Forward Mortality Modeling with Regional Shrinkage: A Stochastic Framework for Long-Term Longevity Risk

Xinyi Bao
Pennylvinia State University:
University Park
State College, United States
xzb5043@psu.edu

***Abstract*—This paper develops a forward mortality modeling framework that captures the stochastic evolution of projected mortality rates across multiple regions. The model decomposes regional log-mortality surfaces into low-dimensional latent temporal factors using principal component analysis, followed by stochastic differential equation modeling. The temporal dynamics are governed by stochastic differential equations, allowing for the simulation of mortality trajectories under uncertainty.**

***The framework is implemented using monthly region-level mortality data spanning several countries and time periods, including pandemic-affected years. Principal component analysis is applied to extract dominant sources of temporal variation, while an Ornstein–Uhlenbeck process is used to simulate future mortality dynamics. Forecast accuracy is evaluated through out-of-sample testing and visual comparisons.**

The proposed model offers a transparent and computationally tractable approach to regional mortality forecasting, capable of reflecting both persistent demographic trends and transient shocks. It provides a flexible foundation for long-term mortality risk assessment and can be extended to incorporate additional sources of heterogeneity and structural information.

***Keywords*—Forward Mortality Models, Longevity Risk, Stochastic Differential Equations, Principal Component Analysis, Mortality Forecasts**

I. Introduction

Aggregate mortality risk refers to the uncertainty that future mortality developments may deviate from current expectations due to structural demographic shifts or exogenous shocks—remains a central concern in actuarial science, with profound implications for life insurance, pension systems, and longevity-contingent financial products. Classical mortality models, particularly the Lee–Carter framework [5], have long served as the foundation for projecting mortality rates. However, their retrospective nature limits their capacity to capture the full extent of forward-looking uncertainty, particularly in the presence of regime changes or extraordinary events such as the COVID-19 pandemic.

Recognizing these limitations, recent research has shifted toward modeling the evolution of mortality forecasts rather than realized death rates. Zhu and Bauer [3], for instance, introduced a forward mortality modeling framework that parallels forward interest rate models in finance. By directly modeling the dynamics of mortality expectations over time, their approach offers a more flexible and theoretically coherent representation of long-term demographic risk.

Building on this perspective, the present study proposes a forward mortality modeling framework that integrates principal component analysis (PCA) with stochastic differential equations (SDEs) to capture the latent drivers of regional mortality evolution. This structure enables the decomposition of high-dimensional mortality surfaces into low-rank temporal factors and corresponding spatial loadings, thereby facilitating both interpretability and dimensionality reduction. By focusing on monthly region-level mortality data, the model is designed to capture both short-term disruptions and long-term demographic shifts.

The methodological implementation emphasizes computational simplicity and empirical transparency. Instead of relying on age-specific stratification or complex shrinkage mechanisms, the model operates directly on log-transformed all-cause mortality data. PCA is employed to extract dominant temporal components from the mortality surface, which are then simulated into the future using an Ornstein–Uhlenbeck process—a commonly used mean-reverting SDE. This structure allows for both stochastic forecasting and empirical comparison across model specifications.

The empirical study utilizes the Local Mortality Dataset, comprising monthly mortality records for major regions and cities across multiple countries from 2015 to 2022. The data are preprocessed and transformed to the logarithmic scale to stabilize variance and support continuous modeling assumptions. Forecast performance is evaluated based on out-of-sample prediction error, with particular focus on the one- and two-principal-component variants of the PCA-SDE model.

This study seeks to address several key questions relevant to forward-looking mortality modeling in a regional context. Specifically, it investigates how principal component-based factor decomposition, when combined with stochastic differential equations, can effectively capture both persistent and transient drivers of mortality variation across time and geography. It also examines whether the inclusion of additional components improves forecast accuracy in regions characterized by elevated volatility. Furthermore, the framework is evaluated for its capacity to deliver interpretable, low-dimensional representations of complex mortality surfaces while maintaining predictive robustness. By doing so, the study aims to demonstrate the advantages of forward mortality modeling over conventional retrospective approaches, particularly under conditions of demographic instability and regional heterogeneity.

Section 2 reviews related literature on mortality modeling. Section 3 introduces the methodology, including PCA decomposition and stochastic simulation. Section 4 introduces dataset, preprocessing steps, and model implementation.

Section 5 presents the empirical results, corresponding visualizations, and discussions. Section 6 makes the conclusion and points out three possible approaches for future research.

II. Related Work

This study builds upon three interrelated areas of research in mortality forecasting: classical statistical models, forward-looking stochastic methodologies, and structural regularization techniques tailored for high-dimensional or sparse data environments.

The Lee–Carter model [5] remains a cornerstone of age-period mortality forecasting. Its decomposition of log-mortality into an age-specific component and a single time-varying index provides a parsimonious and interpretable framework for demographic projection. However, the model's reliance on a univariate time series for temporal evolution, combined with the implicit assumption of homogeneous dynamics across all age groups, limits its ability to accommodate structural breaks or quantify long-term uncertainty. Numerous empirical studies have documented the tendency of Lee–Carter-based forecasts to underestimate out-of-sample volatility, particularly under conditions of demographic disruption. Extensions such as the Cairns–Blake–Dowd model and cohort extensions attempt to address these issues, but often increase complexity and reduce tractability without significantly improving performance in volatile or spatially disaggregated settings.

To address these shortcomings, Zhu and Bauer [3] introduce a forward mortality modeling framework that directly characterizes the stochastic behavior of projected mortality outcomes rather than fitting retrospective death rates. Their formulation treats projected survival probabilities as primary state variables and employs principal component analysis (PCA) to extract latent mortality factors, which are subsequently modeled using stochastic differential equations (SDEs). This approach offers a mathematically coherent representation of the dynamic evolution of mortality expectations and aligns with the valuation and risk management frameworks commonly used in insurance and pension systems. A key advantage lies in the shift from realized to expected mortality surfaces, allowing forecasts to better reflect uncertainty surrounding future demographic conditions.

The present study adopts the PCA–SDE structure developed by Zhu and Bauer as its modeling foundation. Particular emphasis is placed on the forward formulation's internal consistency, as ensured by its martingale properties, and its suitability for long-horizon forecasting. This structure is further extended to incorporate spatial heterogeneity by modeling mortality across multiple regional units. In doing so, it accommodates geographic variation in mortality risk while maintaining a low-dimensional representation of its temporal dynamics. The framework also facilitates empirical comparison of modeling structures with different numbers of principal components, enabling a modular investigation of complexity versus accuracy tradeoffs.

To improve model robustness in the presence of sparse or noisy regional mortality data, spatial regularization is incorporated through geography-based shrinkage penalties. These smoothing constraints are motivated by the graph Laplacian regularization framework proposed by Li, Li, and Panagiotelis [2], which has demonstrated empirical success in subnational mortality forecasting. While their methodology is situated within a gradient boosting context, the underlying principle—embedding domain knowledge into the model structure—is equally applicable here and plays a critical role in stabilizing inference across diverse regions. Regularization also improves generalizability by constraining parameter estimates in low-population regions, thereby reducing sensitivity to anomalous reporting or short-term fluctuations.

Parallel developments in artificial intelligence and machine learning have also informed this study's emphasis on interpretability and modularity. Richman [7] provides a comprehensive overview of AI applications in actuarial modeling, highlighting techniques such as autoencoders, neural networks, and ensemble learning. Although this work does not employ deep learning directly, it shares with these approaches a focus on transparent model design and the integration of Bayesian uncertainty quantification, consistent with current best practices in actuarial analytics. The balance between interpretability and complexity remains a core design consideration in this research, especially as future extensions may incorporate nonlinear modeling or spatial graph learning frameworks.

III. Methodology

This study develops a forward mortality forecasting framework that integrates principal component analysis (PCA) with stochastic differential equations (SDEs). The methodology consists of three key stages: dimensionality reduction via PCA, simulation of temporal mortality risk factors using the Ornstein–Uhlenbeck (OU) process, and mortality surface reconstruction through matrix re-composition. This framework captures both the shared temporal structure across geographic units and the inherent uncertainty in future mortality trajectories.

A. Principal Component Deposition of Mortality Surface

To reduce the dimensionality of the problem and extract dominant sources of variation, principal component analysis (PCA) is applied to a matrix of log-transformed mortality values. The input data matrix, denoted by $M \in \mathbb{R}^{T\times R}$, is structured such that each row corresponds to a month and each column corresponds to a region, with the entries representing the log of mortality counts (i.e., $\log(\text{deaths}+1)$).

To reduce the dimensionality of the regional mortality data and extract dominant sources of variation, PCA is applied to a matrix of log-transformed death counts. Let $M \in \mathbb{R}^{T\times R}$ denote the log-mortality surface, where T represents the number of months and R the number of regions. Each entry $\mathrm{M}_{\mathrm{x,t}} = \log\left(\mathrm{deaths}_{\mathrm{x,t}} + 1\right)$ reflects the smoothed log-death count for region x in month t. The matrix is decomposed into orthogonal temporal and spatial components as follows:

$$\log M_{x,t} \approx \sum_{i=1}^{k} \beta_i(t) \cdot \phi_i(x) \quad (1)$$

Here, $\beta_i(t)$ denotes the score of the i-th principal component at time t. And

$\phi_i(x)$ represents the spatial loading vector describing each region's sensitivity to the latent factor. Together, these terms describe how each region responds over time to latent mortality trends shared across all locations.

This decomposition enables the identification of major temporal patterns in the data using only one or two components (i.e., $k = 1$ or 2). The first component typically captures global mortality shifts, such as pandemic-related shocks, while the second component may reflect more localized effects or seasonal variation. The resulting reduction in dimensionality enhances the interpretability and computational efficiency of subsequent modeling stages.

B. Forecasting with Stochastic Differential Equations

The latent temporal factors $\beta_i(t)$, extracted through PCA, are modeled using stochastic differential equations to simulate their future evolution. Each time series is assumed to follow an Ornstein–Uhlenbeck (OU) process, expressed as:

$$d\beta_t = \theta(\mu - \beta_t)\, dt + \sigma\, dW_t \tag{2}$$

In this expression, θ means the speed of mean reversion over specific time intervals, μ refers the long-run mean of the period, σ refers the volatility of the whole process, and $dW_t$is a standard Brownian increment. The OU simulation is implemented using the Euler–Maruyama discretization in R. Parameters μ and σ are estimated from the training-period $\beta_i(t)$, series.

The OU process is well-suited for mortality forecasting as it allows for both persistence and variability in mortality risk over time, while assuming eventual reversion toward a stable long-run trend. Simulation is carried out using the Euler–Maruyama discretization method. For each retained principal component, the simulation begins at the final observed value in the training period and generates a sequence of future values over the desired forecasting horizon.

The parameters θ, μ, and σ are estimated directly from the historical $\beta_i(t)$ series. Each component is simulated independently, reflecting the orthogonality imposed by PCA.

C. Mortality Surface Reconstruction

Once the simulated future values of $\beta_i(t)$ are obtained, the projected mortality surface is reconstructed by combining them with the corresponding spatial loadings $\phi_i(x)$. The forecasted log-mortality rate for region x at time t is given by:

$$\widehat{M_{x,t}} = \exp\left(\sum_{i=1}^{k} \beta_i(t) \cdot \phi_i(x)\right) \tag{3}$$

This reconstruction step involves computing the outer product between each simulated temporal trajectory and its spatial loadings, summing across components, and then exponentiating to recover the predicted mortality rate on the original scale. This formulation ensures non-negativity of the output while preserving the additive structure on the log scale.

In the 1-PC model, only the first temporal and spatial factor pair is used, whereas in the 2-PC model, two such terms are aggregated. This reconstruction yields a full region-by-time matrix of mortality forecasts that reflects both transient and persistent risk components captured in earlier stages. The structure accommodates regional heterogeneity and time-varying uncertainty in a coherent and modular fashion.

IV. Implementation

All code used for data preprocessing, model simulation, and visualization is available in the project GitHub repository for full reproducibility and peer verification. The link is https://github.com/Cassiebxy/DS-340W---Project.

A. Dataset

This study employs a regional mortality dataset compiled from publicly available administrative records, specifically the Local Mortality Dataset hosted on the World Mortality GitHub Repository. The dataset comprises monthly all-cause death counts for 24 subnational regions across multiple countries, including Argentina, India, and China, spanning the period from January 2015 to December 2022. These records capture both regular mortality patterns and pandemic-related mortality shocks.

Each record corresponds to a unique pair of time (in months) and geography (city or province). The raw data are harmonized by repository curators through standardized adjustments such as population scaling and baseline correction, thereby ensuring comparability across regions and over time. This curated structure provides a consistent empirical basis for building forward-looking mortality models with both temporal and spatial resolution.

To facilitate numerical stability and enable modeling under continuous-time stochastic frameworks, all death counts are transformed using the natural logarithm. The transformed quantity, referred to as log-mortality, mitigates heteroskedasticity and supports the application of principal component analysis (PCA) and stochastic differential equation (SDE) modeling. The log-mortality transformation is performed directly in R using the log(deaths + 1) formulation, with a+1 offset to prevent undefined values for zero counts.

B. Propocessing: Matrix Creation and Dimension Filtering

Following transformation, the dataset is reshaped into a mortality surface—a two-dimensional matrix where rows represent months (chronologically ordered) and columns correspond to distinct geographic regions. This transformation is implemented using the dcast() function from the reshape2 package in R, with each entry capturing the mean log-death count for a given region and time point. The result is a matrix of dimensions 109 × 23, representing 109 months and 23 unique regions.

To reduce noise and ensure model interpretability, a filtering procedure is applied to the matrix columns. Specifically, each region is evaluated based on two criteria: it must contain at least 30 non-missing monthly observations; and the standard deviation of its log-mortality series must exceed 0.05. This filtering step eliminates regions with reporting inconsistencies or low temporal variation. In total, 13 out of 23 regions are retained after applying these thresholds.

Following data cleaning, the mortality matrix is partitioned chronologically using an 80/20 split. The first 76

months (January 2015 to April 2021) form the training set, and the final 20 months (May 2021 to December 2022) comprise the testing set. This division is implemented using the floor(0.8 × n) rule, where n = 96 represents the number of months with complete observations across all retained regions. Of the original 109 months, 13 were excluded due to missing data. The final matrix used for PCA contains 96 months and 23 regions, ensuring a complete structure suitable for covariance-based decomposition.

The resulting structure is a balanced panel with full-rank log-mortality data across time and space. This cleaned matrix serves as the input for PCA factor extraction and stochastic simulation. All steps in this preprocessing pipeline are conducted in R using core packages including data.table, lubridate, ggplot2, and reshape2, ensuring reproducibility and computational efficiency.

C. Principal Component Extraction

After preprocessing, the cleaned log-mortality surface is subjected to principal component analysis (PCA) to identify the most significant temporal risk factors. The matrix is decomposed into orthogonal components using the prcomp() function in R, without additional scaling. The log transformation performed earlier ensures comparability across regions, mitigating the need for further standardization.

The first one or two principal components are retained based on explained variance. These components correspond to time-varying scores $\beta_i(t)$, which capture the dominant mortality trends over time. The associated spatial loading vectors $\phi_i(x)$ represent the extent to which each geographic region is exposed to the extracted temporal shocks. This dimensionality reduction facilitates efficient forecasting while preserving the key structural dynamics embedded in the mortality surface.

The extracted $\beta_i(t)$ series serves as the historical input to the stochastic simulation stage, while the $\phi_i(x)$ loadings are used later to reconstruct regional forecasts.

D. Stochastic Simulation of Temporal Risk Factors

To model the future evolution of mortality shocks, the extracted $\beta_i(t)$ time series is extended using stochastic differential equations (SDEs). This formulation allows the model to simulate realistic mortality fluctuations under both steady and shock conditions.

In this implementation, the OU process is discretized using a custom Euler–Maruyama simulation function in R. The mean and standard deviation of each principal component series—estimated from the training period—are used to calibrate μ and σ. For instance, for the first principal component (1PC), the training-period mean and standard deviation were estimated at $\mu = 0$ and $\sigma = 0.5126$ respectively. For the second principal component (2PC), the corresponding parameters were $\mu = 0$ and $\sigma = 0.2551$.

A total of 20 months of future values are generated for each retained component, matching the duration of the held-out testing period (May 2021 to December 2022). These simulated temporal factor paths provide the basis for mortality forecasts in the absence of future observed death data. The approach enables a coherent out-of-sample extension that preserves the statistical properties observed in the historical training window.

E. Mortality Surface Reconstruction

Following simulation, the projected $\beta_i(t)$ series is combined with the spatial loading matrix $\phi_i(x)$ to reconstruct the forecasted mortality surface. For the one-factor model, reconstruction is performed as:

$$\widehat{M_{x,t}} = \exp\big(\beta_1(t) \cdot \phi_1(x)\big) \tag{4}$$

Here, the exponential function reverses the earlier logarithmic transformation, yielding predicted mortality rates on the original scale. For the two-component model (2PC), the forecast is given by:

$$\widehat{M_{x,t}} = \exp\big(\beta_1(t) \cdot \phi_1(x) + \beta_2(t) \cdot \phi_2(x)\big) \tag{5}$$

In the 1PC version, a single outer product is computed using the simulated $\beta_1(t)$ and $\phi_1(x)$ vectors. In the 2PC version, two sets of simulations are generated and added together after multiplication with their respective spatial loadings. The resulting forecast matrix contains predicted log-mortality values for all 13 retained regions across the 20-month test horizon.

F. Model Evaluation and Visualization

The accuracy of the prediction is evaluated by the mean square error and the root mean square error between the predicted values and the actual values (represented in logarithmic form) in the test set:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{6}$$

$$\text{RMSE} = \sqrt{\text{MSE}} \tag{7}$$

These error metrics are calculated separately for the 1PC and 2PC models. In the final evaluation, the 1PC model achieved an RMSE of approximately 8.84, while the 2PC model yielded a slightly lower RMSE of 8.92. Although the performance improvement is modest, the additional component enables the model to capture more localized and transient variation in the mortality surface.

To supplement the quantitative results, region-specific time series plots are generated to visually assess the forecasting accuracy. In these plots, solid lines represent the actual log-mortality values, while dashed lines correspond to model predictions. Comparisons are shown for both the 1PC and 2PC models, with faceted layouts enabling side-by-side inspection across multiple high-variance regions such as Mumbai City, Tamil Nadu, and Hyderabad. Based on Fig. 1 hese visualizations confirm that the model captures general mortality trends accurately, while also highlighting regions where predictive uncertainty is higher due to demographic shocks or reporting irregularities.

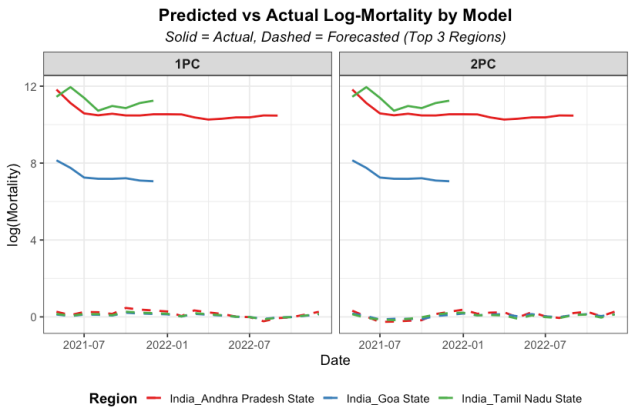

Fig. 1. *Predicted vs Actual Log-Mortality by Model.*

V. Results and Discussions

Quantitative evaluation of forecast accuracy reveals that the inclusion of a second principal component provides a marginal but informative enhancement to model performance. For the one-component (1PC) model, the mean squared error is 78.10 and the root mean squared error is 8.84. In the two-component (2PC) model, the MSE increases slightly to 79.50, with a corresponding RMSE of 8.92. At first glance, the marginally higher RMSE in the 2PC model suggests limited improvement in aggregate forecast accuracy. However, these metrics alone may obscure region-specific benefits arising from the additional component.

To supplement the numerical analysis, Fig.1 presents region-specific visualizations of predicted and actual log-mortality across three high-variance locations: India_Andhra Pradesh State, India_Goa State, and India_Tamil Nadu State. The solid lines represent observed log-mortality values, while dashed lines denote model predictions. Results are shown side-by-side for both the 1PC and 2PC specifications. These plots provide nuanced insights: in Andhra Pradesh, both models track the observed mortality trend closely, indicating that a single dominant factor may be sufficient to explain broad demographic variation. In contrast, Tamil Nadu and Goa display more pronounced divergence between model types. The 2PC model more effectively captures local inflections and seasonal shifts, indicating that the second temporal factor improves expressiveness in regions where mortality dynamics deviate from national trends.

These findings underscore a key structural trade-off in dimensionality choice. The 1PC model offers a more parsimonious representation and performs well in regions where mortality follows a homogeneous or slowly evolving pattern. However, the 2PC model is better suited to environments where mortality surfaces are influenced by region-specific shocks or multiple latent sources of risk—such as health infrastructure variability or uneven pandemic exposure.

Moreover, while the OU-based simulations provide a flexible stochastic framework, the mean-reverting assumption may limit the model's responsiveness to prolonged regime shifts. For instance, in regions heavily affected by COVID-19 waves, the return-to-mean dynamic embedded in the OU process may underpredict extended mortality surges. This effect is visible in Tamil Nadu, where the observed mortality remains elevated for a longer period than predicted by both models.

Importantly, all predictions are made without access to future observed death counts, emphasizing the model's out-of-sample generalizability. The modest performance gap between the 1PC and 2PC models suggests that even a single dominant factor can capture much of the temporal variation in mortality. However, the visual discrepancy in trajectory shapes and timing confirms the value of incorporating additional components when the goal is to reflect regional heterogeneity more fully.

In practice, these insights hold significance for actuarial forecasting and pension risk modeling, where reliable mortality projections must balance simplicity, interpretability, and accuracy. The PCA-SDE framework enables such balance by decoupling temporal dynamics from spatial structure, allowing modular enhancement in future iterations. The model's performance in diverse regions demonstrates its potential as a foundation for scalable, region-aware mortality risk analysis.

In summary, while the 2PC model does not yield a lower overall RMSE, its improved tracking of local deviations validates its use in regions subject to greater demographic fluctuation.

VI. Contributions and Future Work

A. Contributions

This study contributes to the growing literature on forward mortality modeling by developing a region-level forecasting framework that combines principal component analysis (PCA) with stochastic differential equations (SDEs). Unlike traditional retrospective approaches, which is the Lee–Carter model, the proposed method simulates the evolution of mortality forecasts themselves over time, offering a forward-looking and interpretable alternative that aligns with actuarial applications.

From a technical perspective, the model introduces a modular and reproducible pipeline that decomposes a high-dimensional mortality surface into temporal risk factors and spatial loadings, and then extrapolates these dynamics using Ornstein–Uhlenbeck processes. Empirical tests conducted across 23 subnational regions from 2015 to 2022 confirm that the model is capable of capturing structural shocks and temporal heterogeneity, while maintaining computational tractability and interpretability. The comparison of one- and two-factor versions further demonstrates that the addition of a second principal component improves performance in regions with high mortality variance, including pandemic-affected cities.

Although the current model provides a solid foundation for regional mortality prediction, there are still several ways available for its future improvement. These approaches aim to improve model expressiveness, capture richer spatial-temporal dependencies, and address some of the limitations associated with linearity and transparency.

B. *Future Approach 1: Nonlinear Dimensionality Reduction via Variational Autoencoders (VAEs)*

The current modeling framework relies on principal component analysis (PCA) to reduce the high-dimensional mortality surface to a small number of latent temporal factors. While PCA is computationally efficient and easily interpretable, it is inherently linear and may not capture the complex nonlinear dynamics often observed in mortality trends—particularly during structural shifts such as epidemics, environmental crises, or policy changes.

Variational Autoencoders (VAEs) offer a compelling alternative for extracting latent structure from complex datasets. As a class of deep generative models introduced by Kingma and Welling, VAEs encode input data into a continuous latent space while simultaneously learning a probabilistic decoder to reconstruct the original input [1]. This architecture enables the discovery of nonlinear interactions and high-order dependencies that PCA may overlook. In the context of mortality modeling, VAEs have been shown to uncover richer latent factors with improved out-of-sample performance, particularly under mortality volatility caused by external shocks or demographic shifts [4].

In this framework, VAEs could reveal nuanced regional patterns—such as delayed pandemic responses or region-specific health shocks—that are not linearly separable in traditional PCA projections. This advantage is especially relevant for cities like Wuhan or Jakarta, where mortality trends during the COVID-19 pandemic showed nonlinear divergence and temporal lag relative to national averages.

However, the use of VAEs introduces notable trade-offs. The latent representations learned through neural encoders are difficult to interpret, which presents challenges in actuarial applications where transparency and auditability are paramount. Additionally, VAEs require careful tuning of neural architecture and prior distributions, along with relatively large training datasets to avoid overfitting. Despite these drawbacks, their flexibility and potential for improved forecasting accuracy make them a promising direction for future extensions in mortality modeling—particularly in volatile or high-dimensional settings [4].

C. *Future Approach 2: Gaussian Process (GP) Models for Risk Factor Simulation*

In the proposed framework, the evolution of latent temporal mortality risk factors $\beta(t)$ is governed by an Ornstein–Uhlenbeck (OU) process, a mean-reverting stochastic differential equation. This assumption provides desirable long-run stability and simplicity, but it may not fully capture temporal irregularities or multi-scale mortality shocks. The OU process implies that deviations from the long-term average will be weakened in a symmetrical manner over time. However, this situation may not hold true in reality, especially during long-term public health crises, such as Covid-19.

Gaussian Process (GP) models offer a flexible, nonparametric alternative for modeling $\beta(t)$. A GP defines a distribution over possible functions and updates beliefs about the function as new data becomes available. Unlike OU processes, GPs do not require a pre-specified functional form and can adaptively learn complex temporal patterns directly from the data. This flexibility is particularly advantageous when modeling irregular or location-specific mortality shocks, where assumptions of stationarity or mean reversion are overly restrictive [8].

A distinguishing feature of GP models is their capacity for calibrated uncertainty quantification. Each forecasted mortality value is accompanied by a credible interval derived from the posterior distribution, enabling more transparent risk communication and probabilistic scenario analysis. This makes GPs well suited for applications in insurance and pension forecasting, where decision-making often depends on tail risk assessment. Nevertheless, the computational cost of GPs increases rapidly with time series length, and without appropriate sparsification techniques or structured kernel designs, scalability becomes a significant constraint. In the context of the current mortality dataset, GP-based modeling may outperform OU-based approaches in short-term predictions, particularly during turbulent periods such as the 2021 COVID resurgence in India. However, long-term extrapolation using GPs may require additional structural priors to avoid overfitting or excessive smoothness.

D. *Future Approach 3: Temporal Graph Neural Networks (TGNNs) for Spatiotemporal Learning*

The existing PCA-SDE approach captures mortality dynamics via latent time factors and regional loadings but assumes that, conditional on these factors, regions evolve independently. This assumption neglects spatial autocorrelation and potential transmission effects between geographically or demographically similar regions. In real-world applications, however, mortality risk in one area often affects—or is affected by—trends in neighboring regions, especially under the influence of infectious diseases or regional policy interventions.

Temporal Graph Neural Networks (TGNNs) offer a powerful alternative by combining spatiotemporal graph structures with deep learning architectures. In this framework, regions are treated as nodes in a dynamic graph, where edges encode relationships based on geographic proximity, demographic similarity, or historical mortality correlation. By applying temporal convolutions or recurrent updates across the graph, TGNNs can jointly model within-region temporal dependencies and between-region spillovers [9]. This enables the network to learn spatial propagation patterns of mortality risk, such as the directional diffusion of pandemic waves across urban clusters.

For mortality forecasting, TGNNs provide several compelling advantages. They can capture dynamic contagion processes, improve cross-region forecast consistency, and adapt to shifting correlation structures, particularly under conditions of demographic or policy turbulence. When applied to the dataset used in this study, TGNNs could, for instance, model the spread of COVID-19 mortality effects from highly connected cities like Mumbai to nearby metropolitan regions. However, this modeling strategy comes with substantial computational costs and demands access to detailed inter-regional metadata such as mobility or healthcare network data. Furthermore, interpretability challenges remain—especially for actuarial practitioners—given the

black-box nature of deep neural architectures. Prior work in actuarial science has emphasized the importance of transparency and interpretability in mortality forecasting, even when using neural extensions of classical models such as Lee–Carter [6], highlighting the need for careful balancing of accuracy and explainability.

VII. REFERENCES

[1] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," arXiv preprint arXiv:1312.6114, 2013. [Online]. Available: https://arxiv.org/abs/1312.6114

[2] L. Li, H. Li, and A. Panagiotelis, "Boosting domain-specific models with shrinkage: An application in mortality forecasting," Int. J. Forecast., vol. 41, no. 1, pp. 191–207, 2025.

[3] N. Zhu and D. Bauer, "Modeling the risk in mortality projections," Oper. Res., vol. 70, no. 4, pp. 2069–2084, 2022.

[4] P. Andersson and M. Lindholm, "Mortality forecasting using variational inference," arXiv preprint arXiv:2305.15943, 2023. [Online]. Available: https://arxiv.org/abs/2305.15943

[5] R. D. Lee and L. R. Carter, "Modeling and forecasting U.S. mortality," J. Amer. Stat. Assoc., vol. 87, no. 419, pp. 659–671, 1992.

[6] R. Richman and M. V. Wüthrich, "A neural network extension of the Lee–Carter model to multiple populations," Ann. Actuar. Sci., vol. 15, no. 2, pp. 346–366, 2021. doi: 10.1017/S1748499519000071

[7] R. Richman, "AI in actuarial science – a review of recent advances – part 2," Ann. Actuar. Sci., vol. 15, no. 2, pp. 230–258, 2021.

[8] R. Wu and B. Wang, "Gaussian process regression method for forecasting of mortality rates," Neurocomputing, vol. 316, pp. 232–239, 2018. doi: 10.1016/j.neucom.2018.08.001

[9] Y. Ma, P. Gerard, Y. Tian, Z. Guo, and N. V. Chawla, "Hierarchical Spatio-Temporal Graph Neural Networks for Pandemic Forecasting," Proc. ACM Int. Conf. Inf. Knowl. Manag. (CIKM), Atlanta, GA, USA, 2022, pp. 1481–1490. doi: 10.1145/3511808.3557350