes (SIR Model compartments). ReLU activations were employed in all hidden layers, with a Sigmoid activation at the output layer. A dropout rate of $10\%$ was applied to the first and second hidden layers. The computational work was developed using the PyTorch library in Python [10].

Since information on the susceptible and recovered populations was unavailable, only the infected population data were used to compute the data loss. The PINN estimates the time evolution of the three state variables following the SIR formulation; however, the dynamics of S and R were incorporated exclusively for the calculation of the differential equation loss. To ensure physical consistency, a penalty term enforcing the conservation of total population was incorporated, denoted as $MSE_N$ (eq. (5)) and another term for initial conditions was incorporated, denoted as $MSE_{IC}$. The total loss function was thus defined as the sum of four components: $MSE_N$, $MSE_{IC}$, $MSE_I$ (eq. (12)), and $MSE_{\text{Data}}$. Additionally, the PINN was designed to estimate unknown model parameters. The SIR parameters were included as trainable variables, initialized at $\beta_0 = \gamma_0 = 0.01$.

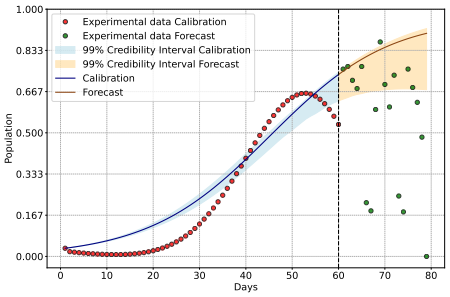
The network was trained for $25000$ epochs using the Adam optimizer with a learning rate of $0.025$. Model

performance was evaluated using Root Mean Squared Deviation (RMSD), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE), yielding values of $0.0132$, $0.0092$, and $0.17\%$, respectively.

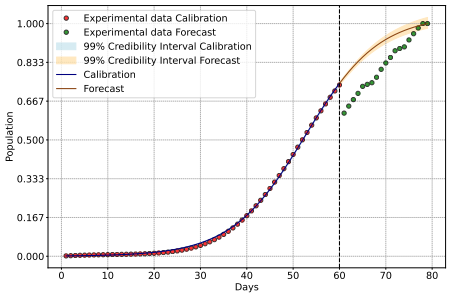
Uncertainty quantification of the solutions was computed individually for each model. For the SIR and GRM models, uncertainty arises naturally from their Bayesian calibration [11]. In both cases, samples were drawn from the estimated posterior distribution. These samples represent parameter sets sampled jointly at each iteration, preserving the dependency structure among parameters. Thus, each sample was used to simulate the corresponding model. However, for the Holt and PINN models, uncertainty had to be introduced explicitly. In the Holt model, uncertainty was estimated using a jackknife resampling applied exclusively to the calibration data. New datasets were generated by randomly omitting $10$ points from this subset. Each resampled dataset was recalibrated, and the resulting optimal parameters were used to solve the model and generate forecasts. In the PINN model, uncertainty was estimated using Monte Carlo Dropout [12]. During inference, dropout remained active, and multiple forward passes were performed to generate a predictive distribution. For each model, regardless of how uncertainty was introduced, the corresponding simulations were used to compute prediction intervals. At each time point, the 0.5th and 99.5th percentiles of these simulations defined the lower and upper bounds of the uncertainty boundaries.

### 4.2 Results

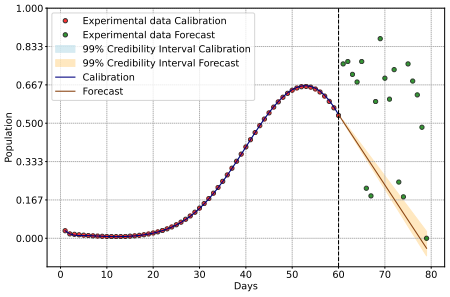
The calibrated parameters were used to simulate each model for both data fitting and forecasting. Figure 2 displays the fitted curves, along with their corresponding uncertainty intervals computed at the $99\%$ credibility level. Among the models, Holt and GRM exhibited the narrowest uncertainty bands during calibration. In particular, the GRM (Fig. 2b) was calibrated using the cumulative case counts. The Holt model performs poorly in forecasting, as it predicts a downward trend inconsistent with the raw data. It even reaches negative values at the end of the trajectory, which is physically unrealistic. In contrast, the SIR model, despite its relatively poor calibration performance, successfully captures the overall dynamics during forecasting. The PINN yields a forecasting behavior similar to that of the SIR model.



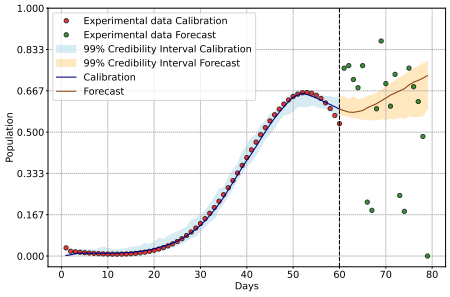

(a) SIR Model ( RMSD: 0.0781, Weight: 0.0177 )




(b) GRM ( RMSD: 0.0184, Weight: 0.0750 )




(c) Holt Model ( RMSD: 0.0017, Weight: 0.8007 )




(d) PINN ( RMSD: 0.0130, Weight: 0.1066 )

Figure 2. Model fitting results with uncertainty quantification.

Ensemble construction required GRM conversion to daily incidence. Holt and PINN models achieved the best fit over the initial phase, dominating the ensemble composition. Model forecasts were combined according to eq. (3), forming the ensemble. As shown in Figure 3, the ensemble demonstrates satisfactory performance fitting initial data. The ensemble reproduces the overall forecasting pattern of the individual models. The remaining models mitigated the influence of Holt, yielding a physically consistent result.

Ensemble uncertainty was quantified via kernel density estimation (KDE) based on the model outputs at

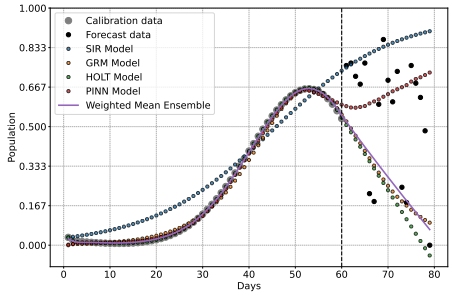
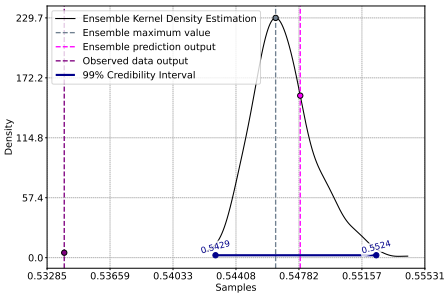


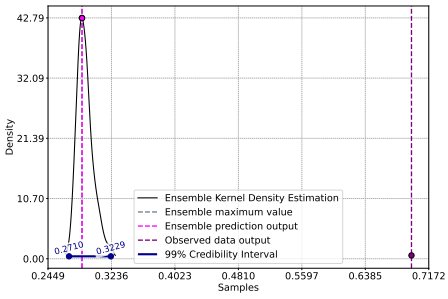

Figure 3. Forecast ensemble and the fitted solutions from individual calibrated models

specific time points of interest: the final day of the calibration period ($t = 60$) and two points within the forecast horizon ($t = 70$ and $t = 79$). For each model, pointwise predictions corresponding to these times were extracted from the simulations, considering the uncertainty. These outputs were then used to estimate individual KDEs for each model. Subsequently, the ensemble KDE was derived by weighting these densities according to the previously computed model weights.

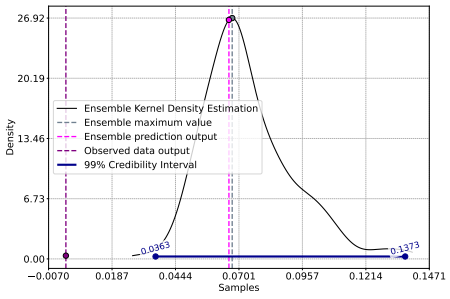
Figure 4 presents the ensemble KDEs at selected time points, with a 99% credibility interval, defined by the 0.5th and 99.5th percentiles. Among all cases, the estimate at $t = 60$ (Figure 4a) exhibits the lowest dispersion and is the only one in which the observed value lies within the credible interval of the ensemble. Moreover, the predicted output at this time is also notably close to the true value. In contrast, the forecasts at $t = 70$ and $t = 79$ failed to capture values near the observed data, which is consistent with previous observations regarding forecast performance. The SIR model exhibits the highest uncertainty propagation, which contributed to the predicted value by the ensemble not coinciding with the maximum of the KDE. Nevertheless, quantifying uncertainty adds significant value to the ensemble by enabling a more informative and robust forecast structure, as it aggregates the variability inherent to each prediction.




(a) Day 60




(b) Day 70




(c) Day 79

Figure 4. Kernel density estimation (KDE) of the ensemble solutions

## 5 Conclusions

This paper introduced an ensemble forecasting framework combining four modeling paradigms calibrated on COVID-19 case data. Due to the high noise and lack of clear trends in the original data, smoothing was applied to enable calibration, though at the cost of some information loss. Individual uncertainties were explicitly quantified and incorporated into the ensemble, enhancing interpretability and robustness. Using KDE to quantify ensemble uncertainty provides a robust estimate of the prediction distribution, enabling pointwise probabilistic interpretations and capturing the variability in model outputs. The ensemble was constructed based on the fit error of each model across the entire time series, a procedure that naturally favored the Holt model due to its superior calibration accuracy. However, the results indicated that the SIR and PINN models better captured the underlying data trends in the forecast. This highlights the importance of incorporating additional criteria that assess model adequacy when constructing the ensemble and will inform our future work. Incorporating models from different paradigms within the Ensemble not only improves forecast reliability but also provides valuable understanding into the complementary strengths of each modeling approach.

**Acknowledgements.** This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – FINANCING CODE 001. R.C.C.A. acknowledges CNPq grant 306588/2022-6.

**Authorship statement.** The authors hereby confirm that they are the sole liable persons responsible for the authorship of this work, and that all material that has been herein included as part of the present paper is either the property (and authorship) of the authors, or has the permission of the owners to be included here.

## References

[1] Z.-H. Zhou. *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall/CRC, 1st edition, 2012.

[2] G. Chowell and R. Luo. Ensemble bootstrap methodology for forecasting dynamic growth processes using differential equations: application to epidemic outbreaks. *BMC Med Res Methodol*, vol. 21, 2021.

[3] D. G. F. Borges, E. R. Coutinho, T. Cerqueira-Silva, M. Grave, A. O. Vasconcelos, L. Landau, A. L. G. A. Coutinho, P. I. P. Ramos, M. Barral-Netto, S. T. R. Pinho, M. E. Barreto, and R. F. S. Andrade. Combining machine learning and dynamic system techniques to early detection of respiratory outbreaks in routinely collected primary healthcare records. *BMC Medical Research Methodology*, vol. 25, n. 1, 2025.

[4] P. Melin, J. C. Monica, D. Sanchez, and O. Castillo. Multiple ensemble neural network models with fuzzy response aggregation for predicting covid-19 time series: The case of mexico. *Healthcare (Basel)*, vol. 8, n. 2, 2020.

[5] M. J. Keeling and P. Rohani. *Modeling infectious diseases in humans and animals*. Princeton University Press, New Jersey, 2011.

[6] G. L. Vasconcelos, G. C. G. C. Duarte-Filho, A. A. Brum, R. Ospina, F. A. G. Almeida, and A. M. S. Macêdo. Análise de curvas epidêmicas da covid-19 via modelos generalizados de crescimento: Estudo de caso para as cidades de recife e teresina. *SciELO Preprints*, 2020.

[7] S. Gelper, R. Fried, and C. Croux. Robust forecasting with exponential and holt–winters smoothing. *Journal of Forecasting*, vol. 29, n. 3, pp. 285–300, 2010.

[8] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, vol. 378, pp. 686–707, 2019.

[9] PyMC. *PyMC — Probabilistic Programming in Python*. PyMC Team. Accessed: 2025-07-22, `https://www.pymc.io`, 2025.

[10] PyTorch. *PyTorch — Deep Learning Framework*. Pytorch Team. Accessed: 2025-07-22, `https://pytorch.org`, 2025.

[11] R. C. Smith. *Uncertainty Quantification: Theory, Implementation, and Applications*. Society for Industrial and Applied Mathematics, USA, 2013.

[12] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pp. 1050–1059. JMLR.org, 2016.