POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Online Anomaly Detection and Probabilistic Interventions for Long-Term Structural Health Monitoring

ZHANWEN XIN
Département de génie civil, géologique et des mines

Thèse présentée en vue de l'obtention du diplôme de *Philosophiæ Doctor*
Génie civil

Août 2026

© Zhanwen Xin, 2026.

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Cette thèse intitulée :

Online Anomaly Detection and Probabilistic Interventions for Long-Term Structural Health Monitoring

présentée par **Zhanwen XIN**
en vue de l'obtention du diplôme de *Philosophiæ Doctor*
a été dûment acceptée par le jury d'examen constitué de :

Nicolas SAUNIER, président
James-A. GOULET, membre et directeur de recherche
Rocio Lilen SEGURA, membre
Sandra POZZER, membre externe

DEDICATION

To my parents, Jiefang Wu, and Liang Xin
To my friends...

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and appreciation to my advisor, James-A. Goulet, for generously sharing his knowledge, for his dedicated supervision, for the inspiration he offered, and for his encouragement throughout this journey — without whom this thesis would not have been possible.

I am grateful to the Natural Sciences and Engineering Research Council of Canada (NSERC) and Hydro-Québec (HQ) for their financial support of this research, and I would like to further thank Hydro-Québec for providing the datasets used in this thesis. My sincere thanks also go to Patrice Côté, Vincent Roy, Benjamin Miquel, and Stéphane Lafrance from Hydro-Québec for their valuable suggestions and support throughout the project.

I would also like to thank the friends and colleagues at Polytechnique Montréal with whom I have had the pleasure of working over the years: Dai Vuong Van, Lucas Alric, Marina Decker, Miquel Florensa, David Wardan, Michel Wu, Luong Ha Nguyen, Bhargob Deka, Zachary Hamida, Ali Fakhri, Shervin Khazaeli, and the other friends of the Groupe de recherche en structures, especially Léa Zheng. Beyond the research group, I am grateful to the friends I made along the way: Fabio Inzirillo, with whom I discovered my passion for climbing; Achille Occhini, who helped me improve my French; and Evan Johnston, with whom I shared many unforgettable memories. I am also deeply grateful to Xiaodong Wang, who first set me on the path of graduate studies and encouraged me every step toward this degree. A special thank you goes to Lucas Alric for always being there, for his inspiration, and for the coffee breaks we shared.

Last but not least, I would like to thank my family for their unconditional support over the years. To my parents, who have always motivated and encouraged me during my studies — thank you.

Declaration of generative AI in the manuscript preparation process. During the preparation of this thesis, the author used Claude to improve the readability and language of the work. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the thesis.

ABSTRACT

Civil structures such as bridges and dams deteriorate slowly and often invisibly, and the interval between the damage starting time and its discovery is precisely where risk accumulates. Structural Health Monitoring (SHM) aims to close this gap by continuously interpreting sensor measurements, where deterioration typically manifests itself as a small, irreversible change in the baseline of the structural response. Detecting such changes over the decades-long service life of a structure poses four simultaneous demands: 1) catching small baseline changes despite much larger reversible environmental and operational fluctuations, 2) maintaining a high detection probability while limiting false alarms, 3) adapting to the new stationary regime once a change has occurred so that subsequent anomalies remain detectable, and 4) doing so at a computational cost compatible with monitoring large networks of structures. No existing method satisfies all of these demands at once.

This thesis addresses that gap by developing online anomaly detection methods within the state-space model (SSM) framework, unified by a common principle: the stationarity of the residual component. Three complementary contributions are proposed. First, the bounded autoregressive (BAR) component constrains the autoregressive residual within the bounds of its stationary distribution, redirecting irreversible changes toward the baseline hidden states where the switching Kalman filter (SKF) can detect them as regime switches. Second, the residual stationarity intervention (RSI) method isolates anomaly-induced drift with an auxiliary drift model, based on which it quantifies the probability of anomaly and, using a Bayesian neural network, predicts the intervention that enables the SSM's adaptation to new regimes. Third, the residual stationarity intervention and classification (RSIC) method builds on the same drift-based detection and additionally infers the anomaly type online through a probabilistic classification, applying the corresponding targeted correction to enable reliable detection of sequential anomalies of different forms.

Each method was verified probabilistically on populations of synthetic time series and validated on real measurements from full-scale infrastructure. Within the SKF framework, the BAR component detects smaller anomalies earlier than an equivalent model using a standard autoregressive component. Operating on SSMs alone, RSI outperforms five established methods, achieving higher detection probabilities and shorter times to detection on time series containing a single trend anomaly, while RSIC extends this advantage to time series with sequential anomalies of different types, outperforming all evaluated methods in that setting.

RÉSUMÉ

Les structures de génie civil telles que les ponts et les barrages se détériorent lentement et souvent de manière invisible, et l'intervalle entre le début d'un dommage et sa découverte est précisément là où le risque s'accumule. La surveillance de l'état des structures (SHM) vise à combler cet écart en interprétant continuellement des mesures du comportement, où la détérioration se manifeste typiquement par un faible changement irréversible dans la tendance de base de la réponse structurale. Détecter de tels changements pose quatre exigences simultanées : 1) identifier les faibles changements de la tendance de base malgré des fluctuations environnementales réversibles, 2) maintenir une probabilité de détection élevée tout en limitant les fausses alarmes, 3) s'adapter aux nouveaux régimes stationnaires afin que les anomalies subséquentes demeurent détectables, et 4) le faire de manière computationnellement efficace. Aucune méthode existante ne satisfait toutes ces exigences à la fois.

Cette thèse aborde cet écart en développant des méthodes de détection d'anomalies en ligne au sein du cadre des modèles d'espace d'états (SSMs), unifiées par un principe commun : la stationnarité de la composante résiduelle. Trois contributions complémentaires sont proposées. Premièrement, la composante autorégressive bornée (BAR) contraint le résidu autorégressif, redirigeant les changements irréversibles vers les états cachés de la tendance de base, où le filtre de Kalman à commutation (SKF) peut les détecter. Deuxièmement, la méthode d'intervention de stationnarité résiduelle (RSI) isole la dérive induite par l'anomalie à l'aide d'un modèle de dérive auxiliaire, d'où elle quantifie la probabilité d'anomalie et, en utilisant un réseau de neurones bayésien, prédit l'intervention qui permet l'adaptation des SSMs aux nouveaux régimes. Troisièmement, la méthode d'intervention et classification de stationnarité résiduelle (RSIC) s'appuie sur la même détection par la dérive et infère en outre le type d'anomalie en ligne et de façon probabiliste, appliquant la correction ciblée afin de permettre la détection fiable d'anomalies séquentielles de formes différentes.

Chaque méthode a été vérifiée de manière probabiliste sur des populations de séries temporelles synthétiques et validée sur des mesures réelles provenant d'infrastructures. Au sein du cadre du SKF, la composante BAR détecte les anomalies de plus faible amplitude plus tôt qu'un modèle équivalent utilisant une composante autorégressive. Opérant sur les SSMs seuls, la RSI surpasse cinq méthodes établies, avec des probabilités de détection plus élevées et des temps à la détection plus courts sur les séries temporelles contenant une seule anomalie de tendance, tandis que la RSIC étend cet avantage aux séries temporelles comportant des anomalies séquentielles de types différents, surpassant toutes les méthodes évaluées.

TABLE OF CONTENTS

DEDICATION iii

ACKNOWLEDGEMENTS iv

ABSTRACT v

RÉSUMÉ vi

LIST OF TABLES x

LIST OF FIGURES xii

LIST OF SYMBOLS AND ACRONYMS xix

LIST OF APPENDICES xxiii

CHAPTER 1 Introduction 1
1.1 Motivation 1
1.2 Research Objectives 3
1.3 Thesis Outline 3
1.4 Co-Authored Papers 4

CHAPTER 2 Literature Review 5
2.1 Structural Health Monitoring and Anomaly Detection 5
2.1.1 Aging Infrastructure and Continuous Monitoring 5
2.1.2 Anomaly Types in SHM 6
2.1.3 Key Challenges 8
2.2 Time-Series Anomaly Detection Methods 9
2.2.1 Distance-based Methods 10
2.2.2 Prediction-Based Methods 12
2.2.3 Deep-Learning Methods 14
2.2.4 Summary 16
2.3 State-Space Models for SHM 16
2.3.1 State-Space Models and the Kalman-Based Methods 16
2.3.2 Component Decomposition 23
2.4 Residual Modeling and Stationarity 26
2.4.1 Autoregressive Residual Model and Online Parameter Estimation . . 26

2.4.2 Residual's Drift and its Partial Absorption 28
2.5 Intervention Models and Regime Adaptation 30
2.5.1 Intervention Analysis 30
2.5.2 Neural-Network Based Intervention Predictors 31
2.6 Evaluation Methodology 33
2.6.1 Standard Evaluation Metrics 33
2.6.2 Evaluation Criteria for the SHM Context 35
2.7 Conclusion 37

CHAPTER 3 Bounded Autoregressive Component in State-Space Models 38
3.1 Introduction 38
3.2 Bounded Autoregressive Component 38
3.2.1 Lower- and Upper-Bounded ReLU and Gaussian Mixture Reduction 39
3.2.2 Mixture Rectified Linear Unit (mReLU) 40
3.2.3 BAR Formulation 43
3.3 Evaluation Metrics 45
3.4 Experiments 47
3.4.1 Synthetic Data 47
3.4.2 Bridge Data 48
3.4.3 Dam Data 51
3.5 Conclusions 57

CHAPTER 4 Anomaly Detection via Residual Stationarity Intervention 59
4.1 Introduction 59
4.2 Methodology 59
4.2.1 Residual's Drift Model 60
4.2.2 Probability of Anomaly 61
4.2.3 Intervention Model 63
4.3 Experiments 66
4.3.1 Synthetic Time Series 67
4.3.2 Real Time Series 71
4.4 Conclusions 76

CHAPTER 5 Classification-Guided Residual Stationarity Intervention 77
5.1 Introduction 77
5.2 Methodology 78
5.2.1 Limitations of RSI 78

5.2.2 Overview of RSIC . . . 79
5.2.3 Intervention Types and Their Inference . . . 80
5.2.4 Tentative Interventions . . . 81
5.2.5 Classification of Intervention Types . . . 82
5.3 Experiments . . . 85
5.3.1 Synthetic Time Series . . . 85
5.3.2 Real Time Series . . . 90
5.4 Conclusions . . . 97

CHAPTER 6 Conclusion . . . 99
6.1 Thesis Conclusion . . . 99
6.2 Limitations . . . 101
6.3 Future Research . . . 102

REFERENCES . . . 104

APPENDICES . . . 117
A.1 Synthetic Data . . . 117
A.1.1 Synthetic Data Generation . . . 117
A.1.2 Stationary Model Matrices in SKF . . . 117
A.1.3 Non-Stationary Model Matrices in SKF . . . 118
A.2 Bridge Data . . . 118
A.2.1 Stationary Model Matrices in SKF . . . 118
A.2.2 Non-Stationary Model Matrices in SKF . . . 119
A.3 Dam Data . . . 119
A.3.1 Stationary Model Matrices in SKF . . . 120
A.3.2 Non-Stationary Model Matrices in SKF . . . 120
F.1 Drift Model Matrices . . . 130
F.2 Model Matrices for Synthetic Time Series Generation . . . 130
G.1 RSI and RSIC . . . 132
G.2 Switching Kalman Filter . . . 133
G.3 Matrix Profile and DAMP . . . 133
G.4 Prophet . . . 134
G.5 LSTMED . . . 134
G.6 TranAD . . . 134
K.1 Maximum Waiting Time . . . 139
K.2 p Quantile . . . 142

LIST OF TABLES

Table 2.1 A confusion matrix in classification problems and its analogue in anomaly detection. True positives (TP) are the instances in which an alarm is triggered in the presence of abnormal data. False positives (FP) are the false alarms in anomaly detection. False negatives (FN) correspond to the missed alarms. True negatives (TN) indicate that no alarm is triggered given that there is no anomaly. 34

Table 3.1 Summary of eight time series measured on a dam with respect to time-series name, measurement type, start & end time, acquisition period, number of observations and SSM components. Crack openings and displacement are notated as C.O. and Disp. Stationary/ non-stationary regimes are referred to as S./ N.S. The abbreviation of res. includes AR or BAR residual component. The *highlighted row* is the time series example showcased in the following sections. 52

Table 3.2 Summary of AR vs. BAR-γ^* performance. γ^* is the optimal parameter corresponding to the highest $F1t$ in the BAR method. **Bolded** numbers indicate the $F1t$ scores improved by more than 80% in BAR compared to AR. 56

Table 4.1 False alarm rate on the test set (#/10 years) of six detection models on two datasets. 69

Table 5.1 Comparison of false alarm rate, baseline estimation errors (in standardized space) across different methods, and RSIC classification performance on synthetic time series with various anomaly combinations: level to level (L→L), level to trend (L→T), trend to level (T→L), and trend to trend (T→T). 89

Table 5.2 Comparison of false alarm rate, baseline estimation errors (in standardized space) across different methods, and RSIC classification performance on ten detrended time series with different combinations of anomalies: level to level (L→L), level to trend (L→T), trend to level (T→L), and trend to trend (T→T). 93

Table H.1 Mean squared errors (MSEs) of the Bayesian neural network (BNN) used as the RSI intervention model on the validation set for different BNN hyperparameters, including network architecture, batch size, patience number, and input length. 135

Table K.1 Performance metrics and classification accuracy with regard to different $\tilde{t}$ on synthetic time series with different combinations of anomalies: level to level (L→L), level to trend (L→T), trend to level (T→L), and trend to trend (T→T). 141

Table K.2 Performance metrics and classification accuracy with regard to different quantile value $\tilde{p}$ on synthetic time series with different combinations of anomalies: level to level (L→L), level to trend (L→T), trend to level (T→L), and trend to trend (T→T). 144

LIST OF FIGURES

Figure 2.1 Proportion of U.S. bridges and dams at or approaching 50 years of service life (2025). Adapted from the American Society of Civil Engineers [1]. 5

Figure 2.2 Taxonomy of nine time-series anomaly types, grouped into point-based and sequence-based categories. Global, contextual, noise, speedup, scale, and flip anomalies are established in the literature [2, 3]. The highlighted baseline-changed types, especially level and trend anomalies, are the focus of this thesis, as they represent potential structural anomalies. 7

Figure 2.3 A time series transitioning from one trend-stationary regime to another at a changepoint. The segment after the changepoint is abnormal with respect to the first regime, yet on its own forms a new stationary regime. 8

Figure 2.4 Illustration of the subtlety of a small trend anomaly. The gradual baseline change is introduced at the red dashed line, and its initial phase is indistinguishable from the normal structural response by visual inspection alone. 8

Figure 2.5 Detection (marked by blue dashed lines) on a long-term time series containing two sequential anomalies: a change in level (teal) followed by a change in trend (maroon). 10

Figure 2.6 The process of the Kalman filter, showing the (a) initialization, (b, f) transition, (c) prediction, (d) observation, (e) update, and (g) recursion steps. 18

Figure 2.7 Information flow of the Kalman filter in a time series from time step 0 to `T`. 18

Figure 2.8 Information flow of the Kalman smoother in a time series from time step `T` to 0. 19

Figure 2.9 A time-series example that switches from a level-stationary regime to a trend-stationary one. The Kalman filter defined only by a constant trend model fails to estimate the hidden states after the regime switching. 20

Figure 2.10 Information flow of the switching Kalman filter between two consecutive time steps, shown for two regimes ($\mathtt{S} = 2$). Adapted from Murphy [4]. 21

Figure 2.11 SKF detection delay on a real time series: part of the anomaly-induced drift in the observation (Obs.) is absorbed by the residual component (Res., third panel), suppressing the anomaly probability $p_{\texttt{anm}}$ (bottom panel) and delaying detection. 22
Figure 2.12 Detectability of the switching Kalman filter (SKF), the tabular Q-learning agent, and the imitation learning agent. Each curve marks the detection boundary as a function of the anomaly magnitude $|h_a|$ and duration w_a. Adapted from Khazaeli [5]. 23
Figure 2.13 Online learning of the AR hyperparameters $\phi^{\texttt{AR}}$ and $(\sigma^{\texttt{AR}})^2$ on synthetic data. Estimates (blue, ± one standard deviation) converge to the true values (red dashed) without offline calibration. 28
Figure 2.14 Delay in the residual estimation when the time series exhibits a non-stationary trend. 29
Figure 2.15 Two categories of time-series anomaly detection metrics. 34
Figure 3.1 Illustration of the core idea behind the BAR component using a synthetic time series y subjected to a trend anomaly starting at the midpoint of the time series. 39
Figure 3.2 Four Gaussian mixture reduction examples. The first and the second plots exhibit the mixture $\tilde{\mathcal{N}}$ for two Gaussian components. The third and the fourth plots show the mixtures $\tilde{\mathcal{N}}$ for three Gaussian components. 40
Figure 3.3 A mReLU transformation example in the context of BAR. The input distribution $f(x^{\texttt{AR}})$ has a mean value that is close to $-b$ within the boundary and a variance that extends beyond the $\pm b$. 41
Figure 3.4 Ten realizations of an autoregressive process with $\phi^{\texttt{AR}} = 0.8$ and $\sigma_W = 0.5$, along with the stationary bounds $\pm b = \pm\gamma\sigma^{\texttt{AR},0}$ for $\gamma \in \{1, 2, 3\}$ following Eq. 3.4. The default value $\gamma = 2$ bounds a 95 % coverage region of the process's stationary distribution. 42
Figure 3.5 Integration process of the BAR component in the SSMs framework. The modifications made by the BAR method are highlighted by dashed lines. 45
Figure 3.6 Time-series-wise TP, FP and FN definition. FP is incremented by 1 when an alarm is triggered before the occurrence of anomalies. TP is incremented by 1 when an alarm is correctly triggered in the abnormal segment. FN is increased by 1 if an alarm is missed within the detection window. When any of these three scenarios happens, analysis on this time series terminates and the confusion matrix is updated. 46

Figure 3.7 Decreasing λ with longer time to detection. l_{dw} is the length of the target detection window. λ drops to zero if the required time to detect an anomaly is longer than l_{dw}. 46
Figure 3.8 A one-year synthetic time series with daily time-step length generated with `LL` and `AR` components. A trend anomaly starts at the midpoint 2020-07-01. 47
Figure 3.9 Hidden states estimation on the synthetic data with the standard `AR` and `BAR` in comparing with the true values. 49
Figure 3.10 Elongation and temperature measurement on a bridge, where time steps are aggregated from 10 minutes to 3.5 days. 49
Figure 3.11 Comparison of hidden states between SKF-AR and SKF-BAR under different anomaly magnitudes in a time series measured on a bridge. Large anomaly indicates a trend anomaly of -0.5 mm/year; the medium is -0.1 mm/year; and the small is -0.025 mm/year. The residual components (`res.`) include the `AR` and the `BAR` components. 50
Figure 3.12 100 synthetic time series (colored) similar to a real time series (black) measured on a dam. 53
Figure 3.13 Performance of the `BAR` method under different γ on synthetic time series with three types of anomalies in comparison with the `AR`. Evaluation metrics are time to detection (Δ_t), detection probability ($p^{\texttt{det}}$), false alarm number (#FA) per ten years and $F1t$ score. The white regions on the top three subfigures indicate the undetectable anomalies and their dark red regions represent all the negative time to detection caused by the false alarms (FA). False alarm rate is drawn in logarithmic scale. A higher value of $F1t$ indicates better performance. Repetition $\texttt{N} = 100$. 54
Figure 3.14 The $F1t$ scores for `BAR`-γ^* and `AR` verified on the synthetic datasets similar to the eight real time series measured on a dam. Acceleration anomalies are overlaid on the synthetic time series. Additional $F1t$ scores with respect to the trend and level anomalies are shown in Appendix D. 55

Figure 3.15 Validation results with regard to the anomaly detection probability ($p^{\texttt{det}}$) and time to detection (Δ_t) on the real time series M08C with acceleration anomalies of different magnitudes by using `BAR`-0.4 and `AR` components. The hyperparameter γ is set to 0.4 in `BAR`. Additional validation results for other time series and anomaly types are reported in Appendix E. 57

Figure 4.1 Overview of the RSI procedure from time t to $t+1$, illustrating the key methodologies covered in this section: residual's drift modeling with an auxiliary SSM, residual stationarity monitoring, and BNN-driven intervention. 60

Figure 4.2 Residual decomposition with the drift model. The posterior hidden states from the drift model are plotted in orange. The residuals plotted with $x^{\texttt{Ld}}$ in the first subfigure are estimated using the base SSM in the example shown in Figure 2.14. 61

Figure 4.3 Illustration of the retrospective intervention. The intervention is applied at the intervention time $t_{\texttt{itv}}$ (gray dashed line), sent back from the detection time $t_{\texttt{det}}$ (red dashed line) by the predicted anomaly duration (Eq. 4.5). *Left two panels (base SSM):* the level $x^{\texttt{L}}$ and trend $x^{\texttt{T}}$ are corrected at $t_{\texttt{itv}}$ by the interventions $I^{\texttt{L}}_{t_{\texttt{itv}}}$ and $I^{\texttt{T}}_{t_{\texttt{itv}}}$ (Eqs. 4.6–4.7), reassigning the non-stationary pattern to the baseline of the base SSM. *Right two panels (drift SSM):* the baseline hidden states of the drift model, $x^{\texttt{Ld}}$ and $x^{\texttt{Td}}$, are reset at $t_{\texttt{itv}}$ following Eqs. 4.8–4.9. Both SSMs are then run forward from $t_{\texttt{itv}}$ to the current step. 66

Figure 4.4 A simple and a complex synthetic time series generated using 2 different SSMs. The baselines and periodic components for them are the same. The residuals of the simple time series are generated with $\phi^{\texttt{AR}} = 0.5$, and those of the complex one are generated with $\phi^{\texttt{AR}} = 0.9$. 68

Figure 4.5 Examples of synthetic anomalies in different time series realizations generated by the SSM with $\phi^{\texttt{AR}} = 0.9$. The starting times of anomalies with varying magnitudes are indicated by the vertical dashed lines. The training, validation (val.), and test sets are marked on top of the time series. 68

Figure 4.6 Comparison of six detection models on simple and complex time series with anomalies with respect to time to detection (Δ_t), detection probability ($p^{\texttt{det}}$), and number of alarms ($\#_{\texttt{ALM}}$) across a range of anomaly magnitudes. The anomaly magnitude and the number of alarms ($\#_{\texttt{ALM}}$) are plotted on a logarithmic scale. Solid lines correspond to the expected value, and the shaded regions indicate the $\pm 1\sigma$ confidence region. 70

Figure 4.7 Normalized weekly time series of crack opening and displacement on dams, with lengths ranging from 9 to 18 years and containing missing data. The red region is the training set, the green one is the validation set, the yellow one corresponds to the anomaly labeled by dam engineers, and the grey one marks the potential anomalies corresponding to changepoints in the baseline responses. 71

Figure 4.8 Detrended stationary time series obtained by removing the baselines estimated by Prophet from the real time series. 72

Figure 4.9 Comparison of six detection models on detrended time series with anomalies, evaluated in terms of time to detection (Δ_t), detection probability ($p^{\texttt{det}}$), and number of alarms ($\#_{\texttt{ALM}}$). The results are averaged across all the abnormal time series derived from the 10 detrended time series. 74

Figure 4.10 Comparison of six methods against confirmed (yellow areas) and potential (grey areas) anomalies on real time series. Grey numbers indicate the ranking of their first valid detection on time series #1. The markers above the time series indicate the alarms triggered by different methods. 75

Figure 5.1 Illustration of the two limitations of the RSI method: (a) early intervention gradually diminishing before the true anomaly starting time, and (b) repeated false alarms triggered by an anomaly only shifting the level. 79

Figure 5.2 Overview of the RSIC procedure using an example where the anomaly is detected at time $t_{\texttt{det}}$, and a trend (T) intervention is identified at time t. 80

Figure 5.3 Tentative intervention estimation, where the BNN intervention $\boldsymbol{I}^{\tau}_{t_{\texttt{det}}}$ is refined via Kalman filtering and smoothing to obtain the prior state $\boldsymbol{X}^{\tau}_{t_{\texttt{det}}|t_{\texttt{det}}-1}$ used for model selection. 82

Figure 5.4 Procedure for classifying intervention type and applying a permanent intervention once a trend (T) intervention is identified at time t. . . . 83

Figure 5.5 Examples of synthetic time series with two types of anomalies: (i) a level anomaly (blue), and (ii) a trend anomaly (orange). The training and validation (val.) sets are identical across all time series, while the test sets correspond to different realizations of the same SSMs with varying anomaly magnitudes and starting times. 86

Figure 5.6 Time to detection (Δ_t) and detection probability ($p^{\texttt{det}}$) of the tested methods on synthetic time series under different anomaly combinations: level to level (L→L), level to trend (L→T), trend to level (T→L), and trend to trend (T→T). 88

Figure 5.7 Time to detection (Δ_t) and detection probability ($p^{\texttt{det}}$) of the tested methods on detrended time series under different anomaly combinations: level to level (L→L), level to trend (L→T), trend to level (T→L), and trend to trend (T→T). 92

Figure 5.8 Comparison of seven methods against confirmed (yellow areas) and potential (grey areas) anomalies on real time series. Grey numbers indicate the ranking of their first valid detection on time series #1. The markers above the time series indicate the alarms triggered by different methods. 95

Figure B.1 Eight time series measuring the cracking opening (C.O.) and displacement (Disp.) on a dam. They present missing data, different data acquisition frequencies, trends and periodic patterns. All the time series are assumed to be anomaly-free given that no structural deterioration was observed during these periods. 122

Figure C.1 BAR-γ and AR performance on synthetic time series A01C. Repetition N = 50. 123

Figure C.2 BAR-γ and AR performance on synthetic time series A16D. Repetition N = 50. 124

Figure C.3 BAR-γ and AR performance on synthetic time series M09C. Repetition N = 100. 124

Figure C.4 BAR-γ and AR performance on synthetic time series M11D. Repetition N = 100. 125

Figure C.5 BAR-γ and AR performance on synthetic time series M12D. Repetition N = 100. 125

Figure C.6 BAR-γ and AR performance on synthetic time series M13D. Repetition N = 100. 126

Figure C.7 BAR-γ and AR performance on synthetic time series M14D. Repetition N = 100. 126
Figure D.1 $F1t$ scores for BAR and AR verified on the synthetic datasets similar to eight real time series. Trend anomalies are overlaid on the synthetic time series. 127
Figure D.2 $F1t$ scores for BAR and AR verified on the synthetic datasets similar to eight real time series. Level anomalies are overlaid on the synthetic time series. 127
Figure E.1 From top to bottom, each row shows the performance of BAR-γ and AR on eight real time series of A01C ($\gamma = 2.8$), A16D ($\gamma = 2.0$), M08C ($\gamma = 0.4$), M09C ($\gamma = 0.4$), M11D ($\gamma = 0.6$), M12D ($\gamma = 0.7$), M13D ($\gamma = 0.9$) and M14D ($\gamma = 0.5$) involving three types of anomalies. . . 129
Figure H.1 Comparison of the RSI detection performance (detection time Δ_t, detection probability $p^{\texttt{det}}$, and number of alarms $\#_{\texttt{ALM}}$) across a range of anomaly magnitudes on validation time series for different (a) γ values and (b) BNN architectures. NN1 denotes the architecture used in all experiments (two fully connected layers with 32 and 64 units), NN2 denotes a single fully connected layer with 32 units, and NN3 denotes two fully connected layers with 256 units in each layer. 136
Figure I.1 Comparison of the prior distributions of the autoregressive (AR) hidden state and its autocorrelation function (ACF) before and after intervention on a non-stationary time series. The abbreviations "obs" and "anm" denote observations and anomalies, respectively. 137
Figure K.1 Time to detection Δ_t and the detection probability $p^{\texttt{det}}$ under different values of maximum waiting time $\tilde{t}$ across four anomaly configurations: level to level (L→L), level to trend (L→T), trend to level (T→L), and trend to trend (T→T). 140
Figure K.2 Time to detection Δ_t and the detection probability $p^{\texttt{det}}$ under different quantile values $\tilde{p}$ across four anomaly configurations: level to level (L→L), level to trend (L→T), trend to level (T→L), and trend to trend (T→T). 143

LIST OF SYMBOLS AND ACRONYMS

Symbols

a	Action of the reinforcement-learning agent or control action of the RSI and RSIC intervention framework
$\mathcal{A}$	Action set in the RSI and RSIC control framework
$\mathbf{A}$	Transition matrix
α	Standardized truncation bound or Beta-distribution hyperparameter
b	Boundary value of the bounded autoregressive component
$\mathcal{B}(\cdot)$	Beta distribution
β	Beta-distribution hyperparameter
$\mathcal{C}$	Model component
$\boldsymbol{c}_t$	TAGI-LSTM cell states at time t
γ	Boundary hyperparameter of the `BAR` component (γ^*: optimal value maximizing *F1t*)
$\mathcal{D}$	Observed data set
`d`	Drift tag referring to the drift model or hidden states
$d^{\texttt{anm}}$	Anomaly duration
`det`	Detection
Δ_t	Time to detection
$\mathbb{E}[\cdot]$	Expectation operator
`E`	Elongation measurement
$\mathbf{F}$	Observation matrix
$F^{-1}_{P^\tau,t}(\cdot)$	Quantile function of the classification variable P^τ_t
F1	F1 score
F1t	F1 temporal score
$f(\cdot)$	Probability density function
`FP`	False positive
`FN`	False negative
$\mathcal{H}$	Model structure (model class)
$\boldsymbol{h}_t$	TAGI-LSTM hidden states at time t
$\eta^{\texttt{anm}}_t$	Anomaly score at time t
$\hat{\eta}^{\texttt{anm}}$	Threshold of anomaly score for triggering alarms
i	Indices for time step or regime

$\boldsymbol{i}_t$	Input covariates at time t
$\boldsymbol{i}_{t_{\mathtt{itv}}}$	Intervention applied at the interventiom time $t_{\mathtt{itv}}$
$\mathtt{itv}$	Intervention
$\boldsymbol{I}_t^{\mathtt{L}},\ \boldsymbol{I}_t^{\mathtt{T}}$	Level or trend intervention vectors at time t
$I^{\mathtt{LcT}}$	Level correction caused by the accumulated change in trend
j	Indices for network layer or regime
$\mathbf{J}_t$	Rauch–Tung–Striebel smoother gain at time t
κ	Intermediate term of the mReLU
$\mathcal{L}$	Likelihood
$\mathtt{L}$	Level or the lower truncation bound
l_{dw}	Length of the target detection window
λ	Decay ratio penalizing the *F1* score or intermediate term of the mReLU
$\mathtt{LcT}$	Level compatible with trend
$\mathcal{M}$	Model
$\mathbf{M}_{t-1,t\|t}^{i,j}$	Joint regime-transition probability from regime i to j
$\boldsymbol{\mu}$	Mean vector
$\mathcal{N}$	Gaussian distribution
n	Number of hidden states
$\mathcal{O}(\cdot)$	Big-O computational-complexity notation
$\mathtt{O}$	LSTM output
p	Period of the periodic component, probability (detection $p^{\mathtt{det}}$, local-level intervention $p^{\mathtt{LL}}$, anomaly $p_{\mathtt{anm}}$), or quantile level (the p-quantile)
$\tilde{p}$	Quantile value for classification in RSIC
$\mathcal{P}$	Set of model parameters
P_t^{τ}	Beta-distributed variable for intervention type τ at time t
$\pi_{t\|t}^{i}$	Posterior probability of regime i at time t
q	Exponent setting the logarithmically-spaced input length ($= 2^q$)
$\mathbf{Q}$	Process-error covariance matrix
$\mathbf{R}$	Observation-error covariance matrix
$\boldsymbol{s}_t$	State of the reinforcement-learning agent at time t
$\mathtt{S}$	Number of regimes
$\mathcal{S}$	State set of the reinforcement-learning agent
$\sigma^{\mathtt{AR}},\ \sigma^{\mathtt{LT}}$	Process-error standard deviation of the autoregressive or local-trend component
$\sigma_W,\ \sigma_V$	Process-error or observation-error standard deviation
$\mathbf{\Sigma}$	Covariance matrix

t	Time index
$\tilde{t}$	Maximum waiting time after a detection
$\mathtt{T}$	Trend, air temperature measurement, or final time step
$\mathtt{TcA}$	Trend compatible with acceleration
$\mathtt{TP}$	True positive
τ	Intervention type in RSIC, $\tau \in \{\mathtt{LL}, \mathtt{LT}\}$, or training sample
$\boldsymbol{\theta}$	Neural-network parameters
$\mathtt{U}$	Upper truncation bound
$\boldsymbol{v}_t$	Observation error vector at time t
$\mathbf{W}^{i,j}_{t-1\|t}$	Regime-collapsing weight in the switching Kalman filter
w	Network weights
ϕ	Standard Gaussian probability density function
$\phi^{\mathtt{AR}}$	Autoregressive coefficient
$\phi_r(\cdot),\ \phi_r^{\pm}(\cdot)$	ReLU and mReLU activation function
$\Phi(\cdot)$	Standard Gaussian cumulative distribution function
$\boldsymbol{X},\ X$	Hidden-state random vector and variable
$\boldsymbol{x},\ x$	Realization of the hidden state
$\boldsymbol{Y},\ Y$	Predicted random vector and variable of the observation
$\boldsymbol{y},\ y$	Realization of the observation
$\boldsymbol{Z}^{(j)}$	Random vector of hidden units of layer j of the Bayesian neural network
z	Regime transition probability $z^{i,j}$, intervention-type pseudo-count indicator $(z^{\mathtt{LL}}, z^{\mathtt{LT}}, \hat{z}^{\tau})$, or neural-network hidden-unit realization $(z^{\mathtt{0}}, \boldsymbol{z}^{(j)})$
ω	Angular frequency of the periodic component or probability of the truncation Gaussian in mReLU

Acronyms

A	Acceleration
ACF	Autocorrelation Function
AGVI	Approximate Gaussian Variance Inference
AR	Autoregressive
ARIMA	Autoregressive Integrated Moving Average
AUC	Area Under the Curve
AUC-PR	Area Under the Precision–Recall Curve
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
BAR	Bounded Autoregressive

BDLM	Bayesian Dynamic Linear Model
BNN	Bayesian Neural Network
CNN	Convolutional Neural Network
DAMP	Discord Aware Matrix Profile
DID	Difference-in-Differences
FA	False Alarm
GMA	Gaussian Multiplicative Approximation
HST	Hydrostatic-Season-Time
KR	Kernel Regression
L	Level
LA	Local Acceleration
LL	Local Level
LSTM	Long Short-Term Memory
LSTMED	Long Short-Term Memory Encoder–Decoder
LT	Local Trend
MSE	Mean Squared Error
PCA	Principal Component Analysis
PD	Periodic (component)
PDF	Probability Density Function
PHM	Prognostics and Health Management
ReLU	Rectified Linear Unit
RL	Reinforcement Learning
ROC	Receiver Operating Characteristic
RSI	Residual Stationarity Intervention
RSIC	Residual Stationarity Intervention and Classification
RTS	Rauch–Tung–Striebel
SHM	Structural Health Monitoring
SKF	Switching Kalman Filter
SKF-AR	Switching Kalman Filter with Autoregressive component
SKF-BAR	Switching Kalman Filter with Bounded Autoregressive component
SSM	State-Space Model
T	Trend
TAGI	Tractable Approximate Gaussian Inference
TranAD	Transformer-based Anomaly Detection
VUS-PR	Volume Under the Surface (Precision–Recall)
VUS-ROC	Volume Under the Surface (Receiver Operating Characteristic)

LIST OF APPENDICES

Appendix A	SSM Matrices and Parameters for BAR	117
Appendix B	Overview of the Dam Datasets for Evaluating BAR Performance	122
Appendix C	BAR-γ Performance on Synthetic Time Series in Dam Dataset	123
Appendix D	*F1t* Score Summary on 8 Dam Time Series	127
Appendix E	BAR-γ Performance on Real Dam Dataset	128
Appendix F	SSM Matrices and Parameters for RSI	130
Appendix G	Experimental Setup for the RSI and RSIC Experiments	132
Appendix H	Hyperparameter Sensitivity in RSI	135
Appendix I	Residual Changes under RSI Intervention	137
Appendix J	Training Process of BNN in RSIC	138
Appendix K	Hyperparameter Sensitivity in RSIC	139

CHAPTER 1 Introduction

1.1 Motivation

Civil infrastructure such as bridges, dams, and roads underpins both the economic development and the safety of the people who rely on it. In Canada, the scale of this dependence is considerable: the country's core public transportation and water infrastructure was estimated to carry a replacement value of about $2.6 trillion [6]. When a structure fails, the consequences reach beyond economic loss. In 2006, a section of the De la Concorde overpass in Laval, Quebec, collapsed onto the highway below, killing five people and injuring six [7]. The commission of inquiry attributed the collapse to a chain of deficiencies and deterioration accumulated over more than 35 years [8]. Dam failures can be even more devastating. In 2019, the Brumadinho tailings dam in Brazil collapsed, killing more than 250 people [9]. Similarly to the De la Concorde overpass, this failure was the outcome of a slow, unnoticed deterioration over the years preceding the collapse [9].

Separating the deterioration of infrastructure from normal responses is complex. For instance, the response in concrete dams combines reversible loading effects, such as those induced by the water level in the reservoir and by the seasonal temperature [10], with irreversible material degradation such as the alkali-aggregate reaction, which causes progressive swelling and cracking of the concrete over decades [11]. More generally, deterioration is inevitable as structures age, and the interval between the moment it begins and the moment it becomes known is precisely where risk accumulates. *Structural Health Monitoring* (SHM) has emerged as the research field dedicated to closing this gap, by continuously providing information about the condition of a structure throughout its life [12].

This continuous stream of information is made possible by the sensors installed on the structure, whose measurements keep track of the structure's behavior. Deterioration and damage typically take the form of a gradual, irreversible change in the underlying baseline of the structural responses. Such a departure from the structure's normal response is referred to as an *anomaly*. The nature of an anomaly can vary from a shift in level, a change in trend, to other even more complex changes in kinematic form [13], each pointing to different potential deterioration mechanisms. Because civil structures are designed to remain in service for decades [14, 15], a structure will typically experience several such anomalies of different forms occurring at unknown times over its lifespan. Continuous monitoring of infrastructure requires 1) determining whether an anomaly has occurred, 2) identifying the nature of the anomaly, i.e., the form of the abnormal baseline change, and 3) detecting these anomalies

at an early stage, since early detection provides more time to react and more options for maintenance.

Early detection, however, cannot come at the expense of its reliability. A method that raises an alarm at the slightest fluctuation would catch every anomaly without delay, but it would also flag countless false alarms, drowning genuine abnormal events and leaving the monitoring system uninformative. A useful method must therefore balance its *detectability*, the ability to swiftly catch abnormal changes, against its *false alarm rate*, raising a clean alarm only when an anomaly is truly present [16]. Such false alarms can arise from two main sources: 1) mistaking a reversible variation for an irreversible baseline change; and 2) after the detection of an anomaly, continuing to flag as abnormal the baseline that has shifted to a *new stationary regime*, characterized by another constant kinematic form such as a level or a trend [13]. Avoiding the first requires separating genuine deterioration from the reversible fluctuations driven by environmental and operational conditions; avoiding the second requires adapting to the new stationary regime once a change is detected, so that subsequent observations and anomalies can be interpreted relative to it.

Beyond detectability and false alarms, computational cost is another important consideration, driven by two factors. The first concerns the growing length of the data. As noted above, civil structures are monitored continuously over decades, producing long time series on which retraining at each new observation is impractical. This points to *online* methods, such as state-space models (SSMs), which update their assessment recursively from the latest observation alone rather than reprocessing the entire past time series [17]. The second concerns the number of monitored structures. Infrastructure owners often manage large portfolios of continuously monitored assets, so even a modest computational cost per structure can accumulate into a substantial operational burden. For example, Hydro-Québec monitors over 400 large dams equipped with thousands of sensors, illustrating the scale at which anomaly detection methods may need to operate. The method must therefore remain computationally efficient to enable large-scale deployment.

Taken together, these considerations define the challenging setting in which long-term SHM must operate. Monitoring aging civil infrastructure calls for methods that can catch small baseline changes as they arise, without being misled by the far larger reversible fluctuations around them. Once a change is detected, the monitoring system must adapt to the new regime so that subsequent observations are interpreted correctly. All of this must happen at a computational cost compatible with processing ever-growing time series across large networks of structures. No existing method meets all of these demands at once, and closing this gap is what motivates this thesis.

1.2 Research Objectives

This thesis aims at developing online anomaly detection methods within the existing SSMs framework, enabling the early detection of small, slowly developing baseline changes in long-term SHM, while maintaining a low false alarm rate and a computational cost compatible with network-scale deployment. The core objectives of this thesis are:

- Develop a bounded autoregressive (BAR) component that enforces the stationarity of the residuals in SSMs in order to improve the detectability of small baseline anomalies.
- Formulate an intervention-based method that leverages the drift in the residuals to detect anomalies and accelerate the adaptation of SSMs to new regimes.
- Enhance the intervention framework with an online classification of anomaly types in order to apply the appropriate corrections and reduce false alarms under sequential anomalies.

For each of these objectives, the proposed methodologies are verified probabilistically on populations of synthetic time series and validated using real data recorded on full-scale structures.

1.3 Thesis Outline

The content of this thesis is organized as follows: Chapter 2 presents a literature review that introduces anomaly detection in the context of SHM and its challenges, existing time-series anomaly detection methods, as well as the evaluation metrics and criteria suited to SHM. It also details the theory behind SSMs and Kalman-based methods, with particular attention given to the autoregressive residual component in SSMs and to existing intervention techniques. Chapter 3 introduces the BAR component, which constrains the residual hidden state within the bounds of its stationary distribution in order to improve the switching Kalman filter's detectability of small anomalies. Chapter 4 presents the residual stationarity intervention (RSI) method, which detects anomalies by monitoring the stationarity of the residuals and applies interventions predicted by a Bayesian neural network to accelerate the adaptation of SSMs to new regimes. Chapter 5 extends this framework with the residual stationarity intervention and classification (RSIC) method, which infers the anomaly type online and applies the corresponding targeted intervention, addressing the detection of anomalies of different forms. For each chapter, the proposed methodologies are verified on synthetic time series and validated on real ones measured on infrastructure. Finally, Chapter 6 provides overall conclusions, the limitations of this work, and directions for future research.

1.4 Co-Authored Papers

The list of co-authored journal articles that are part of this thesis is:

- Z. Xin and J.-A. Goulet, "Enhancing structural anomaly detection using a bounded autoregressive component," *Mechanical Systems and Signal Processing*, vol. 212, p. 111279, 2024.
- Z. Xin and J.-A. Goulet, "Time series anomaly detection with residuals stationarity intervention on state-space models," *Advanced Engineering Informatics*, vol. 72, p. 104485, 2026.
- Z. Xin and J.-A. Goulet, "Anomaly detection in state-space models via residual stationarity intervention and classification," *Advanced Engineering Informatics*, submitted, 2026.

The list of conference papers is:

- Z. Xin and J.-A. Goulet, "Anomaly detection with switching Kalman filter and imitation learning," in *Proceedings of the 12th International Conference on Bridge Maintenance, Safety and Management (IABMAS 2024)*, Copenhagen, Denmark, 2024.

The list of presentations and posters is:

- Z. Xin, "Changepoint detection in time series using reinforcement learning," *"Stats in a Flash" 180-second Thesis Competition*, CANSSI Quebec, Concordia University, Montreal, Canada, March 2024.
- Z. Xin and J.-A. Goulet, "Kalman-based anomaly detection and intervention via imitation learning," *São Paulo School of Advanced Science on Structural Safety*, São Paulo, Brazil, May 2025.
- Z. Xin and J.-A. Goulet, "Online changepoint detections in state-space models through residual stationarity monitoring," *46th International Symposium on Forecasting*, Montreal, Canada, June–July 2026.

CHAPTER 2 Literature Review

This chapter reviews the theoretical foundations and existing methods relevant to anomaly detection in long-term Structural Health Monitoring (SHM). Section 2.1 introduces SHM, the types of anomalies encountered in structural responses, and the key challenges associated with their detection. Section 2.2 surveys existing time-series anomaly detection methods. Section 2.3 reviews state-space models and Kalman-based methods. Section 2.4 discusses autoregressive residual modeling and its impact on anomaly detection performance. Section 2.5 presents existing intervention analysis and regime-adaptation techniques. Section 2.6 discusses the standard evaluation metrics in the SHM context.

2.1 Structural Health Monitoring and Anomaly Detection

Section 2.1.1 first reviews the state of aging infrastructure and its long-term monitoring. Section 2.1.2 then surveys the taxonomy of anomaly types encountered in structural responses. Finally, Section 2.1.3 synthesizes the key challenges that persist in the field.

2.1.1 Aging Infrastructure and Continuous Monitoring

The safety and integrity of civil infrastructure are vital for both public safety and the economic development of society. Structural Health Monitoring (SHM) aims at providing continuous, automated assessment of the condition of structures such as bridges, dams, and buildings over their operational lifespan [12]. In many countries, infrastructure is operating beyond its originally intended service life, and the resources available for inspection and rehabilitation are insufficient to address the full extent of the deterioration [18]. In the United States alone, as illustrated in Figure 2.1, 51% of the bridges have passed or are approaching 50 years of service life, and 70% of the dams now exceed that same threshold [1].

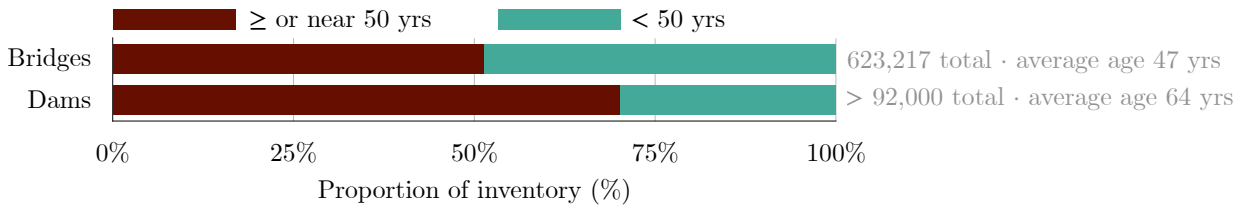


Figure 2.1 Proportion of U.S. bridges and dams at or approaching 50 years of service life (2025). Adapted from the American Society of Civil Engineers [1].

The widening gap between the amount of aging infrastructure and the capacity to inspect it exposes the limits of the methods used to assess structural condition. Traditional monitoring approaches rely primarily on visual inspections [19] and physics-based assessments, such as finite element modeling [20]. While these methods provide valuable structural insight, they are labor-intensive, hard to scale, and unsuited to continuous monitoring. Visual inspection is subjective and can miss subtle subsurface defects, while the accuracy of physics-based assessments depends on a careful calibration against experimental test data [21].

Data-driven SHM addresses these limitations by deploying networks of sensors that continuously record physical quantities such as strains, displacements, and crack openings [22]. The resulting time series, spanning from months to decades, capture both reversible effects, such as environmental conditions and operational loads [23], and irreversible changes associated with structural degradation [24]. Data-driven SHM enables continuous monitoring: rather than capturing a structure's condition only at isolated moments, it follows the structure continuously, so that slow changes become visible as they develop instead of being noticed only during the following inspections. It also relies on the measurements themselves rather than on subjective judgment or other calibration of physical models. Finally, because modern sensors are inexpensive, the approach can be extended across large infrastructure networks at a scale difficult to achieve with physics-based models [25].

2.1.2 Anomaly Types in SHM

Anomalies in time series are broadly categorized as *point-based* or *sequence-based* [2]. Point-based anomalies refer to isolated observations that deviate sharply from the surrounding data, and are further categorized into *global outliers*, which are abnormal with respect to the full time series, and *contextual outliers*, which are abnormal only within a local temporal context. Sequence-based anomalies, by contrast, correspond to contiguous segments of data that exhibit a pattern inconsistent with the historical behavior of the series [3]. These anomaly types are illustrated in Figure 2.2.

In the context of SHM, point-based anomalies are most commonly associated with sensor malfunctions or transient external disturbances rather than with genuine structural degradation [26]. The deterioration processes of civil infrastructure typically develop over periods ranging from months to years, manifesting as gradual, irreversible changes in the structural baseline rather than as instantaneous deviations [13]. Consequently, the anomalies of primary interest in SHM are sequence-based, and specifically those that alter the long-term irreversible baseline of a time series.

Three types of baseline anomalies are of particular relevance. A *level* (`L`) anomaly corresponds

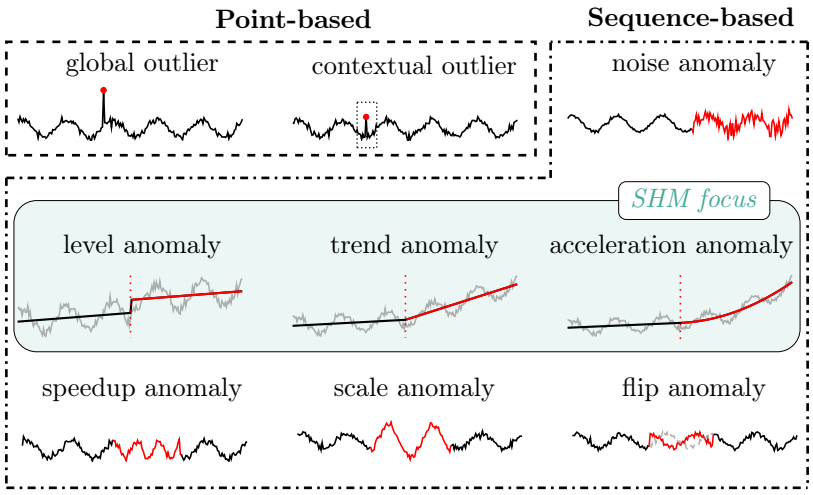


Figure 2.2 Taxonomy of nine time-series anomaly types, grouped into point-based and sequence-based categories. Global, contextual, noise, speedup, scale, and flip anomalies are established in the literature [2, 3]. The highlighted baseline-changed types, especially level and trend anomalies, are the focus of this thesis, as they represent potential structural anomalies.

to a sudden, permanent shift in the level of the time series, reflecting an abrupt structural change such as a settlement event or a sudden loss of stiffness. A *trend* (`T`) anomaly corresponds to a permanent change in the trend of the series, capturing progressive deterioration processes that continuously accumulate over time. An *acceleration* (`A`) anomaly reflects a change from a constant-speed regime to a constant-acceleration regime, indicative of accelerating degradation. The level and trend anomalies are the focus of this thesis because they are the simplest forms of baseline change and of primary interest in continuous SHM [13, 27].

The baseline anomalies can be described more formally in terms of *regimes* and *changepoints*. A *regime* is a segment of the time series over which the baseline follows a constant behavior, characterized by a fixed level, trend, or acceleration. Each baseline anomaly type corresponds to a transition between two such regimes. The time step at which this transition occurs is the *changepoint*, which marks the starting time of the *abnormal region* with respect to the preceding one. The abnormal region, considered on its own, constitutes a *new stationary regime* governed by its own constant baseline. This is illustrated in Figure 2.3, where the series evolves under a first trend-stationary regime until the changepoint, after which it transitions to a second trend-stationary regime with a different trend, corresponding to an trend anomaly. The second regime is abnormal with respect to the first because it carries a different trend, yet on its own its trend remains constant, so it forms a new stationary regime. These terms are reused throughout the thesis, where our goal is to detect the abnormal region

after the changepoint and to adapt to the new stationary regime once the detection is made.

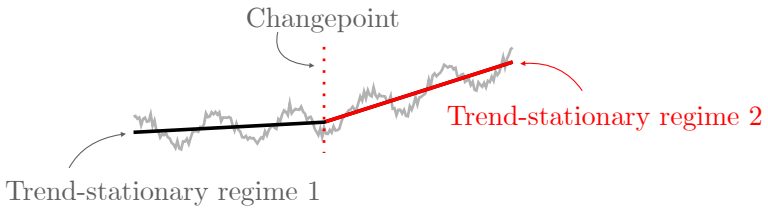


Figure 2.3 A time series transitioning from one trend-stationary regime to another at a changepoint. The segment after the changepoint is abnormal with respect to the first regime, yet on its own forms a new stationary regime.

2.1.3 Key Challenges

Despite significant advances in sensing technology and data-driven modeling, several fundamental challenges persist in the reliable detection of structural anomalies from long-term SHM data.

Subtlety and Slow Development. The level and trend anomalies characterizing the degradation of infrastructure differ markedly in how easily they can be detected. A level anomaly produces an abrupt shift in the baseline and is therefore relatively easy to detect [16]. A trend anomaly, by contrast, gradually alters the baseline. When its magnitude is small, it is among the most challenging changes to detect in a time series [28]. As illustrated in Figure 2.4, in the early stage of the development of a trend anomaly, the small baseline change is visually indistinguishable from the normal structural response and is thus difficult to detect for most algorithms.

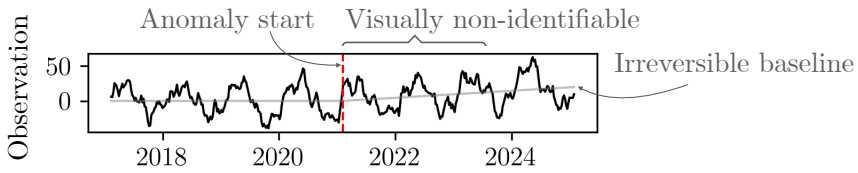


Figure 2.4 Illustration of the subtlety of a small trend anomaly. The gradual baseline change is introduced at the red dashed line, and its initial phase is indistinguishable from the normal structural response by visual inspection alone.

Environmental and Operational Variability. A reliable detection method must separate the irreversible baseline changes that signal deterioration from the reversible effects

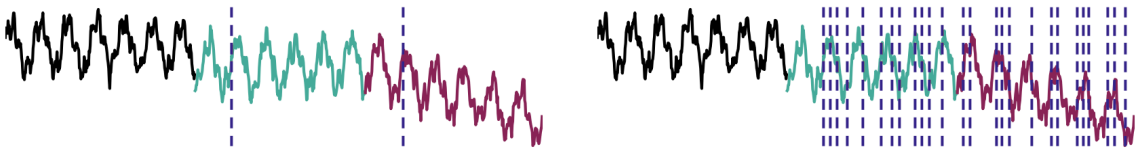
induced by environmental and operational conditions [29]. Structures are continuously influenced by factors such as temperature, humidity, and water level, whose effects are reversible [30]. A method that mistakes the recurring fluctuations they induce for genuine baseline changes will continuously trigger false alarms [31]. On the other hand, the magnitude of the reversible effects is often much larger than the changes in the baseline, so they can mask genuine structural deterioration and cause missed alarms.

Online Processing and Scalability. In operational SHM deployments, data are collected continuously. In order to be deployed at a large scale, it is important that monitoring systems process incoming observations in an online manner by updating their estimates incrementally as new data arrive without having to reprocess the full historical record [32]. Offline methods requiring retraining on the complete dataset become computationally prohibitive as decades of data accumulate [33]. Furthermore, infrastructure networks typically comprise a large number of structures, each instrumented with multiple sensors. For example, in Canada, Hydro-Québec owns over 400 large dams that are instrumented with tens of thousands of sensors of various kinds. Scalable anomaly detection therefore requires methods with a low computational cost per time step, and that is independent of the time-series length, enabling simultaneous monitoring of thousands of time series in real time.

Sequential Anomalies and Long-Term Robustness. In practice, a structure is likely to experience multiple anomalies over its service life, whether from progressive material degradation, discrete damage events, or planned structural interventions [34]. After detecting an initial anomaly, a method must be able to adapt to the new regime and remain capable of detecting subsequent ones without triggering a continuous stream of false alarms [35]. As illustrated in Figure 2.5, a method capable of adapting to new regimes detects each anomaly cleanly (a), whereas a method that fails to adapt its baseline model after a detection would interpret the persistent deviation as a new anomaly at every subsequent time step, producing a continuous stream of false alarms and rendering the monitoring system uninformative (b). Balancing sensitivity to new anomalies with stability in new regimes is therefore a central challenge in long-term SHM.

2.2 Time-Series Anomaly Detection Methods

The detection of anomalies in time series has been studied extensively in the context of structural health monitoring (SHM), and in the closely related field of prognostics and health management (PHM) [36], which extends beyond civil infrastructure to mechanical systems



(a) Clean detection (b) Excessive false alarms

Figure 2.5 Detection (marked by blue dashed lines) on a long-term time series containing two sequential anomalies: a change in level (teal) followed by a change in trend (maroon).

and equipment across aerospace, manufacturing, energy, transportation, and healthcare. In these contexts, anomaly detection enables the identification of outliers caused by sensor faults [37], operational activities [38], and, most importantly, the extraction of the baseline irreversible degradation that reflects structural deterioration [13]. This section presents a review of existing time-series anomaly detection methods, organized into three categories: distance-based methods (Section 2.2.1), prediction-based methods (Section 2.2.2), and deep-learning methods (Section 2.2.3).

2.2.1 Distance-based Methods

Distance-based methods [39], detect anomalies by measuring the distance between a target segment of the time series and a reference set of past observations, without relying on an explicit generative model of the normal behavior [40]. Their main advantage is that they do not require a parametric model of the structural response, making them applicable to time series whose underlying dynamics are difficult to specify. However, because they do not explicitly separate the structural baseline from environmental and periodic effects, the distances between subsequences reflect a mixture of structural changes, seasonal variations, and environmental fluctuations, making it difficult to attribute a large distance unambiguously to a genuine structural anomaly. Moreover, the absence of an underlying model limits these methods to the task of detecting anomalies, without the ability to decompose the signal or forecast future structural conditions, both of which are essential for decision-making in long-term SHM.

A representative method is the k-nearest neighbors (k-NN), which represents the time series as a collection of overlapping subsequences and measures, for each subsequence, its average distance to its k nearest neighbors in a reference set constructed from past observations [41]. A subsequence is flagged as abnormal when this average distance exceeds a specified threshold. In SHM, Sarmadi and Karamodin [42] coupled the Mahalanobis distance with a one-class k-NN to estimate the nearest neighbors of a given data point and remove the variation

caused by environmental effects, labeling observations far from their nearest neighbor as anomalies. Although effective under varying environmental conditions, the approach is sensitive to hyperparameter selection and requires a large amount of data. Other distance-based formulations have been used to quantify damage severity, for instance through a damage index computed from the dissimilarity between the probability distributions of transmissibility functions [43], and to identify structural components that respond differently to traffic input through spectral-based clustering on a large-scale bridge [44]. While these methods can mitigate environmental effects and identify outliers when the training data are not strongly imbalanced, they do not separate the variation caused by environmental effects from the meaningful time-series baseline, whose changes reflect structural deterioration or damage. In the following two paragraphs, we review in detail two prominent distance-based methods, Matrix Profile and Discord Aware Matrix Profile, which serve as comparison baselines in the following chapters.

Matrix Profile. Matrix Profile [45] stores, for each subsequence in a time series, the z-normalized Euclidean distance to its nearest neighbor, together with the index of that neighbor. For anomaly detection, subsequences with unusually large nearest-neighbor distances are identified as discords. The rationale is that if a subsequence's closest match remains highly dissimilar, the subsequence is likely abnormal. The method is computationally efficient and does not require a training phase, making it straightforward to apply. However, it presents two notable limitations in the context of SHM. First, it is designed primarily for offline operation, where the full time series is available to construct the pairwise distance structure. Second, if an abnormal pattern recurs more than once, which is likely in long-term SHM, each occurrence may find another abnormal occurrence as its nearest neighbor, causing the anomaly to be missed [46].

Discord Aware Matrix Profile. Discord Aware Matrix Profile (DAMP) [47] extends Matrix Profile to an online setting by computing left-discords at each time step, comparing the current subsequence only against past observations. An improved iterative searching and pruning technique contributes to the scalability of DAMP for long time series or high sampling frequencies. Rather than sequentially detecting emerging regime switches, DAMP identifies the subsequences that are most globally dissimilar from previously seen patterns, which makes it fundamentally different from regime-switching detection methods. In practice, however, DAMP outputs a ranking of discords, and identifying the actual anomalies relies on a fixed top-k cutoff. Choosing k requires prior knowledge of the number of anomalies in the series, which is generally not available in long-term SHM context.

2.2.2 Prediction-Based Methods

Prediction-based methods [48], detect anomalies by quantifying the inconsistency between the observations and the behavior predicted by a model of the normal structural response. Their main advantages are that (1) the model makes the structure of the normal response explicit, and for model families such as regression models and state-space models, it can be decomposed into interpretable components, such as trend, seasonality, and regression effects, that carry direct engineering interpretability [49, 50]; and (2) it can typically be used generatively to simulate plausible realizations of the structural responses [51]. Their main limitation is that the detection performance depends on the quality of the underlying model. Constructing the underlying model requires a preliminary analysis of the time series to determine which components are appropriate, and the training procedure typically involves hyperparameters whose misspecification can degrade both the model fit and the anomaly detectability [17]. In the following, we review four representative prediction-based methods: regression-based forecasting models and Prophet, which operate offline, and the switching Kalman filter and Bayesian online changepoint detection, which operate online.

Regression-Based Forecasting Models. A broad class of prediction-based methods forecasts future observations from the time-series context and triggers alarms when observations fall outside the prediction's confidence intervals. Salazar et al. [52] used boosted regression trees to predict future observations with confidence intervals given the measurement history, and verified the approach on a finite element model and on data from a real dam. For hydraulic structures, the hydrostatic-season-time (HST) model [53] captures the interaction between hydrostatic forces, seasonal variations, and time-dependent behavior; Gamse et al. [10] employed Bayesian model class selection to update HST models online as new measurements arrive, allowing the influence of environmental variables such as temperature and water level to be modeled. Gamse et al. [54] further analyzed the residual periodicities using the Lomb-Scargle normalized periodogram, obtaining a slight statistical improvement compared to the optimal HST model. Turrisi et al. [55] applied a cointegration-based approach with a predefined threshold to a complex stadium structure. Although regression-based forecasting is commonly used in practice, it is highly sensitive to the predefined threshold and to outliers, and is widely known to be prone to false alarms and to require constant human supervision under changing conditions.

Prophet. Prophet [56] fits time series with a parametric regression model that incorporates trend, seasonality, and holiday effects. It is reviewed separately because it is a well-established

time-series anomaly detection method and is used for comparison in the experiments presented in this thesis. It partitions the time series into segments and estimates a separate piecewise linear trend for each, flagging anomalies when the change in trend between consecutive segments exceeds a threshold. Its interpretable structure and ease of use make it a strong baseline for offline tasks such as detrending, and it has been widely adopted in practice across domains such as business forecasting [56,57], energy [58,59], and public health [60,61]. However, Prophet is fundamentally an offline method, and adapting it to an online setting requires retraining the full regression model at each time step as new observations arrive, which becomes computationally prohibitive as data accumulate. This undermines its applicability to large-scale and real-time SHM deployments.

Switching Kalman Filter. The switching Kalman filter (SKF) [4] detects anomalies by estimating, at each time step, the posterior probability of the stationary and non-stationary models given the observations. At each time step, each possible regime transition is processed by a separate Kalman filter, and the resulting estimates are merged into a single posterior per regime by weighting them according to their joint likelihood and transition probability. An alarm is triggered when the probability of the non-stationary regime exceeds a specified threshold. The performance of the SKF depends on the quality of the underlying SSMs, which in turn relies on the tuning of their hyperparameters and on the training process for components that will be reviewed in the next sections, such as a Bayesian LSTM for seasonality (Section 2.3.2), or an autoregressive component for residuals (Section 2.4.1). A suboptimal selection of hyperparameters or poorly trained components can therefore degrade the detection performance of the SKF, whose limitations are further discussed in Section 2.3.1.

Bayesian Online Changepoint Detection. Bayesian online changepoint detection identifies anomalies by estimating the posterior distribution over a run length, defined as the number of time steps elapsed since the parameters of the underlying probabilistic distribution last changed [62]. At each time step, the posterior is updated recursively as new observations arrive, and an alarm is triggered when the posterior probability of a short run length, indicating a recent changepoint, exceeds a threshold. Because its underlying model is a simple probabilistic distribution, this method performs well on time series governed by fixed, stationary dynamics, but requires coupling with other models for complex ones [63]. Furthermore, maintaining the posterior distribution over all possible run lengths requires storing and updating a growing number of hypotheses at each time step, making the method computationally expensive for long time series.

2.2.3 Deep-Learning Methods

Deep-learning methods leverage deep network architectures to learn representations of normal behavior from training data [64, 65]. Two main paradigms are used: reconstruction-based methods, which train a generative model to reconstruct normal time-series segments and flag anomalies when the reconstruction error exceeds a threshold [66], and classification-based methods, which train a model to directly distinguish between normal and abnormal patterns [67]. Their main advantage is their capacity to capture complex, high-dimensional temporal dependencies from large volumes of training data without requiring explicit model specification [68].

Reconstruction-Based Methods. Within the reconstruction-based paradigm, a variety of architectures have been explored in SHM. Generative adversarial networks (GANs) have been used to generate time series of damaged structures from measurements on healthy ones, showing potential to recognize different damage types but with generation accuracy on unseen components too low to scale to complex structures [69]; other GAN frameworks avoid requiring prior structural information and detect anomalies accurately provided they are sufficiently distinct from the majority of patterns [70]. Autoencoder models trained on guided-wave data spanning regular, irregular, and damage variations have achieved a higher area under the receiver operating characteristic curve than principal component analysis (PCA) [71, 72]. Comparative studies of PCA-based reconstruction have shown that aggregating data to daily or multi-day frequencies improves the suppression of environmental and operational variations along with the sensitivity to damage [73], and that probabilistic PCA handles missing data better, detects smaller anomalies, and is promising for anomaly localization [74]. Hybrid pipelines have also been proposed, such as a Bi-LSTM model to identify anomalies caused by data acquisition or baseline drift followed by a conditional GAN to repair the abnormal data, with models trained on anomaly-free data achieving lower reconstruction error than those trained with data including anomalies [75]. Among the reconstruction-based methods, two architectures are used as baselines in this work:

(a) LSTM-encoder-decoder (LSTMED) [76] combines a LSTM network with an encoder-decoder architecture: the encoder compresses a sliding window of observations into a fixed-size latent representation, and the decoder reconstructs the original sequence from it. It is trained in an unsupervised manner to minimize the reconstruction error on the training time series, so that abnormal subsequences deviating from the patterns seen during training produce a higher reconstruction error at inference time, and an alarm is triggered when this error exceeds a threshold.

(b) TranAD [77] replaces the LSTM with a Transformer architecture, using attention-based encoders to capture broader temporal trends. It adopts a two-phase adversarial training procedure, where the second phase amplifies reconstruction errors to improve sensitivity to subtle deviations, and focus score-based self-conditioning is used to stabilize training and enable robust multi-modal feature extraction. As in LSTMED, an alarm is triggered based on the deviation between the reconstructed output and the observations.

Classification-Based Methods. The classification-based paradigm maps a segment of the time series directly to a predefined anomaly type. In SHM, this has been realized in several ways. Tang et al. [78] divided bridge measurements into segments and used a convolutional neural network (CNN) to predict an acquisition- or drift-related anomaly type for each segment. Zhang et al. [79] applied the support vector data description to distinguish sensor-fault anomalies, which are typically detected through local features, from extreme-load anomalies, which exhibit correlation across several sensors. Jian et al. [80] combined relative frequency histograms with a one-dimensional CNN to classify acquisition-related faults such as missing data, outliers, and noise. Classification-based methods face two recurring difficulties: (1) the tradeoff in segment length, where short segments capture noisy or meaningless patterns while long segments may contain multiple patterns and delay detection; (2) the scarcity of anomalies, which makes it laborious to collect and manually label a representative set of abnormal samples. This second difficulty can be alleviated by generating synthetic anomalies to augment the training set; however, if the synthetic anomalies are not representative of those encountered in practice, models may still fail to map unseen anomalies to the correct label.

Deep-learning methods can leverage large volumes of training data to capture complex temporal patterns that simpler models may fail to represent. In SHM applications, however, the amount of training data available from a single time series is typically limited, which reduces this advantage and often causes them to underperform relative to simpler methods [81]. A further limitation for long-term SHM is their inability to generalize to structural regimes on which they have not been trained [82]. After a first anomaly shifts the time series to a new regime, deep-learning methods may trigger a continuous stream of false alarms, rendering the monitoring system uninformative for detecting subsequent anomalies. This lack of adaptability makes deep-learning methods, in their current form, poorly suited for long-term SHM, where multiple sequential anomalies are expected.

2.2.4 Summary

The methods reviewed in this section span distance-based, prediction-based, and deep-learning paradigms, yet most are tailored to specific problems and validated only on synthetic data, laboratory experiments, or real datasets in which the anomalies investigated are far larger than those of practical interest in SHM. Distance-based methods identify outliers but generally do not separate environmental variation from the structural baseline, nor do they forecast future conditions. Many prediction-based methods provide interpretable decompositions but depend on appropriate model specification and can be sensitive to threshold choice and to outliers. Deep-learning methods capture complex dependencies but require large training sets and are difficult to adapt across the successive structural regimes expected over a long monitoring horizon. Across all three paradigms, existing methods fall short of being simultaneously scalable across SHM systems, sensitive to small irreversible baseline changes, and robust against false alarms. Addressing these shortcomings requires a framework, state-space models, which is reviewed in the next section.

2.3 State-Space Models for SHM

The methods developed in this thesis build on state-space models (SSMs), a scalable framework capable of quantifying uncertainty, processing data in an online manner, and separating irreversible structural changes from reversible external effects. Section 2.3.1 reviews the SSMs framework, the Kalman filter and other Kalman-based methods. Section 2.3.2 describes the component decomposition in SSMs used to extract interpretable structural information from time series.

2.3.1 State-Space Models and the Kalman-Based Methods

A state-space model (SSM) describes a dynamic system through a hidden state vector $\boldsymbol{x}$ and four model matrices [83]: the transition matrix $\mathbf{A}$, the observation matrix $\mathbf{F}$, the process-error covariance matrix $\mathbf{Q}$, and the observation error covariance $\mathbf{R}$. The transition equation describes how the hidden states evolve over time,

$$\boldsymbol{x}_t = \mathbf{A}\boldsymbol{x}_{t-1} + \boldsymbol{w}_t, \quad \mathbf{W} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}), \tag{2.1}$$

and the observation equation maps the hidden states to the observation,

$$\boldsymbol{y}_t = \mathbf{F}\boldsymbol{x}_t + \boldsymbol{v}_t, \quad \mathbf{V} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}), \tag{2.2}$$

where $\boldsymbol{w}_t$ and $\boldsymbol{v}_t$ are process and observation errors at time t. The framework to obtain these four $\{\mathbf{A}, \mathbf{F}, \mathbf{Q}, \mathbf{R}\}$ from a component-based decomposition is presented in Section 2.3.2.

Kalman Filter. The Kalman filter [84] estimates the hidden states recursively through a *prediction* and an *update step*, without requiring the storage of the full observation history.

Throughout this thesis, we adopt a compact notation for the hidden-state's moments, in which: 1) the subscript, e.g., $_{t|t-1}$ or $_{t|t}$, comprises two indices, where the first denotes the time step of the estimate and the second denotes the time step up to which observations are used; and 2) the superscript denotes either the hidden-state vector of random variable, such as $^{\boldsymbol{X}}$, or a specific component within it, for example $^{\mathtt{AR}}$ for the autoregressive component reviewed in Section 2.3.2.

The *prediction step* computes the prior distribution of the hidden states at time t using all observations up to $t-1$ following

$$\boldsymbol{\mu}^{\boldsymbol{X}}_{t|t-1} = \mathbf{A}\boldsymbol{\mu}^{\boldsymbol{X}}_{t-1|t-1}, \tag{2.3}$$

$$\boldsymbol{\Sigma}^{\boldsymbol{X}}_{t|t-1} = \mathbf{A}\boldsymbol{\Sigma}^{\boldsymbol{X}}_{t-1|t-1}\mathbf{A}^\intercal + \mathbf{Q}, \tag{2.4}$$

and the corresponding prior distribution for the predicted observation is

$$\boldsymbol{\mu}^{\boldsymbol{Y}}_{t|t-1} = \mathbf{F}\boldsymbol{\mu}^{\boldsymbol{X}}_{t|t-1}, \tag{2.5}$$

$$\boldsymbol{\Sigma}^{\boldsymbol{Y}}_{t|t-1} = \mathbf{F}\boldsymbol{\Sigma}^{\boldsymbol{X}}_{t|t-1}\mathbf{F}^\intercal + \mathbf{R}. \tag{2.6}$$

The *update step* corrects the prior using the incoming observation $\boldsymbol{y}_t$ according to

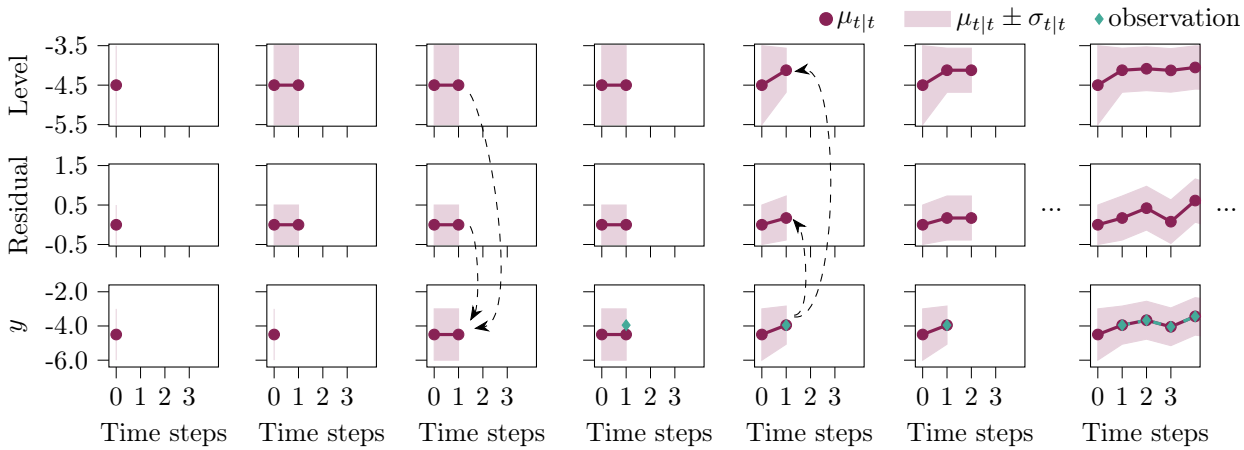
$$\boldsymbol{\mu}^{\boldsymbol{X}}_{t|t} = \boldsymbol{\mu}^{\boldsymbol{X}}_{t|t-1} + \boldsymbol{\Sigma}^{\boldsymbol{XY}}_{t|t-1}(\boldsymbol{\Sigma}^{\boldsymbol{Y}}_{t|t-1})^{-1}(\boldsymbol{y}_t - \boldsymbol{\mu}^{\boldsymbol{Y}}_{t|t-1}), \tag{2.7}$$

$$\boldsymbol{\Sigma}^{\boldsymbol{X}}_{t|t} = \boldsymbol{\Sigma}^{\boldsymbol{X}}_{t|t-1} - \boldsymbol{\Sigma}^{\boldsymbol{XY}}_{t|t-1}(\boldsymbol{\Sigma}^{\boldsymbol{Y}}_{t|t-1})^{-1}(\boldsymbol{\Sigma}^{\boldsymbol{XY}}_{t|t-1})^\intercal, \tag{2.8}$$

where the cross-covariance matrix is $\boldsymbol{\Sigma}^{\boldsymbol{XY}}_{t|t-1} = \boldsymbol{\Sigma}^{\boldsymbol{X}}_{t|t-1}\mathbf{F}^\intercal$. The posterior distribution $\boldsymbol{X}_{t|t} \sim \mathcal{N}(\boldsymbol{\mu}^{\boldsymbol{X}}_{t|t}, \boldsymbol{\Sigma}^{\boldsymbol{X}}_{t|t})$ is then carried forward to the next time step recursively.

The procedure of analyzing a time series with the Kalman filter is illustrated in Figure 2.6. The analysis begins with the prior knowledge of the observation and hidden states, including the level and the residual (Figure 2.6a). These states are transferred to the first time step using the transition model (Eqs. 2.3-2.4, Figure 2.6b), and aggregated through the observation model to produce a predicted observation (Eqs. 2.5-2.6, Figure 2.6c). The prediction is then compared with the true observation (Figure 2.6d), and the resulting difference is used to

update the state estimates (Eqs. 2.7-2.8, Figure 2.6e). The updated states are subsequently transferred to the next time step (Figure 2.6f), and the procedure is repeated recursively until the end of the series (Figure 2.6g).



(a) Init., t_0 (b) Trans. (c) Pred. (d) Obs. (e) Update (f) Trans. (g) ...

Figure 2.6 The process of the Kalman filter, showing the (a) initialization, (b, f) transition, (c) prediction, (d) observation, (e) update, and (g) recursion steps.

The Kalman filter operation from time step $t-1$ to t can be summarized as

$$\left(\boldsymbol{\mu}^{\boldsymbol{X}}_{t|t}, \boldsymbol{\Sigma}^{\boldsymbol{X}}_{t|t}, \mathcal{L}^{\boldsymbol{y}_t}_{t|t-1}\right) = \text{Filter}\left(\boldsymbol{\mu}^{\boldsymbol{X}}_{t-1|t-1}, \boldsymbol{\Sigma}^{\boldsymbol{X}}_{t-1|t-1}, \boldsymbol{y}_t; \mathbf{A}, \mathbf{F}, \mathbf{Q}, \mathbf{R}\right), \tag{2.9}$$

where $\mathcal{L}^{\boldsymbol{y}_t}_{t|t-1} = \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}^{\boldsymbol{Y}}_{t|t-1}, \boldsymbol{\Sigma}^{\boldsymbol{Y}}_{t|t-1})$, is the likelihood of observing $\boldsymbol{y}_t$ given the model predictions. The information flow of the Kalman filter from the initial time step, 0, until the last time step $\mathtt{T}$ is illustrated in Figure 2.7.

$$\begin{array}{lccccccccc} t & & 0 & & 1 & & t & & \mathtt{T}-1 & & \mathtt{T} \\ \text{Filter:} & & \begin{matrix}\boldsymbol{\mu}_0 \\ \boldsymbol{\Sigma}_0\end{matrix} & \rightarrow & \begin{matrix}\boldsymbol{\mu}_{1|1} \\ \boldsymbol{\Sigma}_{1|1}\end{matrix} & \rightarrow \cdots \rightarrow & \begin{matrix}\boldsymbol{\mu}_{t|t} \\ \boldsymbol{\Sigma}_{t|t}\end{matrix} & \rightarrow \cdots \rightarrow & \begin{matrix}\boldsymbol{\mu}_{\mathtt{T}-1|\mathtt{T}-1} \\ \boldsymbol{\Sigma}_{\mathtt{T}-1|\mathtt{T}-1}\end{matrix} & \rightarrow & \begin{matrix}\boldsymbol{\mu}_{\mathtt{T}|\mathtt{T}} \\ \boldsymbol{\Sigma}_{\mathtt{T}|\mathtt{T}}\end{matrix} \end{array}$$

Figure 2.7 Information flow of the Kalman filter in a time series from time step 0 to $\mathtt{T}$.

The Kalman filter is particularly well suited to the challenges of SHM outlined in Section 2.1. Its online, recursive structure satisfies the requirement for continuous, computationally efficient processing of streaming sensor data. Its Bayesian formulation provides uncertainty quantification for all hidden state estimates. Its interpretable decomposition of the observed

signal into physically meaningful components allows the separation of reversible environmental effects from irreversible structural changes. These properties have established Kalman filter as a primary framework for data-driven SHM [85–87].

Kalman Smoother. The Kalman filter produces estimates of the hidden states conditioned on all past observations up to the current time step t. When the full time series with length $\mathtt{T}$ is available, or when retrospective analysis is required, the Rauch-Tung-Striebel (RTS) smoother [88] enables refining the estimate at t by incorporating future observations up to $\mathtt{T}$. The information flow of the Kalman smoother is shown in Figure 2.8.

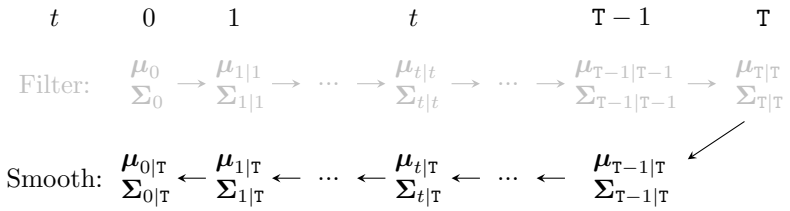


Figure 2.8 Information flow of the Kalman smoother in a time series from time step $\mathtt{T}$ to 0.

Starting from the final filtered state $\boldsymbol{X}_{\mathtt{T}|\mathtt{T}}$, it computes the smoothed posterior $\boldsymbol{X}_{t|\mathtt{T}}$ backward in time according to

$$\begin{aligned}\boldsymbol{\mu}^{\boldsymbol{X}}_{t|\mathtt{T}} &= \boldsymbol{\mu}^{\boldsymbol{X}}_{t|t} + \mathbf{J}_t(\boldsymbol{\mu}^{\boldsymbol{X}}_{t+1|\mathtt{T}} - \boldsymbol{\mu}^{\boldsymbol{X}}_{t+1|t}),\\ \boldsymbol{\Sigma}^{\boldsymbol{X}}_{t|\mathtt{T}} &= \boldsymbol{\Sigma}^{\boldsymbol{X}}_{t|t} + \mathbf{J}_t(\boldsymbol{\Sigma}^{\boldsymbol{X}}_{t+1|\mathtt{T}} - \boldsymbol{\Sigma}^{\boldsymbol{X}}_{t+1|t})\mathbf{J}_t^{\intercal},\end{aligned}$$

where the smoother gain is $\mathbf{J}_t = \boldsymbol{\Sigma}^{\boldsymbol{X}}_{t|t}\mathbf{A}^{\intercal}(\boldsymbol{\Sigma}^{\boldsymbol{X}}_{t+1|t})^{-1}$. The Kalman smoother operation from $t+1$ to t is summarized as

$$\left(\boldsymbol{\mu}^{\boldsymbol{X}}_{t|\mathtt{T}}, \boldsymbol{\Sigma}^{\boldsymbol{X}}_{t|\mathtt{T}}\right) = \text{Smooth}\left(\boldsymbol{\mu}^{\boldsymbol{X}}_{t+1|\mathtt{T}}, \boldsymbol{\Sigma}^{\boldsymbol{X}}_{t+1|\mathtt{T}}, \boldsymbol{\mu}^{\boldsymbol{X}}_{t|t}, \boldsymbol{\Sigma}^{\boldsymbol{X}}_{t|t}, \boldsymbol{\mu}^{\boldsymbol{X}}_{t+1|t}, \boldsymbol{\Sigma}^{\boldsymbol{X}}_{t+1|t}; \mathbf{A}\right).$$

In the context of SHM, the smoother is used to obtain refined estimates of hidden states from future observations.

Switching Kalman Filter. The Kalman filter and smoother are effective at extracting baselines and predicting future states in stationary time series, but cannot be applied to non-stationary time series where the observations transition between multiple regimes. Figure 2.9 illustrates a time series that switches from a level-stationary regime to a trend-stationary one, showing how the Kalman filter defined only by a constant trend model fails to capture the changes in trend and provides poor hidden-state estimation.

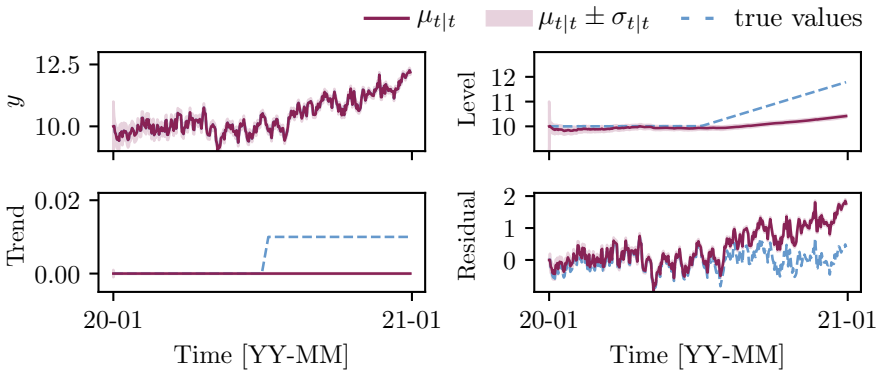


Figure 2.9 A time-series example that switches from a level-stationary regime to a trend-stationary one. The Kalman filter defined only by a constant trend model fails to estimate the hidden states after the regime switching.

The switching Kalman filter (SKF) [4] addresses this by modeling both a normal and an abnormal regime, and estimating, at each time step t, the posterior probability of each regime. Given the posterior hidden vector's mean $\boldsymbol{\mu}^i_{t-1|t-1}$ and its covariance $\boldsymbol{\Sigma}^i_{t-1|t-1}$ for the i-th regime, the SKF runs a separate Kalman filter for each possible transition from regime i to regime j,

$$\left(\boldsymbol{\mu}^{i,j}_{t|t}, \boldsymbol{\Sigma}^{i,j}_{t|t}, \mathcal{L}^{i,j}_{t|t-1}\right) = \text{Filter}\left(\boldsymbol{\mu}^i_{t-1|t-1}, \boldsymbol{\Sigma}^i_{t-1|t-1}, \boldsymbol{y}_t; \mathbf{A}^j, \mathbf{F}^j, \mathbf{Q}^j, \sigma^2_V\right),$$

yielding $\mathtt{S}^2$ filtered estimates at each time step, where $\mathtt{S}$ is the number of regimes. A merge step then collapses these estimates into a single posterior for each regime j by computing the joint probability

$$\begin{aligned}\mathbf{M}^{i,j}_{t-1,t|t} &= \Pr(S_{t-1} = i, S_t = j | \boldsymbol{y}_{1:t}), \\ &= \frac{\mathcal{L}^{i,j}_{t|t-1} \cdot z^{i,j} \cdot \pi^i_{t-1|t-1}}{\sum_i \sum_j \mathcal{L}^{i,j}_{t|t-1} \cdot z^{i,j} \cdot \pi^i_{t-1|t-1}},\end{aligned}$$

where $\pi^i_{t-1|t-1}$ is the probability for the i-th regime at time $t-1$, and $z^{i,j}$ is the transition probability from regime i to regime j. The regime probability at t is then obtained following

$$\pi^j_{t|t} = \sum_i \mathbf{M}^{i,j}_{t-1,t|t}.$$

The posterior mean and covariance for regime j are then estimated as

$$\boldsymbol{\mu}_{t|t}^{j} = \sum_{i=1}^{\mathtt{S}} \boldsymbol{\mu}_{t|t}^{i,j} \mathbf{W}_{t-1|t}^{i,j},$$
$$\boldsymbol{\Sigma}_{t|t}^{j} = \sum_{i=1}^{\mathtt{S}} \mathbf{W}_{t-1|t}^{i,j} (\boldsymbol{\Sigma}_{t|t}^{i,j} + \boldsymbol{\mu}\boldsymbol{\mu}^{\intercal}),$$

where $\mathbf{W}_{t-1|t}^{i,j} = \mathbf{M}_{t-1,t|t}^{i,j} / \pi_{t|t}^{j}$ and $\boldsymbol{\mu} = \boldsymbol{\mu}_{t|t}^{i,j} - \boldsymbol{\mu}_{t|t}^{j}$. The information flow of the SKF between two consecutive time steps is illustrated in Figure 2.10 for the case of two regimes ($\mathtt{S} = 2$), which results in $\mathtt{S}^2 = 4$ parallel Kalman filters followed by a merge step.

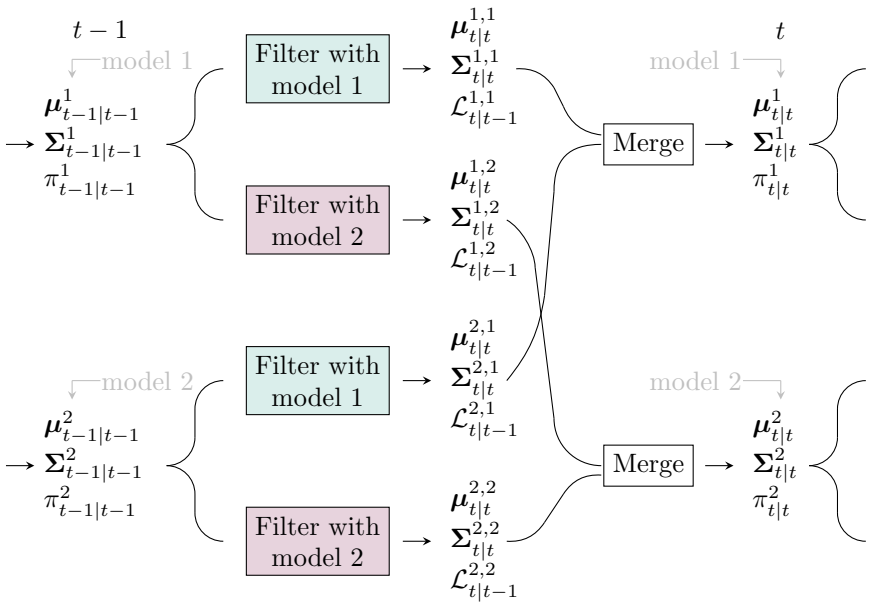


Figure 2.10 Information flow of the switching Kalman filter between two consecutive time steps, shown for two regimes ($\mathtt{S} = 2$). Adapted from Murphy [4].

In the SHM context, the SKF is typically configured with a stationary model representing the normal structural regime and a non-stationary model representing the abnormal regime, and an anomaly is flagged when the probability of the non-stationary regime exceeds a specified threshold [13]. Despite its appealing formulation, the SKF presents several practical limitations for long-term SHM. First, it requires the analyst to specify the non-stationary model in advance, which demands prior knowledge about the nature of future anomalies that is rarely available in practice. Second, when a component of the stationary model absorbs part of the anomaly-induced drift, the SKF may fail to detect a regime switch, resulting in delayed alarms. This limitation is illustrated in Figure 2.11, where the residual component absorbs part of the drift caused by a trend change, suppressing the anomaly probability

($p_{\texttt{anm}}$) and delaying the detection. Third, after a detection, the SKF typically requires a period ranging from several weeks to years to adapt its baseline to the new stationary regime, during which false or missed alarms may occur. These limitations motivate the development of the intervention-based methods presented in this thesis.

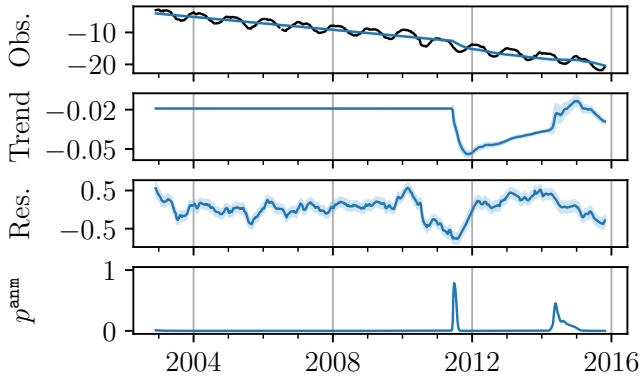


Figure 2.11 SKF detection delay on a real time series: part of the anomaly-induced drift in the observation (Obs.) is absorbed by the residual component (Res., third panel), suppressing the anomaly probability $p_{\texttt{anm}}$ (bottom panel) and delaying detection.

SKF Coupled with Reinforcement Learning. As discussed above, the SKF typically detects anomalies with a delay, because the anomaly-induced drift is partially absorbed by the components supposed to model the reversible behavior before the regime switch is identified. To reduce this delay, Khazaeli et al. [89,90] proposed to leverage the drift that appears in the hidden states to train anomaly detection agents that detect a regime switch earlier than the SKF alone. Two reinforcement learning (RL) paradigms were explored: tabular Q-learning and imitation learning.

Both paradigms share the same state and action space, derived from the SKF. The state is defined as $\boldsymbol{s}_t \in \mathcal{S} = \left\{\mu^{\texttt{T}}_{t|t}, \pi^{\texttt{ns}}_{t|t}\right\}$, where $\mu^{\texttt{T}}_{t|t}$ is the mean of the trend hidden state and $\pi^{\texttt{ns}}_{t|t}$ is the probability of the non-stationary regime at time t. The action is whether or not to trigger an alarm, $a_t \in \mathcal{A} = \{a^0 : \text{trigger an alarm},\ a^1 : \text{do not trigger an alarm}\}$. The two paradigms train agents with different signals obtained from synthetic time series with known anomaly positions.

In the tabular Q-learning [89], the agent learns from a reward defined according to manual anomaly labels, which gives credits to alarms triggered within an anomaly region and penalizes false or missed alarms. In the imitation learning [90], the agent instead learns directly from expert demonstrations, where each $\mathcal{D}_i = \left(\boldsymbol{s}_i, a_i^{\text{expert}}\right)$ pairs a state with the expert action $a_i^{\text{expert}} \in \mathcal{A}$ defined according to the manual anomaly label.

Figure 2.12 compares the detectability of the SKF, the tabular Q-learning, and the imitation learning agents, expressed as the region of anomaly magnitude $|h_a|$ and duration w_a for which each method reaches a true positive rate of at least 90%. Both RL-based agents enlarge this region relative to the SKF, detecting anomalies with smaller magnitude and longer duration, which confirms that exploiting the pre-detection drift enhances the SKF's detectability.

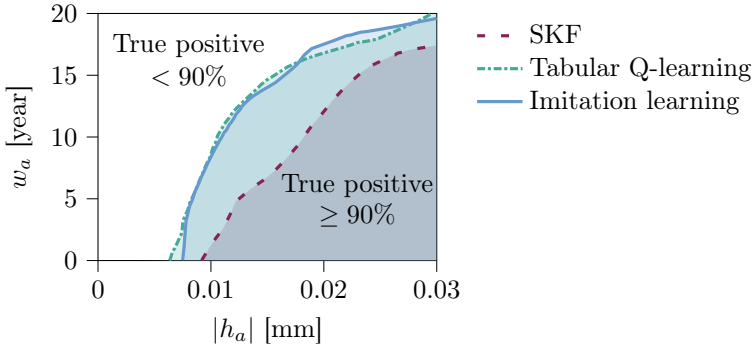


Figure 2.12 Detectability of the switching Kalman filter (SKF), the tabular Q-learning agent, and the imitation learning agent. Each curve marks the detection boundary as a function of the anomaly magnitude $|h_a|$ and duration w_a. Adapted from Khazaeli [5].

Although these learning-based agents can reduce the detection delay of the SKF, they share several limitations for long-term SHM. First, the two methods are limited to low-dimensional hidden states, and exploit neither the data history nor the drift in the residual component. Second, they are designed for time series containing a single anomaly and cannot handle the sequential anomalies that arise over a structure's service life. Third, the design of the RL environment requires manual feature engineering, which is difficult to scale across a SHM system. These limitations further motivate the intervention-based methods presented in this thesis.

2.3.2 Component Decomposition

A central advantage of SSMs in SHM is their ability to decompose a time series into interpretable components, each capturing a distinct feature of the structural response. This decomposition is achieved by modularly specifying the hidden state vector $\boldsymbol{x}$ and the model matrices, and assembling the full SSM as a block-diagonal combination of individual components. Three categories of components are distinguished: baseline components, periodic components, and residual components.

Baseline Components. Baseline components describe the underlying long-term tendency of the time series after excluding short-term fluctuations and reversible periodic patterns. The local level (LL) component models a stationary level, while the local trend (LT) component extends this by additionally modeling the rate of change of the level. The LT component is formulated as

$$\boldsymbol{x}^{\mathtt{LT}} = \begin{bmatrix} x^{\mathtt{L}} \\ x^{\mathtt{T}} \end{bmatrix}, \quad \mathbf{A}^{\mathtt{LT}} = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix}, \quad \mathbf{Q}^{\mathtt{LT}} = (\sigma^{\mathtt{LT}})^2 \begin{bmatrix} \frac{\Delta t^4}{4} & \frac{\Delta t^3}{2} \\ \frac{\Delta t^3}{2} & \Delta t^2 \end{bmatrix}, \quad \mathbf{F}^{\mathtt{LT}} = \begin{bmatrix} 1 & 0 \end{bmatrix},$$

where Δt is the time-step length and $\sigma^{\mathtt{LT}}$ is the process-error standard deviation of the LT component, controlling the non-stationarity of the trend. When $\sigma^{\mathtt{LT}} = 0$, the baseline evolves deterministically according to its initial level and trend. The local acceleration (LA) component further augments the hidden state with an acceleration term. Baseline components are directly associated with the anomaly types of interest in SHM: a sudden change in $x^{\mathtt{L}}$ corresponds to a level anomaly, and a change in $x^{\mathtt{T}}$ corresponds to a trend anomaly, as illustrated in Figure 2.2.

Periodic Components. Structural responses typically exhibit strong periodic patterns driven by seasonal environmental cycles. These patterns must be captured by the model; otherwise, they could contaminate other components and trigger false alarms. Harmonic periodic phenomena can be represented using Fourier forms with a period p and angular frequency $\omega = 2\pi\Delta t/p$,

$$\boldsymbol{x}^{\mathtt{PD}} = \begin{bmatrix} x^{\mathtt{PD1}} \\ x^{\mathtt{PD2}} \end{bmatrix}, \quad \mathbf{A}^{\mathtt{PD}} = \begin{bmatrix} \cos\omega & \sin\omega \\ -\sin\omega & \cos\omega \end{bmatrix}, \quad \mathbf{Q}^{\mathtt{PD}} = \left(\sigma^{\mathtt{PD}}\right)^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{F}^{\mathtt{PD}} = \begin{bmatrix} 1 & 0 \end{bmatrix}.$$

For non-harmonic periodic patterns, kernel regression [91] and, more recently, the analytically tractable Bayesian Long Short-Term Memory (LSTM) networks [92] provide more flexible representations. The LSTM-based recurrent component is particularly suited to long-term SHM, as it can capture complex, non-stationary seasonal patterns without requiring manual feature engineering. The corresponding hidden state vector, along with the transition and observation matrices for the LSTM component, is defined as

$$\boldsymbol{x}^{\mathtt{LSTM}} = \begin{bmatrix} z^{\mathtt{0}} \end{bmatrix}, \ \mathbf{A}^{\mathtt{LSTM}} = \begin{bmatrix} 1 \end{bmatrix}, \ \mathbf{Q}^{\mathtt{LSTM}} = \begin{bmatrix} (\sigma^{\mathtt{LSTM}})^2 \end{bmatrix}, \ \mathbf{F}^{\mathtt{LSTM}} = \begin{bmatrix} 1 \end{bmatrix},$$

where the pattern hidden state $z^{\mathtt{0}}$ is predicted not through the transition model in Eqs. 2.3-

2.4, but instead by a TAGI-LSTM network, summarized as

$$z^{\texttt{0}} = \text{TAGI-LSTM}(\boldsymbol{i}_t, \boldsymbol{h}_{t-1}, \boldsymbol{c}_{t-1}, \boldsymbol{\theta}_{t-1}),$$

where $\boldsymbol{i}_t$ denotes the input covariates, $\boldsymbol{h}_{t-1}$ and $\boldsymbol{c}_{t-1}$ represent the TAGI-LSTM hidden states and cell states, and $\boldsymbol{\theta}_{t-1}$ denotes the network parameters from the previous time step $t-1$.

Residual Components. After accounting for the baseline and periodic components, the remaining discrepancy between the model predictions and the observations is captured by a residual component. The simplest residual can be modeled by a white noise component, represented as a fixed Gaussian distribution with zero mean and constant variance. The Approximate Gaussian Variance Inference (AGVI) method [93] can be coupled with a TAGI-LSTM method to model heteroscedastic errors. When the residuals are correlated across consecutive time steps, a first-order autoregressive (AR) component is a standard choice,

$$\boldsymbol{x}^{\texttt{AR}} = \left[x^{\texttt{AR}}\right], \quad \mathbf{A}^{\texttt{AR}} = \left[\phi^{\texttt{AR}}\right], \quad \mathbf{Q}^{\texttt{AR}} = \left[(\sigma^{\texttt{AR}})^2\right], \quad \mathbf{F}^{\texttt{AR}} = \left[1\right],$$

where the autoregression coefficient $\phi^{\texttt{AR}} \in [0, 1)$ controls the temporal dependency between consecutive residuals, and $\sigma^{\texttt{AR}}$ is its process-error standard deviation. The stationary variance of the AR process is

$$\left(\sigma^{\texttt{AR},0}\right)^2 = \frac{\left(\sigma^{\texttt{AR}}\right)^2}{1 - \left(\phi^{\texttt{AR}}\right)^2}. \tag{2.10}$$

The constraint $\phi^{\texttt{AR}} \in [0, 1)$ ensures that the process remains stationary, meaning that its variance does not grow unbounded over time. When $\phi^{\texttt{AR}}$ approaches one, the AR component becomes increasingly persistent and prone to capturing trends that should be attributed to the baseline, a behavior that is particularly problematic for anomaly detection. The autoregressive residual component and its stationarity, which are central to this thesis, are reviewed further in Section 2.4.

Assembling Components into a Full Model. The individual components are combined into a single SSM by concatenating their hidden state vectors and arranging their matrices in block-diagonal form. For example, a model comprising a local trend, an LSTM, and an

autoregressive component is assembled as

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}^{\texttt{LT}} \\ \boldsymbol{x}^{\texttt{LSTM}} \\ \boldsymbol{x}^{\texttt{AR}} \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} \mathbf{A}^{\texttt{LT}} & 0 & 0 \\ 0 & \mathbf{A}^{\texttt{LSTM}} & 0 \\ 0 & 0 & \mathbf{A}^{\texttt{AR}} \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} \mathbf{Q}^{\texttt{LT}} & 0 & 0 \\ 0 & \mathbf{Q}^{\texttt{LSTM}} & 0 \\ 0 & 0 & \mathbf{Q}^{\texttt{AR}} \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} \mathbf{F}^{\texttt{LT}}, \mathbf{F}^{\texttt{LSTM}}, \mathbf{F}^{\texttt{AR}} \end{bmatrix}.$$

This modular structure allows the SSM to be tailored to the characteristics of each time series, while maintaining a unified Kalman filter inference procedure across all components.

2.4 Residual Modeling and Stationarity

Section 2.4.1 presents the autoregressive residual model and its online hyperparameter estimation, and Section 2.4.2 discusses how anomalies induce a gradual drift in the residuals that delays their detection.

2.4.1 Autoregressive Residual Model and Online Parameter Estimation

In SSMs, after accounting for the baseline and seasonality, the remaining discrepancy between the model prediction and the observation is modeled by a residual component, which includes both the observation and modeling errors. Prediction errors are typically correlated across time steps, because if the model underestimates the structural response at one time step, it is very likely to underestimate it again at the next one. An autoregressive (`AR`) model can represent such temporally correlated residuals, as it explicitly models the dependency between the current error and its recent history.

As reviewed in Section 2.3.2, the first-order `AR` component has two hyperparameters: the autoregression coefficient $\phi^{\texttt{AR}}$, which controls the temporal dependency between consecutive residuals, and the process-error standard deviation $\sigma^{\texttt{AR}}$. These hyperparameters are typically set to fixed priors or estimated offline by maximum likelihood [94]. Alternatively, they can be learned online, jointly with the other hidden states, which avoids the slow offline optimization.

Online Learning of $\phi^{\texttt{AR}}$ Using Gaussian Multiplicative Approximation. The autoregression coefficient $\phi^{\texttt{AR}}$ can be learned online [95] by augmenting it as a latent variable in the hidden state vector alongside $x^{\texttt{AR}}$, so that $\boldsymbol{X} = [X^{\phi}, X^{\texttt{AR}}]^{\intercal}$. The model matrices become

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} 0 & 0 \\ 0 & (\sigma^{\texttt{AR}})^2 \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} 0 & 1 \end{bmatrix},$$

where X^{ϕ} is treated as a time-varying random variable with no process-error, and only $X^{\texttt{AR}}$ is directly observable. However, the transition of $X^{\texttt{AR}}$ involves the product $X^{\phi}X^{\texttt{AR}}$, which is a nonlinear operation. The Gaussian Multiplicative Approximation (GMA) [95] addresses this by approximating the distribution of the product of two jointly Gaussian random variables as a Gaussian, propagating the first and second moments of the product analytically. Consider a Gaussian random vector

$$\boldsymbol{X} \sim \mathcal{N}\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}; \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} (\sigma_1)^2 & \text{cov}(X_1, X_2) \\ \text{cov}(X_2, X_1) & (\sigma_2)^2 \end{bmatrix}\right).$$

The GMA operation $\text{GMA}^{X_2}_{X_1X_2}(\boldsymbol{X})$ proceeds in two steps: it first forms the product X_1X_2, and then replaces the hidden state X_2 by this product. Here, the superscript X_2 denotes the hidden state being replaced, and the subscript X_1X_2 denotes the product that replaces it. Applying this operation yields

$$\text{GMA}^{X_2}_{X_1X_2}(\boldsymbol{X}) \sim \mathcal{N}\left(\begin{bmatrix} x_1 \\ x_1x_2 \end{bmatrix}; \begin{bmatrix} \mu_1 \\ \mathbb{E}(X_1X_2) \end{bmatrix}, \begin{bmatrix} (\sigma_1)^2 & \text{cov}(X_1, X_1X_2) \\ \text{cov}(X_1X_2, X_1) & (\sigma_{X_1X_2})^2 \end{bmatrix}\right),$$

where Deka et al. [95] provide closed-form equations for the covariance between X_1 and the product term, $\text{cov}(X_1, X_1X_2)$, and the variance of the product term, $(\sigma_{X_1X_2})^2$.

The transition equation 2.1 of the Kalman filter is then modified to

$$\boldsymbol{x}_t = \mathbf{A}\tilde{\boldsymbol{x}}_{t-1} + \boldsymbol{w}_t,\ \mathbf{W} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}),$$

where $\tilde{\boldsymbol{x}}_{t-1}$ is the realization from the new posterior $\tilde{\boldsymbol{X}}_{t-1|t-1} \sim \mathcal{N}(\boldsymbol{\mu}^{\tilde{\boldsymbol{X}}}_{t-1|t-1}, \boldsymbol{\Sigma}^{\tilde{\boldsymbol{X}}}_{t-1|t-1})$ whose moments are obtained through $\text{GMA}^{X^{\texttt{AR}}}_{X^{\phi}X^{\texttt{AR}}}(\boldsymbol{X}_{t-1|t-1})$. The update step follows the standard Kalman filter equations. This allows $\phi^{\texttt{AR}}$ to be inferred jointly with $x^{\texttt{AR}}$ at each time step as new observations arrive, without requiring any offline calibration.

Online Learning of $\sigma^{\texttt{AR}}$ Using Approximate Gaussian Variance Inference. The process-error standard deviation $\sigma^{\texttt{AR}}$ can also be learned online using Approximate Gaussian Variance Inference (AGVI) [93], which extends the GMA framework to treat the process-error variance as another Gaussian latent variable. Together with the GMA-based learning of $\phi^{\texttt{AR}}$, this eliminates both hyperparameters of the `AR` component, making the residual model fully adaptive without requiring offline optimization. Figure 2.13 illustrates this on a synthetic dataset. Given the observations shown in the top panel, the GMA and AGVI methods jointly infer the autoregression coefficient $\phi^{\texttt{AR}}$ (middle panel) and the process-error variance $(\sigma^{\texttt{AR}})^2$

(bottom panel) online. Both estimates start from their initial priors and converge toward the true values within the first few years, with the confidence region shrinking as more data is processed.

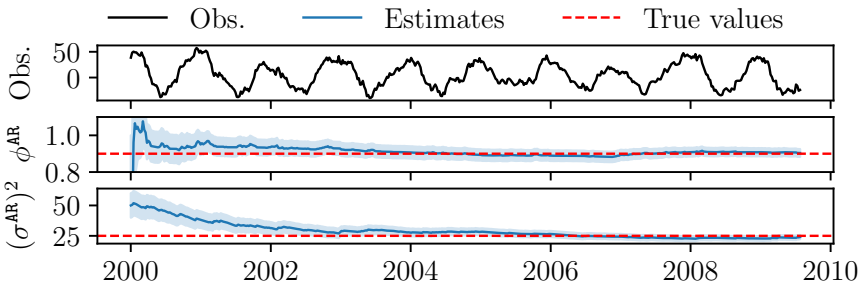


Figure 2.13 Online learning of the `AR` hyperparameters $\phi^{\mathtt{AR}}$ and $(\sigma^{\mathtt{AR}})^2$ on synthetic data. Estimates (blue, ± one standard deviation) converge to the true values (red dashed) without offline calibration.

2.4.2 Residual's Drift and its Partial Absorption

A key limitation of the `AR` component in SSMs is its tendency to drift away from its stationary range when the time series undergoes a change, as illustrated in Figure 2.11. This drift, which should be attributed to the irreversible baseline, delays the detection of the anomaly. The `AR` component becomes more prone to assimilate such drift if $\phi^{\mathtt{AR}}$ has a value close to one, indicating a high degree of temporal dependency and an increased complexity in the residuals. A value of $\phi^{\mathtt{AR}}$ close to one can arise from two main sources:

(a) *Modeling errors*: When the baseline or periodic components are not correctly defined, the patterns normally captured by these components may be partially attributed to the `AR` component and eventually lead to a $\phi^{\mathtt{AR}}$ value close to one, indicating the presence of strong dependencies between residual values;

(b) *Small time steps*: When the time-step length is short, e.g., one hour, consecutive model predictions are likely to be similar. The model errors describing the discrepancy between the model predictions and the structural states are thus highly correlated. In order to capture such a high correlation, $\phi^{\mathtt{AR}}$ typically approaches one. On the other hand, with long time-step lengths, e.g., one month, the model errors become mostly independent from each other and $\phi^{\mathtt{AR}}$ tends to be close to zero.

Nevertheless, even when $\phi^{\mathtt{AR}}$ approaches one, the `AR` component can only partially absorb the anomaly-induced drift. An example of such a *partial residual absorption* is shown in

Figure 2.14. The zero-valued trend in the time series (black) shifts to a positive trend after the red dashed line. This change in trend has a negligible impact on the time series during the first six months, when the prior estimates of the trend and the LSTM remain almost identical to those in the stationary period (before 2021). A small drift appears in the residual beginning in mid-2021 but soon returns to the stationary range by the end of the same year. The drift is not distinguishable from the stationary period until mid-2022.

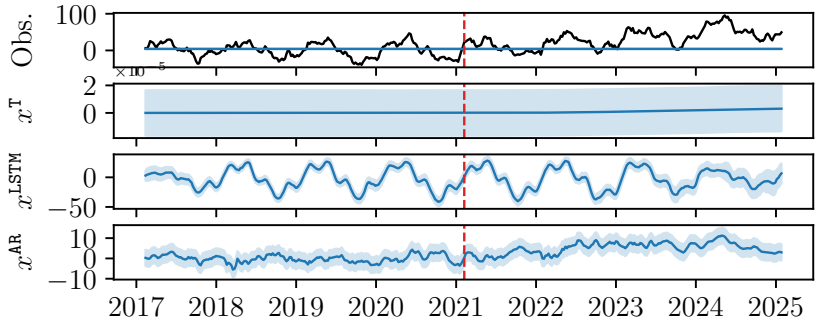


Figure 2.14 Delay in the residual estimation when the time series exhibits a non-stationary trend.

Three main reasons account for such a partial residual absorption in non-stationary time series:

(a) *Small initial changes*: When an anomaly first develops, the induced change in the trend is small enough that its contribution to the residuals remains negligible and indistinguishable from stationary fluctuations;

(b) *Baseline absorption by the LSTM*: When the model includes an LSTM component, the LSTM can accommodate minor to moderate baseline changes within its predictions before it eventually fails to generalize, partially masking the emerging drift from the `AR` component;

(c) *Limited transition rate*: The transition from $x_{t-1}^{\mathtt{AR}}$ to $x_t^{\mathtt{AR}}$ being limited by the autoregressive coefficient $\phi^{\mathtt{AR}} \in [0, 1)$, where smaller values of $\phi^{\mathtt{AR}}$ require longer time to capture a drift.

Together, these effects explain why the residual's drift induced by a structural anomaly develops gradually and remains difficult to distinguish from stationary fluctuations in the early stages of anomaly development, resulting in detection delays that are particularly problematic for slow-developing structural deteriorations.

2.5 Intervention Models and Regime Adaptation

Section 2.5.1 reviews the intervention analysis framework and its limitations, and Section 2.5.2 presents neural-network based approaches for predicting interventions in order to correct models and to adapt to non-stationary situations.

2.5.1 Intervention Analysis

In SHM, identifying the presence of an anomaly is only the first step: once a change is detected, the model must be updated and adapted to the new regime so that monitoring can continue without generating persistent false alarms [35]. This adaptation, referred to as intervention, consists of correcting the hidden states of the model to account for the effect of the anomaly [96]. It is particularly critical in long-term SHM, where structures are subjected to multiple sequential anomalies over their service life, and where the ability to detect subsequent events depends directly on how well the model has adapted itself to each preceding one.

Shifts in the level and trend caused by external factors occurring at known times have been studied within the *intervention analysis* framework introduced by Box and Tiao [97]. In this framework, deterministic transfer functions are used to represent structural changes in the data with parametric functions, such as intervention magnitude, whose parameters are estimated offline via maximum likelihood. This approach is commonly applied to Autoregressive Integrated Moving Average (ARIMA) models [98]. Li et al. [99] further summarized two complementary statistical approaches for estimating intervention effects: the *Difference-in-Differences* (DID) model, which estimates the intervention magnitude by comparing pre- and post-intervention differences between a treated and a control group [100], and *segmented regression*, which fits separate regression models over predefined segments of the time series [101]. Beyond these offline approaches, several online intervention methods have been proposed, including Gaussian-Process State-Space Models (GPSSMs) [102] and Kalman-based techniques [103–106].

Despite their variety, existing intervention-analysis methods generally rely on counterfactual reasoning, where the intervention magnitude is estimated by comparing pre-intervention model predictions with the abnormal observations, under the assumption that the intervention time and number are known a priori. Such assumptions are rarely satisfied in online SHM applications, where anomalies emerge gradually and their starting times are typically unknown, motivating the development of methods capable of predicting the necessary intervention without prior knowledge of when or how many anomalies have occurred.

2.5.2 Neural-Network Based Intervention Predictors

Alternatively, rather than being estimated through counterfactual reasoning as it is the case in the classical intervention analysis, the interventions can be predicted directly from the observed time series using neural networks [107–109]. For example, Li et al. [108] formulate changepoint detection as a supervised classification problem and show that deep neural networks can match or outperform classical changepoint detectors under challenging noise conditions. However, adapting this idea to long-term SHM requires addressing two obstacles. The first is compatibility with the monitoring framework. The methods cited above [107–109] operate outside the SSMs framework, and their deterministic outputs cannot be supplied directly to the hidden states of an SSM. Because the hidden states in SSMs are modeled as Gaussian distributions, any predicted intervention must itself be expressed as a Gaussian variable to remain consistent with the Kalman filter steps reviewed in Section 2.3.1. This requirement points to *Bayesian Neural Networks* (BNNs), whose outputs are probability distributions rather than point estimates [110], and which can therefore deliver interventions directly to the hidden states of SSMs.

The second obstacle is the difficulty of the SHM mapping. Neural-network intervention predictors are most mature in clinical and biomedical contexts [107, 109, 111], where the relationship between an intervention type and its resulting pattern is well characterized and extensively studied. In SHM, by contrast, mapping a structural damage to a specific abnormal data pattern is considerably harder, because damages often manifest themselves gradually and subtly, the structural responses are highly specific to each structure's geometry, materials, and environmental variations frequently produce larger response changes than the damage itself [30, 112]. A successful predictor must therefore learn this mapping under weak supervision while remaining robust to reversible environmental effects. Under these conditions the predicted correction is inevitably uncertain, and because an intervention modifies the hidden states directly, a confident but erroneous correction would cause false alarms and compromise the detection of subsequent anomalies [113]. Quantifying the uncertainty of the predicted intervention would allow the Kalman update to weight the correction by its reliability, which further motivates the use of a BNN over a neural network producing a single deterministic estimate.

Bayesian Neural Networks and Tractable Approximate Gaussian Inference. In a BNN, the network parameters $\boldsymbol{\theta}$ are treated as random variables rather than fixed weights, and inference is performed by computing the posterior distribution over $\boldsymbol{\theta}$ given the observed

data $\mathcal{D}$ and the model structure $\mathcal{H}$ [114],

$$P(\boldsymbol{\theta}|\mathcal{D},\mathcal{H}) = \frac{P(\mathcal{D}|\boldsymbol{\theta},\mathcal{H}) \cdot P(\boldsymbol{\theta}|\mathcal{H})}{P(\mathcal{D}|\mathcal{H})},$$

which is the standard Bayesian update of a prior $P(\boldsymbol{\theta}|\mathcal{H})$ by the likelihood $P(\mathcal{D}|\boldsymbol{\theta},\mathcal{H})$, normalized by the evidence $P(\mathcal{D}|\mathcal{H})$. However, exact Bayesian inference over the parameters of large neural networks is typically intractable [65]. The network output depends nonlinearly on its parameters, which makes the posterior non-Gaussian and the marginalization integral analytically intractable.

Tractable Approximate Gaussian Inference (TAGI) [115] addresses this by maintaining Gaussian distributions over both the network parameters and hidden units, approximating their posterior as Gaussian, and exploiting two structural properties to reduce the computational cost:

(a) Diagonal covariance matrices are used for both the parameters $\boldsymbol{\theta}$ and the hidden units $\boldsymbol{Z}^{(j)}$ at each layer j, avoiding the storage and inversion of full covariance matrices.

(b) The hidden units of two non-adjacent layers are assumed to be conditionally independent given the intermediate layer, $\boldsymbol{Z}^{(j-1)} \perp \boldsymbol{Z}^{(j+1)}|\boldsymbol{z}^{(j)}$, so that parameters and hidden units are updated layer by layer rather than jointly, keeping the covariance matrices small.

By defining the short-hand notations $\left\{\boldsymbol{\theta}^{+}, \boldsymbol{Z}^{+}\right\} \equiv \left\{\boldsymbol{\theta}^{(j+1)}, \boldsymbol{Z}^{(j+1)}\right\}$, $\{\boldsymbol{\theta}, \boldsymbol{Z}\} \equiv \left\{\boldsymbol{\theta}^{(j)}, \boldsymbol{Z}^{(j)}\right\}$, and denoting $\boldsymbol{Y}$ as the observation vector, the hidden units at the output layer $\boldsymbol{Z}^{(0)}$ is updated following

$$\begin{aligned}
f\left(\boldsymbol{z}^{(0)}|\boldsymbol{y}\right) &= \mathcal{N}\left(\boldsymbol{z}^{(0)}; \boldsymbol{\mu}_{\boldsymbol{Z}^{(0)}|\boldsymbol{y}}, \boldsymbol{\Sigma}_{\boldsymbol{Z}^{(0)}|\boldsymbol{y}}\right), \\
\boldsymbol{\mu}_{\boldsymbol{Z}^{(0)}|\boldsymbol{y}} &= \boldsymbol{\mu}_{\boldsymbol{Z}^{(0)}} + \boldsymbol{\Sigma}^{\intercal}_{\boldsymbol{Y}\boldsymbol{Z}^{(0)}} \boldsymbol{\Sigma}^{-1}_{\boldsymbol{Y}} \left(\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{Y}}\right), \\
\boldsymbol{\Sigma}_{\boldsymbol{Z}^{(0)}|\boldsymbol{y}} &= \boldsymbol{\Sigma}_{\boldsymbol{Z}^{(0)}} - \boldsymbol{\Sigma}^{\intercal}_{\boldsymbol{Y}\boldsymbol{Z}^{(0)}} \boldsymbol{\Sigma}^{-1}_{\boldsymbol{Y}} \boldsymbol{\Sigma}_{\boldsymbol{Y}\boldsymbol{Z}^{(0)}},
\end{aligned}$$

the parameters at each layer are updated according to

$$\begin{aligned}
f\left(\boldsymbol{\theta}|\boldsymbol{y}\right) &= \mathcal{N}\left(\boldsymbol{\theta}; \boldsymbol{\mu}_{\boldsymbol{\theta}|\boldsymbol{y}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}|\boldsymbol{y}}\right), \\
\boldsymbol{\mu}_{\boldsymbol{\theta}|\boldsymbol{y}} &= \boldsymbol{\mu}_{\boldsymbol{\theta}} + \mathbf{J}_{\boldsymbol{\theta}} \left(\boldsymbol{\mu}_{\boldsymbol{Z}^{+}|\boldsymbol{y}} - \boldsymbol{\mu}_{\boldsymbol{Z}^{+}}\right), \\
\boldsymbol{\Sigma}_{\boldsymbol{\theta}|\boldsymbol{y}} &= \boldsymbol{\Sigma}_{\boldsymbol{\theta}} + \mathbf{J}_{\boldsymbol{\theta}} \left(\boldsymbol{\Sigma}_{\boldsymbol{Z}^{+}|\boldsymbol{y}} - \boldsymbol{\Sigma}_{\boldsymbol{Z}^{+}}\right) \mathbf{J}^{\intercal}_{\boldsymbol{\theta}}, \\
\mathbf{J}_{\boldsymbol{\theta}} &= \boldsymbol{\Sigma}_{\boldsymbol{\theta}\boldsymbol{Z}^{+}} \boldsymbol{\Sigma}^{-1}_{\boldsymbol{Z}^{+}},
\end{aligned}$$

and the hidden units at the layers other than the output one are computed by

$$
\begin{aligned}
f\left(\boldsymbol{z}|\boldsymbol{y}\right) &= \mathcal{N}\left(\boldsymbol{z}; \boldsymbol{\mu}_{Z|y}, \boldsymbol{\Sigma}_{Z|y}\right), \\
\boldsymbol{\mu}_{Z|y} &= \boldsymbol{\mu}_{Z} + \mathbf{J}_{Z}\left(\boldsymbol{\mu}_{Z^{+}|y} - \boldsymbol{\mu}_{Z^{+}}\right), \\
\boldsymbol{\Sigma}_{Z|y} &= \boldsymbol{\Sigma}_{Z} + \mathbf{J}_{Z}\left(\boldsymbol{\Sigma}_{Z^{+}|y} - \boldsymbol{\Sigma}_{Z^{+}}\right)\mathbf{J}_{Z}^{\intercal}, \\
\mathbf{J}_{Z} &= \boldsymbol{\Sigma}_{ZZ^{+}}\boldsymbol{\Sigma}_{Z^{+}}^{-1}.
\end{aligned}
$$

By following these operations, TAGI maintains the same computational complexity as methods trained with backpropagation [115]. It allows quantifying the epistemic uncertainty in the network parameters, modeling heteroscedastic aleatory uncertainty [116], and propagating these uncertainties to obtain the prior and posterior predictions for the outputs. TAGI has been employed in an LSTM network coupled with SSMs, where the baseline components remain and the LSTM network predicts the recurrent patterns and dependencies upon explanatory time series [92]. It is also used in a regression setting to obtain better prior knowledge of the deterioration rate with the information shared from similar bridges [117]. TAGI, whose effectiveness has been demonstrated in various machine learning tasks (e.g., adversarial attack generation, optimization, and reinforcement learning) [118], provides a suitable approach for training the intervention model in this study.

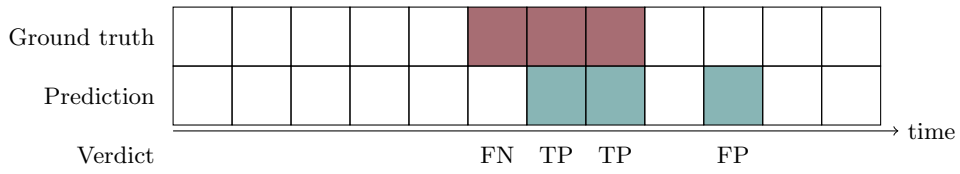
2.6 Evaluation Methodology

This section reviews the metrics used to evaluate time-series anomaly detection methods and assesses their suitability for long-term SHM. Section 2.6.1 reviews the standard metrics and discusses their limitations. Section 2.6.2 then presents the evaluation criteria that are most relevant in the SHM context.

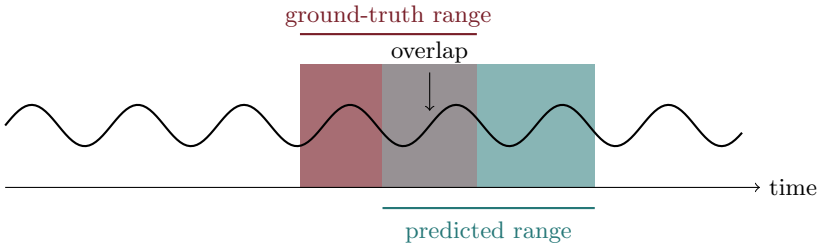
2.6.1 Standard Evaluation Metrics

Following the recent benchmark of Liu and Paparrizos [119], the metrics commonly used to evaluate time-series anomaly detection methods can be organized into two categories, as illustrated in Figure 2.15: *point-wise* measures, which treat each time step as an independent classification decision, and *range-wise* measures, which account for the fact that anomalies occur over a sequence of data rather than at isolated points.

Point-Wise Measures. Point-wise measures compare the predicted and ground-truth labels at each time step independently, counting true positives (`TP`), true negatives (`TN`), false



(a) Point-wise measures.



(b) Range-wise measures.

Figure 2.15 Two categories of time-series anomaly detection metrics.

positives (FP), and false negatives (FN) over individual points. A typical confusion matrix [72] adapted to the context of SHM is presented in Table 2.1. Given an observation, when the model prediction matches the ground truth, the corresponding prediction is evaluated as either a TP or TN depending on whether an anomaly is present or not. Conversely, if the model prediction is incorrect, this prediction is labeled as either a FN or FP.

Table 2.1 A confusion matrix in classification problems and its analogue in anomaly detection. True positives (TP) are the instances in which an alarm is triggered in the presence of abnormal data. False positives (FP) are the false alarms in anomaly detection. False negatives (FN) correspond to the missed alarms. True negatives (TN) indicate that no alarm is triggered given that there is no anomaly.

		ground truth	
		+ (abnormal)	− (normal)
model prediction	+ (alarm)	TP	FP
	− (no alarm)	FN	TN

We can rely on the *F1* score [120], the harmonic mean of precision and recall, to evaluate the overall accuracy while considering the TP, FP and FN, following

$$F1 = \frac{2\,\mathtt{TP}}{2\,\mathtt{TP} + \mathtt{FP} + \mathtt{FN}}, \qquad \text{Precision} = \frac{\mathtt{TP}}{\mathtt{TP} + \mathtt{FP}}, \qquad \text{Recall} = \frac{\mathtt{TP}}{\mathtt{TP} + \mathtt{FN}}, \tag{2.11}$$

where TP, FP, and FN are the numbers of true positive, false positive, and false negative

points. Since these counts require binary labels, the continuous anomaly score must first be thresholded, so the *F1* score depends on the chosen threshold. To remove this dependence, the *area under the receiver operating characteristic curve* (AUC-ROC) [72] and the *area under the precision-recall curve* (AUC-PR) [121] integrate performance over all thresholds. Of the two, AUC-ROC tends to be overly optimistic when the classes are strongly imbalanced, which is typical in time-series anomaly detection, whereas AUC-PR is more informative in this case. Across all point-wise measures, the common limitation is that they quantify *which* points are flagged but never *when* relative to the starting time of anomalies.

Range-Wise Measures. Range-wise measures were introduced to reflect the sequential nature of time series anomalies. The *range-based F1* (R-based-*F1*) [122] generalizes precision and recall to ranges, scoring a detection through an existence reward, an overlap reward and a cardinality factor that penalizes fragmented predictions. The *affiliation-based F1* [123] evaluates anomaly detection using local distance-based precision and recall computed within affiliation zones around each ground-truth event. By rewarding temporal proximity rather than exact overlap, it is robust to small timing errors and provides an interpretable event-level assessment. Finally, the *Volume Under the Surface* measures, VUS-ROC and VUS-PR [124], extend traditional AUC metrics for range anomaly detection by introducing tolerance regions around anomaly intervals and evaluating performance across a range of tolerance widths. The resulting ROC and precision-recall surfaces are then integrated over the tolerance dimension to obtain a volume-under-the-surface score. While range-wise measures better reflect the sequential structure of anomalies, they were designed for benchmarks in which anomalies are clearly distinguishable and when the objective is to rank detectors by overall accuracy.

2.6.2 Evaluation Criteria for the SHM Context

Two properties that are central to long-term SHM remain outside the scope of all the measures reviewed above. First, none of them explicitly quantifies the *time to detection*: point-wise measures ignore timing entirely, while range-wise measures incorporate temporal position, proximity, or tolerance only as a weighting or robustness mechanism, without reporting the elapsed delay between the starting time of an anomaly and its detection. Second, they summarize performance over a fixed labeled set rather than characterizing the detection sensitivity to anomalies having different magnitudes as well as the robustness against false alarms. To be suitable for long-term SHM, an evaluation framework must therefore respond to the following criteria.

Tradeoff Between Detectability and False Alarms. A method that triggers an excessive number of false alarms will detect small baseline changes but will also be uninformative, whereas a conservative method avoids false alarms at the expense of missing anomalies. A fair evaluation must therefore report both sides of this tradeoff [16]. Most standard metrics, such as the *F1* score, account for this factor by aggregating true positives, false positives, and false negatives into a single value. A single score, however, does not reveal where a method lies along the tradeoff: a low *F1* score can be attributed either to excessive false alarms or to missed alarms, two behaviors with very different consequences in SHM. The detectability and false-alarm rates should therefore be reported separately rather than collapsed into one number.

Time to Detection. For slow-developing deterioration, detecting an anomaly is necessary but not sufficient. A detection that arrives years or even decades after the anomaly started to develop provides little value for decision-making [125]. The *time to detection*, defined as the delay between the starting time of an anomaly and the first time step at which a detection is triggered, is an essential evaluation criterion for anomaly detection methods in SHM. A short time to detection leaves room for inspection and preventive maintenance before the deterioration progresses to a critical state, whereas a long delay narrows the available options and increases the risk and cost of reparation. A meaningful evaluation of anomaly detection methods in SHM must therefore measure the time to detection directly.

Probabilistic Evaluation on Synthetic Anomalies. The criteria above should not be assessed on a single anomaly of fixed magnitude starting at one particular time step. When evaluated this way, a method may appear to perform well or poorly by chance, for instance if the chosen starting time happens to coincide with an outlier or an abnormal reversible pattern that facilitates or hinders detection. To obtain a reliable assessment, an anomaly detection method must instead be evaluated over a population of anomalies having different magnitudes, with their starting times randomly positioned across the time series. Computing each evaluation metric over these randomized samples yields not only its average value, which reflects the expected performance regardless of where the anomaly occurs, but also its variability across starting times. This variability enables quantifying the reliability of the method and the confidence that can be placed in its detections, which is essential for risk-informed decisions on infrastructure.

2.7 Conclusion

This literature review has presented the methods and challenges that motivate the contributions of this thesis. The degradation of civil infrastructure manifests itself as small-magnitude irreversible changes in time series that develop over months to years and are hidden by reversible environmental effects having a much larger magnitude. This combination of subtlety, non-stationarity, and slow development makes long-term SHM a demanding setting for anomaly detection, and the comparative review of the preceding section showed that no existing paradigm meets all of its requirements at once: distance-based, prediction-based, and deep-learning methods each address part of the problem while leaving the others open.

Against this backdrop, SSMs emerge as the framework most naturally suited to long-term SHM. They combine online recursive inference, hidden-state uncertainty quantification, and an interpretable component decomposition that separates reversible environmental effects from the irreversible structural changes of interest. Within this framework, however, two limitations remain problematic. First, the autoregressive residual component tends to absorb the anomaly-induced drift, hindering the detection of the small baseline changes that characterize early-stage deterioration. Second, once an anomaly is detected, the model must adapt to the new regime, yet existing intervention methods assume that the intervention time and type are known a priori, which is rarely the case in practice.

Beyond the methods themselves, the way they are evaluated is also inadequate for long-term SHM. The point- and range-wise metrics were designed to rank detectors on benchmarks of clearly distinguishable anomalies, and none of them captures the criteria that matter most in the long-term SHM context: the time to detection of slow-developing changes, the reliability of a method across anomaly magnitudes and starting times, and the tradeoff between detectability and false alarms. Assessing these criteria calls instead for a probabilistic evaluation framework based on multiple anomalies having different magnitudes and starting times.

These gaps define the three research objectives of this thesis: to detect anomalies in time series by leveraging the residual's drift rather than being hindered by it; to predict the necessary SSM interventions from the data history without prior knowledge of their timing or type; and to evaluate these methods within a probabilistic framework that reflects the detectability, timeliness, and reliability. The methods developed to address these objectives are presented and evaluated in Chapters 3, 4, and 5.

CHAPTER 3 Bounded Autoregressive Component in State-Space Models

3.1 Introduction

In the switching Kalman filter (SKF), the autoregressive (`AR`) component modeling the stationary prediction errors has a tendency to wrongfully capture irreversible abnormal patterns that should be attributed to the baseline, and thus hinders their detectability. This chapter presents a new *bounded autoregressive* (`BAR`) component, which imposes constraints on the estimation of an `AR` hidden state while remaining compatible with the SKF framework (SKF-BAR). For that purpose, the existing Rectified Linear Unit (ReLU) is extended to a two-sided bounded form and use it to constrain the distribution of the `BAR` hidden state, yielding a *mixture Rectified Linear Unit* (mReLU) formulation. A modified *F1* score is also proposed for the probabilistic evaluation of the new `BAR` component.
This chapter is structured as follows. Section 3.2 extends the Rectified Linear Unit (ReLU) into a mixture form (mReLU) that constrains a distribution within two bounds, and builds on it to develop the `BAR` component and its integration within the state-space models (SSMs) framework. Section 3.3 proposes a time-series-wise evaluation metric, the *F1t* score, that integrates the true positive, false positive, and false negative counts together with the time to detection. Section 3.4 reports three sets of experiments, on synthetic data, on time series measured on a bridge, and on time series recorded on dams. The main contributions of this chapter include:

- a `BAR` component that enforces the stationarity of the residual component in SSMs and enhances the SKF in anomaly detection;
- a new evaluation metric that considers the true positive, false negative, and false positive counts, as well as the time to detection;
- a verification of the SKF-BAR detection model through a probabilistic evaluation over a population of synthetic time series, and a validation of its performance on real time series.

3.2 Bounded Autoregressive Component

This section introduces the bounded autoregressive (`BAR`) component for modeling constrained error processes in the SSMs framework. The core idea behind the `BAR` component is that a stationary autoregressive process should remain, by definition, within the bounds

defined by its stationary distribution. When an irreversible anomaly occurs, the standard AR component tends to drift beyond these bounds by absorbing the non-stationary patterns that should instead be attributed to the baseline components. The BAR component prevents this behavior by constraining the AR hidden state within an interval $[-b, b]$ derived from its stationary distribution, so that persistent changes are redirected to the baseline hidden states, where they can be detected as regime switches. Figure 3.1 illustrates this contrast on a synthetic time series subjected to a trend anomaly: after the anomaly starting at the midpoint, the estimation of the AR hidden state drifts beyond the bounds of its stationary distribution and away from the true values, whereas the BAR hidden state remains within the constraint region $\pm b$.

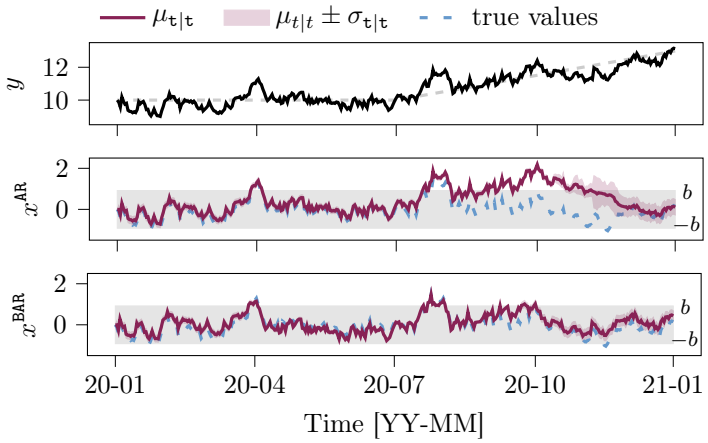


Figure 3.1 Illustration of the core idea behind the BAR component using a synthetic time series y subjected to a trend anomaly starting at the midpoint of the time series.

The moments of the BAR hidden states as well as its covariance with respect to other hidden states are approximated using the mReLU function developed in Section 3.2.2 based on the existing ReLU activation function and the theory of Gaussian mixture reduction in Section 3.2.1. The formulations for the BAR component are presented in Section 3.2.3.

3.2.1 Lower- and Upper-Bounded ReLU and Gaussian Mixture Reduction

The rectified Linear Unit (ReLU) is one of the most popular activation functions in deep learning models [126, 127]. A ReLU function bounds its input values to the positive range by assigning zero to all negative input values, i.e., $\phi_r(x) = \max(0, x)$. With the superposition of a minimum and a maximum function, one can obtain an activation function that bounds input values to a closed interval. Similar to a RuLU-6 [128] that bounds the input to [0, 6], a general formulation of a lower- and upper-bounded ReLU that symmetrically constrains

input values between $-b$ and b is given by

$$\phi_r^{\pm}(x) = \min\left(\max\left(-b, x\right), b\right).$$

A Gaussian mixture combines multiple Gaussian distributions, each represented by $\{\pi_i, \mu_i, \sigma_i^2\}$, where π_i is the probability of each distribution, and μ_i and σ_i^2 denote the mean and variance. This approach enables calculating the exact moments for the combined distribution following

$$\tilde{\mu} = \sum_{i=1}^{n} \pi_i \mu_i, \tag{3.1}$$

$$\tilde{\sigma}^2 = \sum_{i=1}^{n} \pi_i \left(\sigma_i^2 + (\mu_i - \tilde{\mu})^2\right). \tag{3.2}$$

The Gaussian mixture reduction [129] approximates this combined distribution by a Gaussian distribution using the moments defined by Eqs. 3.1 and 3.2. Figure 3.2 presents examples of Gaussian mixture reduction for two and three components.

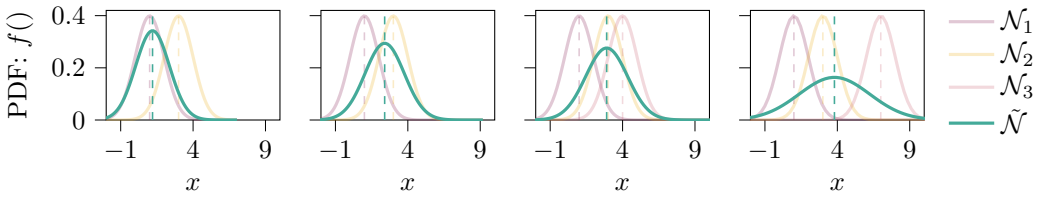


Figure 3.2 Four Gaussian mixture reduction examples. The first and the second plots exhibit the mixture $\tilde{\mathcal{N}}$ for two Gaussian components, which are $\mathcal{N}_1 = \{\pi_1 = 0.9, \mu_1 = 1, \sigma_1^2 = 1^2\}$, $\mathcal{N}_2 = \{\pi_2 = 0.1, \mu_2 = 3, \sigma_2^2 = 1^2\}$ for the first plot, and $\mathcal{N}_1 = \{\pi_1 = 0.3, \mu_1 = 1, \sigma_1^2 = 1^2\}$, $\mathcal{N}_2 = \{\pi_2 = 0.7, \mu_2 = 3, \sigma_2^2 = 1^2\}$ for the second plot. The third and the fourth plots show the mixtures $\tilde{\mathcal{N}}$ for three Gaussian components, which are $\mathcal{N}_1 = \{\pi_1 = 0.2, \mu_1 = 1, \sigma_1^2 = 1^2\}$, $\mathcal{N}_2 = \{\pi_2 = 0.5, \mu_2 = 3, \sigma_2^2 = 1^2\}$, and $\mathcal{N}_3 = \{\pi_3 = 0.3, \mu_3 = 4, \sigma_3^2 = 1^2\}$ for the third plot, and $\mathcal{N}_1 = \{\pi_1 = 0.2, \mu_1 = 1, \sigma_1^2 = 1^2\}$, $\mathcal{N}_2 = \{\pi_2 = 0.5, \mu_2 = 3, \sigma_2^2 = 1^2\}$, and $\mathcal{N}_3 = \{\pi_3 = 0.3, \mu_3 = 7, \sigma_3^2 = 1^2\}$ for the fourth plot.

3.2.2 Mixture Rectified Linear Unit (mReLU)

The `BAR` component relies on a Gaussian mixture reduction between a truncated Gaussian and two real values obtained from a lower- and upper-bounded ReLU function. Figure 3.3 illustrates an example of the mReLU transformation, where the input is a Gaussian distribution $f(x^{\texttt{AR}})$ and the output is approximated by another Gaussian distribution $f(x^{\texttt{BAR}})$ given $X^{\texttt{BAR}} = \phi_r^{\pm}(X^{\texttt{AR}})$. Starting from the plot on the second row, $f(x^{\texttt{AR}})$ is truncated into three

segments by two boundary values $-b$ and b. By passing the truncated segments through a ReLU function, the intervals $(-\infty, -b]$ and $[b, +\infty)$ are collapsed to $-b$ and b, as shown in the second plot on the first row. These can be regarded as two Gaussian components with variances equal to zero, $\{\pi_1 = F_{X^{\texttt{AR}}}(-b), \mu_1 = -b, \sigma_1^2 = 0\}$ and $\{\pi_2 = 1 - F_{X^{\texttt{AR}}}(b), \mu_2 = b, \sigma_2^2 = 0\}$, following the same notation for the Gaussian mixture reduction mentioned in Section 3.2.1. The probability density function (PDF) for the middle interval $f(\tilde{x}^{\texttt{AR}})$ remains unchanged by the ReLU and is approximated by matching its moments to a Gaussian PDF illustrated by a solid line in the third plot on the first row, which acts as the third component $\{\pi_3 = F_{X^{\texttt{AR}}}(b) - F_{X^{\texttt{AR}}}(-b), \mu_3, \sigma_3^2\}$. The Gaussian mixture reduction represents the distribution of the BAR hidden state $f(x^{\texttt{BAR}})$ as shown in the rightmost plot.

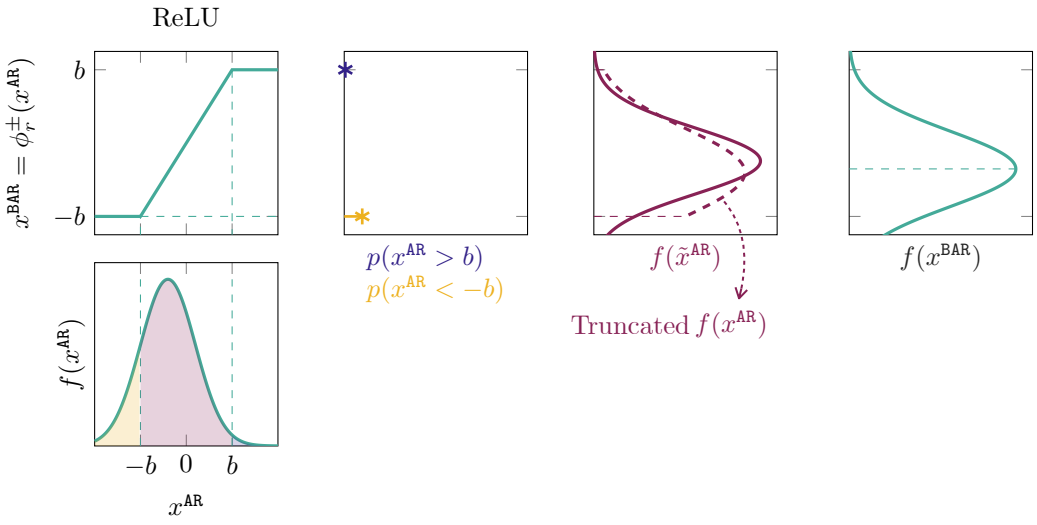


Figure 3.3 A mReLU transformation example in the context of BAR. The input distribution $f(x^{\texttt{AR}})$ has a mean value that is close to $-b$ within the boundary and a variance that extends beyond the $\pm b$.

The calculations of the moments μ_3, σ_3, $\mu^{\texttt{BAR}}$, and $\sigma^{\texttt{BAR}}$, as well as the covariance between $X^{\texttt{BAR}}$ and other hidden states are carried in the mReLU function which can be summarized as

$$\left\{\mu^{\texttt{BAR}}, \sigma^{\texttt{BAR}}, \text{cov}\left(X^{\texttt{BAR}}, X^{\cdot}\right)\right\} = \text{mReLU}\left(\mu^{\texttt{AR}}, \sigma^{\texttt{AR}}, \text{cov}(X^{\texttt{AR}}, X^{\cdot}), \gamma\right), \tag{3.3}$$

where $\text{cov}\left(X^{\texttt{BAR}}, X^{\cdot}\right)$ is the covariance between BAR and other hidden states, $\text{cov}(X^{\texttt{AR}}, X^{\cdot})$ is the covariance between AR and other hidden states, and γ is a hyperparameter that defines the boundaries according to the autoregressive process's stationary standard deviation $\sigma^{\texttt{AR},0}$

as defined by Eq. 2.10, so that the boundary value is

$$b = \gamma\sigma^{\texttt{AR},0} = \gamma\sqrt{\frac{\sigma_W^2}{1-(\phi^{\texttt{AR}})^2}}. \tag{3.4}$$

Figure 3.4 illustrates a set of realizations from a first-order autoregressive process along with the boundaries $\pm b$ obtained for $\gamma \in \{1, 2, 3\}$. As γ increases, the realizations are increasingly unlikely to exceed the boundaries. When the constant $\gamma = 2$, the bounded area corresponds to a 95 % coverage region for the autoregressive process's stationary distribution, and it is used as the default value in the `BAR` component. The impact of other γ values on the anomaly detection performance is further examined in Section 3.4.3.

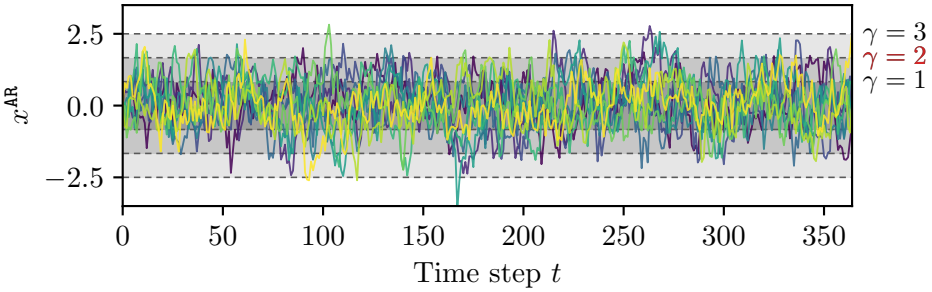


Figure 3.4 Ten realizations of an autoregressive process with $\phi^{\texttt{AR}} = 0.8$ and $\sigma_W = 0.5$, along with the stationary bounds $\pm b = \pm\gamma\sigma^{\texttt{AR},0}$ for $\gamma \in \{1, 2, 3\}$ following Eq. 3.4. The default value $\gamma = 2$ bounds a 95 % coverage region of the process's stationary distribution.

The truncated Gaussian between $-b$ and b is approximated by a Gaussian PDF whose moments match those of a truncated Gaussian, for which the expected value is

$$\begin{aligned}
\alpha_{\texttt{L}} &= -\tfrac{b+\mu^{\texttt{AR}}}{\sigma^{\texttt{AR}}}, \\
\alpha_{\texttt{U}} &= \tfrac{b-\mu^{\texttt{AR}}}{\sigma^{\texttt{AR}}}, \\
\omega &= \Phi(\alpha_{\texttt{U}}) - \Phi(\alpha_{\texttt{L}}) = \Pr(-b < X^{\texttt{AR}} < b), \\
\beta &= \tfrac{\phi(\alpha_{\texttt{U}})-\phi(\alpha_{\texttt{L}})}{\omega}, \\
\tilde{\mu}^{\texttt{AR}} &= \mu^{\texttt{AR}} - \beta\sigma^{\texttt{AR}},
\end{aligned}$$

the variance is

$$\begin{aligned}
\kappa &= 1 - \tfrac{\alpha_{\texttt{U}}\phi(\alpha_{\texttt{U}})-\alpha_{\texttt{L}}\phi(\alpha_{\texttt{L}})}{\omega} - \beta^2, \\
(\tilde{\sigma}^{\texttt{AR}})^2 &= \kappa(\sigma^{\texttt{AR}})^2,
\end{aligned}$$

and the linearized covariance between $\tilde{X}^{\texttt{AR}}$ and another variable $X^{\cdot}$ is

$$\text{cov}(\tilde{X}^{\texttt{AR}}, X^{\cdot}) = \kappa^{1/2}\text{cov}(X^{\texttt{AR}}, X^{\cdot}).$$

A Gaussian mixture reduction is performed on the three Gaussian components, $\{\pi_1 = F_{X^{\texttt{AR}}}(-b), \mu_1 = -b, \sigma_1^2 = 0\}$, $\{\pi_2 = 1 - F_{X^{\texttt{AR}}}(b), \mu_2 = b, \sigma_2^2 = 0\}$ and $\{\pi_3 = F_{X^{\texttt{AR}}}(b) - F_{X^{\texttt{AR}}}(-b), \mu_3 = \tilde{\mu}^{\texttt{AR}}, \sigma_3^2 = (\tilde{\sigma}^{\texttt{AR}})^2\}$, to obtain the output of `BAR` hidden state

$$X^{\texttt{BAR}} = -b\Phi(\alpha_{\texttt{L}}) + \omega\tilde{X}^{\texttt{AR}} + b(1 - \Phi(\alpha_{\texttt{U}})).$$

Following Eq. 3.1, the expected value of $X^{\texttt{BAR}}$ is a mixture of the expected values of the three Gaussian components according to their probabilities so that

$$\mu^{\texttt{BAR}} = -b\Phi(\alpha_{\texttt{L}}) + \omega\tilde{\mu}^{\texttt{AR}} + b(1 - \Phi(\alpha_{\texttt{U}})).$$

By setting $\sigma_1 = 0$ and $\sigma_2 = 0$ in Eq. 3.2, the variance of $X^{\texttt{BAR}}$ is given by

$$(\sigma^{\texttt{BAR}})^2 = \omega(\tilde{\sigma}^{\texttt{AR}})^2 + \omega(\tilde{\mu}^{\texttt{AR}} - \mu^{\texttt{BAR}})^2 + \Phi(\alpha_{\texttt{L}})(b + \mu^{\texttt{BAR}})^2 + (1 - \Phi(\alpha_{\texttt{U}}))(b - \mu^{\texttt{BAR}})^2,$$

and the covariance between $X^{\texttt{BAR}}$ and other hidden states $X^{\cdot}$ is

$$\lambda = (\omega\kappa)^{1/2},$$
$$\text{cov}(X^{\texttt{BAR}}, X^{\cdot}) = \lambda\text{cov}(\tilde{X}^{\texttt{AR}}, X^{\cdot}).$$

3.2.3 BAR Formulation

This section outlines the integration of the `BAR` component within the SSMs framework. The model matrices for the `BAR` component are

$$\boldsymbol{x_t} = \begin{bmatrix} x^{\texttt{AR}} \\ x^{\texttt{BAR}} \end{bmatrix}, \mathbf{A} = \begin{bmatrix} \phi^{\texttt{AR}} & 0 \\ 0 & 0 \end{bmatrix}, \mathbf{Q} = \begin{bmatrix} (\sigma_W^{\texttt{AR}})^2 & 0 \\ 0 & 0 \end{bmatrix}, \mathbf{F} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}^{\intercal}, \mathbf{R} = \sigma_V^2.$$

The `BAR` approach only changes the transition model in the prediction step presented in Section 2.3.1 by bounding its autoregressive output using mReLU and by incorporating the bounded values to the predicted hidden states. Following Eq. 2.1, the moments for the output

of the transition model can be calculated as

$$\tilde{\boldsymbol{\mu}}^{\boldsymbol{X}}_{t|t-1} = \mathbf{A}\boldsymbol{\mu}^{\boldsymbol{X}}_{t-1} = \begin{bmatrix} \phi^{\texttt{AR}} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mu^{\texttt{AR}}_{t-1} \\ \mu^{\texttt{BAR}}_{t-1} \end{bmatrix} = \begin{bmatrix} \phi^{\texttt{AR}}\mu^{\texttt{AR}}_{t-1} \\ 0 \end{bmatrix}, \tag{3.5}$$

$$\begin{aligned} \tilde{\boldsymbol{\Sigma}}^{\boldsymbol{X}}_{t|t-1} &= \mathbf{A}\boldsymbol{\Sigma}^{\boldsymbol{X}}_{t-1}\mathbf{A}^{\intercal} + \mathbf{Q}, \\ &= \begin{bmatrix} \phi^{\texttt{AR}} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} (\sigma^{\texttt{AR}}_{t-1})^2 & \operatorname{cov}(X^{\texttt{AR}}_{t-1}, X^{\texttt{BAR}}_{t-1}) \\ \operatorname{cov}(X^{\texttt{AR}}_{t-1}, X^{\texttt{BAR}}_{t-1}) & (\sigma^{\texttt{BAR}}_{t-1})^2 \end{bmatrix} \begin{bmatrix} \phi^{\texttt{AR}} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} (\sigma^{\texttt{AR}}_{W})^2 & 0 \\ 0 & 0 \end{bmatrix}, \\ &= \begin{bmatrix} (\phi^{\texttt{AR}})^2(\sigma^{\texttt{AR}}_{t-1})^2 + (\sigma^{\texttt{AR}}_{W})^2 & 0 \\ 0 & 0 \end{bmatrix}. \end{aligned} \tag{3.6}$$

Note that $X^{\texttt{BAR}}$ is not involved in the transition model as indicated by the components of the $\mathbf{A}$ matrix that are all 0 except in the `AR` position $[\mathbf{A}]_{1,1}$. On the other hand, $X^{\texttt{BAR}}$ is the only observable hidden state. This is because the `BAR` method constrains a variable that is analogous to the autoregressive component without modifying the original variable, whose entire distribution remains and is transferred through time steps. Without this procedure, the approximation errors caused by the truncation would accumulate at each time step, leading to a monotonically increasing bias and monotonically decreasing variance [130].

The moments of $X^{\texttt{BAR}}_t$ as well as its covariance with respect to $X^{\texttt{AR}}_t$ are computed by bounding the `AR` hidden state using the mReLU function defined in Eq. 3.3. They are then added to Eq. 3.5 and Eq. 3.6 to obtain the predicted hidden states with constrained residuals, whose moments conditional on all the observations up to $t-1$ are defined by

$$\begin{aligned} \boldsymbol{\mu}^{\boldsymbol{X}}_{t|t-1} &= \tilde{\boldsymbol{\mu}}^{\boldsymbol{X}}_{t|t-1} + \begin{bmatrix} 0 \\ \mu^{\texttt{BAR}}_{t|t-1} \end{bmatrix} = \begin{bmatrix} \phi^{\texttt{AR}}\mu^{\texttt{AR}}_{t-1} \\ \mu^{\texttt{BAR}}_{t|t-1} \end{bmatrix}, \\ \boldsymbol{\Sigma}^{\boldsymbol{X}}_{t|t-1} &= \tilde{\boldsymbol{\Sigma}}^{\boldsymbol{X}}_{t|t-1} + \begin{bmatrix} 0 & \operatorname{cov}(X^{\texttt{BAR}}_t, X^{\texttt{AR}}_t) \\ \operatorname{cov}(X^{\texttt{BAR}}_t, X^{\texttt{AR}}_t) & (\sigma^{\texttt{BAR}}_{t|t-1})^2 \end{bmatrix}, \\ &= \begin{bmatrix} (\phi^{\texttt{AR}})^2(\sigma^{\texttt{AR}}_{t-1})^2 + (\sigma^{\texttt{AR}}_{W})^2 & \operatorname{cov}(X^{\texttt{BAR}}_t, X^{\texttt{AR}}_t) \\ \operatorname{cov}(X^{\texttt{BAR}}_t, X^{\texttt{AR}}_t) & (\sigma^{\texttt{BAR}}_{t|t-1})^2 \end{bmatrix}, \end{aligned}$$

which are then passed to the observation model described by Eq. 2.2, and finally to the update step following Eq. 2.7 and Eq 2.8 as presented in Section 2.3.1. The process of integrating the `BAR` component in the SSMs framework is illustrated by the flowchart in Figure 3.5 with the modified steps highlighted by dashed lines.

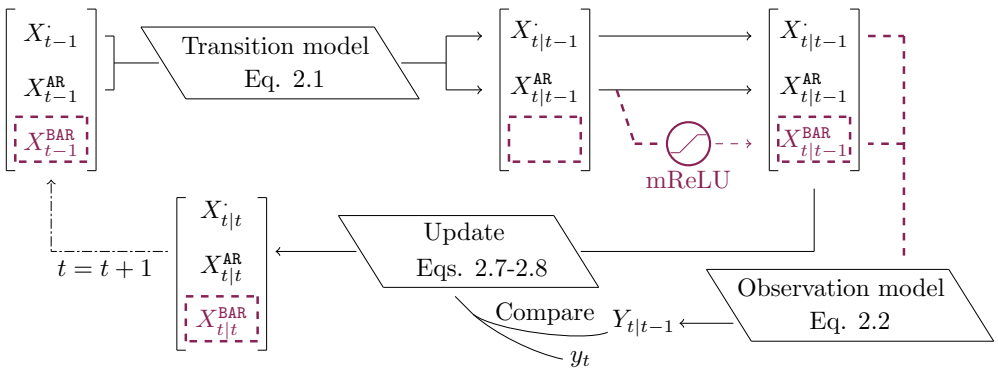


Figure 3.5 Integration process of the BAR component in the SSMs framework. The modifications made by the BAR method are highlighted by dashed lines.

3.3 Evaluation Metrics

Anomaly detection is to some extent a classification problem, as each observation or segment must ultimately be assigned to one of two classes, normal or abnormal. Metrics commonly used for evaluating classification models, such as the confusion matrix [72] and the *F1* score [120] reviewed in Section 2.6.1, can be adapted to the anomaly detection task. However, in the context of time series anomaly detection, it is not feasible to evaluate the anomaly detectability point-wise because it would be biased by the length of anomalies, where models gain more reward for detecting long anomalies rather than short ones [131]. Take the baseline-related anomalies in Figure 2.2 as an example, after an anomaly is introduced in a time series where hidden states are shifted, all the following observations are labeled as abnormal. A model could then go through more TP instances when applied to a longer abnormal segment and thus wrongfully achieve a higher *F1* score. As a result, by using the existing *F1* definition, the performance of the anomaly detection model is not only a function of its detectability, but it is also biased by the length of abnormal segment.

Instead of updating the confusion matrix for each observation, we propose to assess it on a *time-series* basis, whose concept is illustrated in Figure 3.6. Given a time series with an anomaly, e.g., an acceleration anomaly, if a model triggers an alarm before the anomaly starts, its prediction will be labeled as a FP whose count will be incremented by 1. On the other hand, if an alarm is correctly triggered within the abnormal region without any false alarm prior to it, the TP count is incremented by 1. Yet, if no alarm is triggered throughout an entire time series subjected to an anomaly, or if an alarm is triggered after a pre-defined detection window, the model performance will be assessed as a missed alarm with 1 added to the FN count. The evaluation on a time series is terminated as soon as any of the FP, TP or FN count is updated.



(a) False alarm (b) True alarm (c) Missed alarm

Figure 3.6 Time-series-wise TP, FP and FN definition. FP is incremented by 1 when an alarm is triggered before the occurrence of anomalies. TP is incremented by 1 when an alarm is correctly triggered in the abnormal segment. FN is increased by 1 if an alarm is missed within the detection window. When any of these three scenarios happens, analysis on this time series terminates and the confusion matrix is updated.

With such an evaluation process, the model's performance is independent of the length of abnormal time series because each one only has a maximum increment of 1 in either TP, FP, or FN. Nonetheless, that determination of TP counts ignores the time delay in detecting anomalies because alarms triggered at any time within the anomaly region have the same increment in TP. To account for the time to detection in the anomaly detectability, we propose to use a decay ratio λ that linearly penalizes the *F1* score with an increase in the *time to detection*, Δ_t. This new metric is called the *F1 temporal* (*F1t*) score, such that $F1t = \lambda F1$.

Figure 3.7 shows the relation between λ and Δ_t, which is the average time to detection for all the time series. λ is one if a model triggers an alarm before or at the moment when the anomaly starts, while it linearly decreases to zero as the required time to detection reaches the end of a pre-defined detection window l_{dw}. This approach is aware that a time to detection smaller than zero corresponds to a false alarm. Since false alarms are penalized through the FP in the calculation of *F1*, λ remains one for $\Delta_t < 0$ to avoid a double penalization. A model can achieve a maximum *F1t* score of 1 if it detects all the anomalies at their exact starting time without triggering any false alarms. On the other hand, any delay in detection, false alarms or missed alarms will lead to a reduction in the *F1t* score.

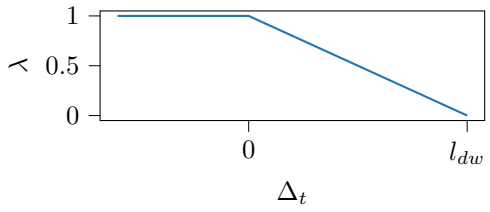


Figure 3.7 Decreasing λ with longer time to detection. l_{dw} is the length of the target detection window. λ drops to zero if the required time to detect an anomaly is longer than l_{dw}.

3.4 Experiments

By coupling the BAR component with a Switching Kalman Filter (SKF) reviewed in Section 2.3.1, the SKF-BAR is tested against the existing SKF-AR framework on three distinct case studies. As presented in Section 3.2, the choice of the hyperparameter γ decides the boundaries imposed on the residual estimation: the bigger it is, the less constraints it imposes, so that the BAR component gradually converges towards a standard AR component. In the first experiments, we demonstrate the performance of SKF-BAR with the default value $\gamma = 2$, corresponding to a 95% coverage of the stationary distribution, on 1) a synthetic time series with a fixed anomaly in Section 3.4.1, and 2) a real time series measured on a bridge with synthetic anomalies of three different magnitudes in Section 3.4.2. These two case studies are intended to show the average performance of SKF-BAR when deployed at scale without tuning γ to a specific value. A more complete evaluation is then conducted in Section 3.4.3 on eight independent time series recorded on dams, showing the sensitivity of the SKF-BAR method to a range of anomaly magnitudes, different anomaly types and starting times, as well as different values of γ. The BAR component is implemented in the open-source library Canari [https://github.com/Bayes-Works/canari].

3.4.1 Synthetic Data

A synthetic time series is generated using a SSM comprising a local level (LL) and an autoregressive (AR) component, covering the period from 2020-01-01 to 2021-01-01 with a daily time-step length. The model matrices for this experiment can be found in Appendix A.1.1. A trend anomaly is introduced starting at the midpoint (2020-07-01) of the time series as depicted in Figure 3.8.

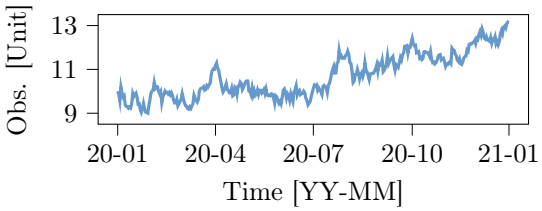


Figure 3.8 A one-year synthetic time series with daily time-step length generated with LL and AR components. A trend anomaly starts at the midpoint 2020-07-01.

A SKF involving a *stationary* and a *trend-stationary regime* is used to detect the simulated anomaly. The stationary regime is modelled using LL + AR/BAR components, while the trend-stationary regime is modelled by a local trend (LT) instead of a local level component. The

model matrices for these regimes are summarized in Appendix A.1.2 and A.1.3. The probability of detecting the synthetic anomaly, denoted as $p^{\texttt{anm}}$, is represented by the probability of the trend-stationary regime.

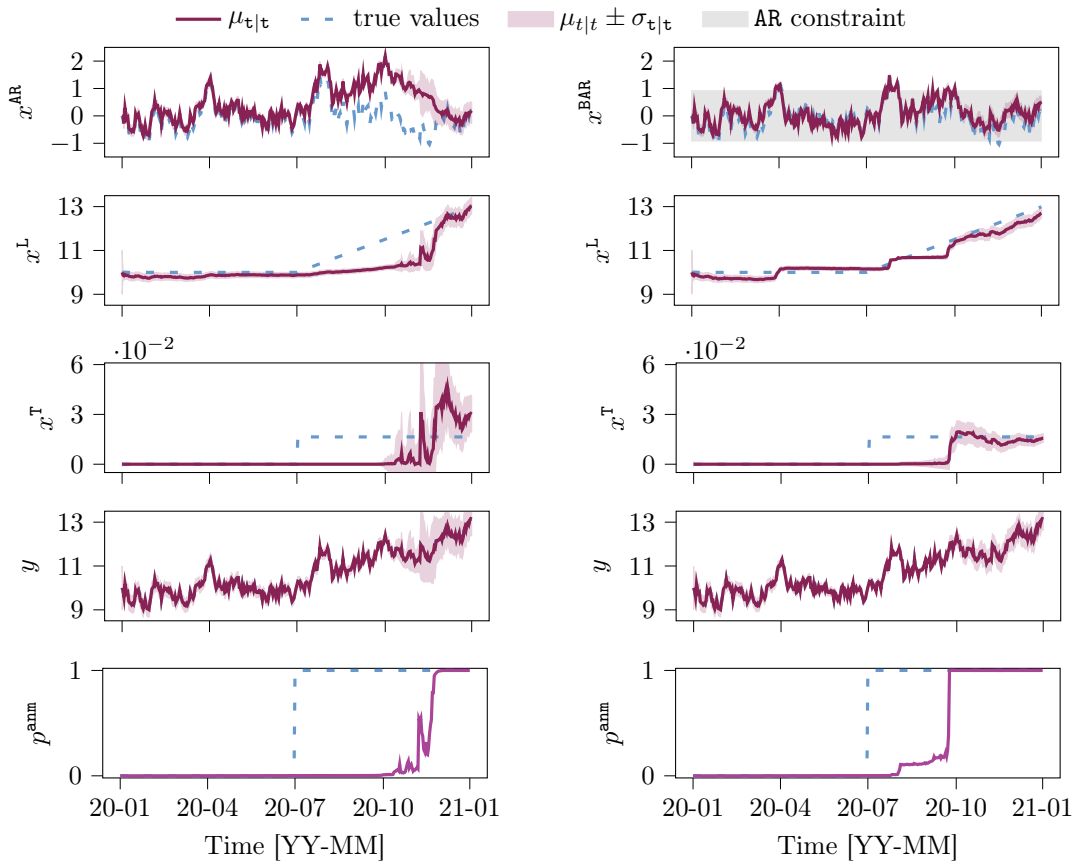
Figure 3.9 compares the estimations for the SKF-AR and the SKF-BAR with the true values. In the first plot from Figure 3.9a, the estimations of the `AR` hidden state gradually drift away from the true values after the introduction of the anomaly, implying that the changes in the trend caused by the anomaly are wrongfully captured by the residual component, which was initially only intended to model a stationary error-process. This drift results in delays in the anomaly detection quantified by $p^{\texttt{anm}}$, as well as inaccuracies in the hidden state estimation for the `LL` and `LT` components.

On the other hand, the `BAR` approach in Figure 3.9b constrains the drift and enable the SKF to quickly catch up with the changes in the `L` and `T` hidden states. Additionally, after the anomaly is introduced, its predicted responses y exhibit a smaller variance compared to those from the SKF-AR model. As a result, the SKF-BAR model detects the anomaly with $p^{\texttt{anm}} \geq 0.5$ almost two months earlier than the SKF-AR. This delay in the time to detection can be further reduced for the SKF-BAR model by imposing a smaller γ parameter to further constrain the `AR` hidden state. The tuning of the γ parameter will be discussed in Section 3.4.3.

3.4.2 Bridge Data

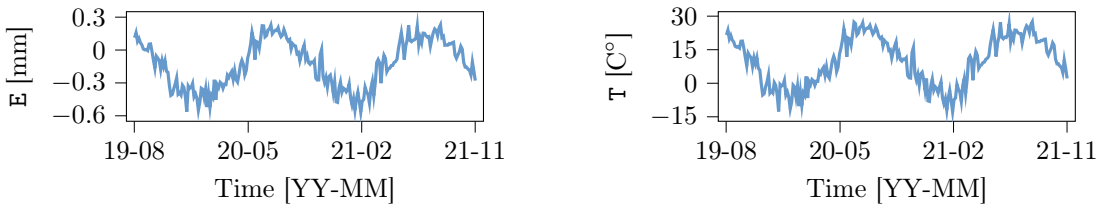
The time series studied in this section comprise measurements of elongation (`E`) and air temperature (`T`) obtained from a bridge. Both raw time series have a time-step length of 10 minutes, which is aggregated during preprocessing to 3.5 days as demonstrated in Figure 3.10. Elongation measurements exhibit a dependency on the air temperature, which can be accounted for in the observation model described in Section 2.3.1. The two time series, covering the period between 2019-08-16 and 2021-11-16, are considered to be stationary, given that no structural interventions or anomalies were recorded during this period, and the temperature remained stable. Three trend anomalies with magnitudes of -0.5, -0.1 and -0.025 mm/year are introduced on the elongation measurements starting at the timestamp 2021-01-22 as depicted by the true values for the $x^{\texttt{T}}$ hidden state in Figure 3.11.

The SSM components for the temperature measurement are $\{\texttt{LL}^{\texttt{T}} + \texttt{PD}^{\texttt{T}} + \texttt{AR}^{\texttt{T}}\}$ for both the stationary and trend-stationary regimes because there is no regime switch in the temperature. A local level component ($\texttt{LL}^{\texttt{T}}$) is used to describe the constant level of the temperature, a periodic component ($\texttt{PD}^{\texttt{T}}$) is used for modeling its yearly periodic pattern, and an autoregressive component (`AR`) is used to describe the model's errors. For the elongation measurement,



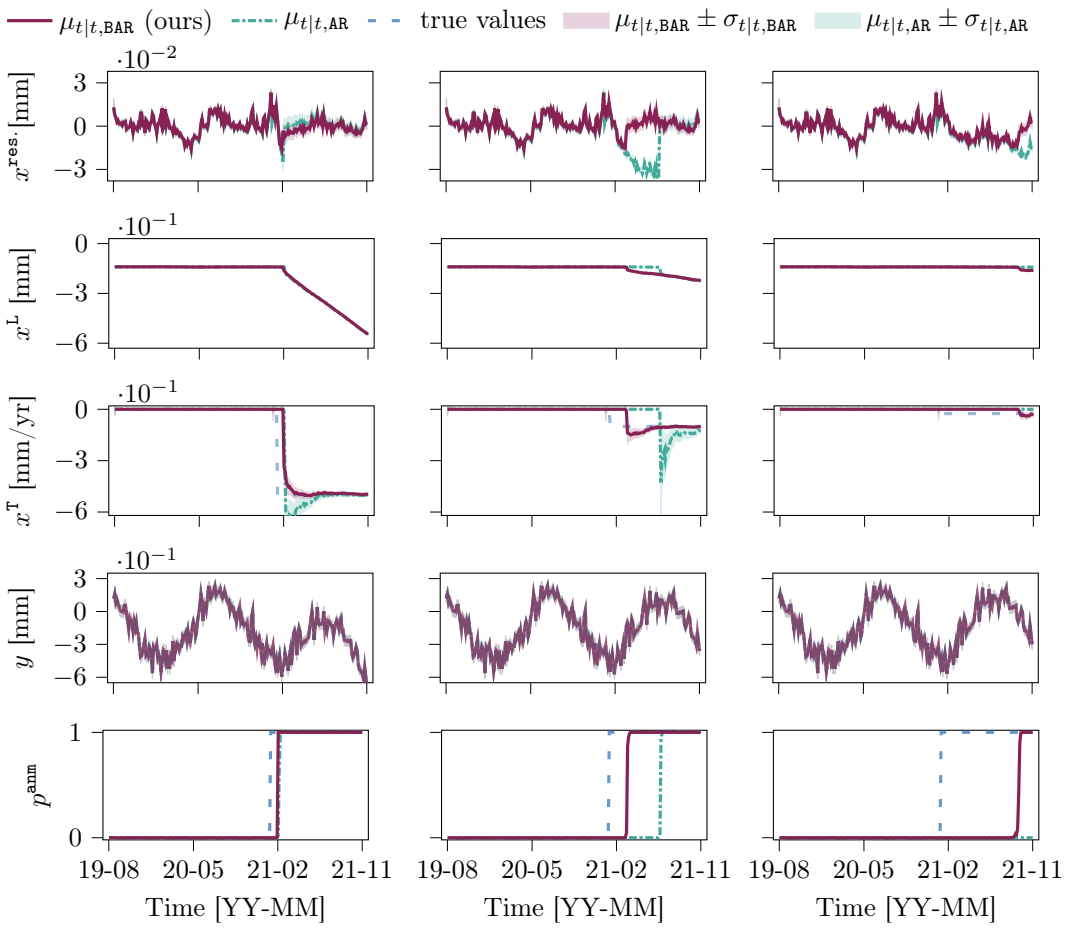
(a) SKF-AR (b) SKF-BAR (ours)

Figure 3.9 Hidden states estimation on the synthetic data with the standard `AR` and `BAR` in comparing with the true values.



(a) Elongation (b) Temperature

Figure 3.10 Elongation and temperature measurement on a bridge, where time steps are aggregated from 10 minutes to 3.5 days.



(a) Large anomaly (b) Medium anomaly (c) Small anomaly

Figure 3.11 Comparison of hidden states between SKF-AR and SKF-BAR under different anomaly magnitudes in a time series measured on a bridge. Large anomaly indicates a trend anomaly of -0.5 mm/year; the medium is -0.1 mm/year; and the small is -0.025 mm/year. The residual components (`res.`) include the `AR` and the `BAR` components.

$\{\texttt{LL}^{\texttt{E}} + \texttt{AR}/\texttt{BAR}^{\texttt{E}}\}$ and $\{\texttt{LT}^{\texttt{E}} + \texttt{AR}/\texttt{BAR}^{\texttt{E}}\}$ components are employed for the stationary and trend-stationary regimes without using a periodic component, which is accounted for through the dependency upon the temperature measurements. The constant level of the elongation for the stationary regime is described by a local level component ($\texttt{LL}^{\texttt{E}}$), while its constant speed for the trend-stationary regime is modelled by a local trend component ($\texttt{LT}^{\texttt{E}}$) with the model errors captured by either an `AR` or a `BAR` component. All model matrices are listed in Appendix A.2.

Figure 3.11 compares the results for the SKF-BAR and the SKF-AR models on the elongation time series with synthetic anomalies. When the anomaly is large enough to be seen visually as shown in Figure 3.11a, both models yield similar estimations with a slight delay in the anomaly detection. Yet, the SKF-BAR outperforms the SKF-AR in capturing the shift in trend $x^{\texttt{T}}$ without overestimation at the beginning when there is a regime switch. As the anomaly magnitude decreases, it becomes almost impossible to identify it visually as shown in Figure 3.11b-c, and it takes a longer time for both models to detect it. SKF-AR detects the second anomaly of -0.1 mm/year after a delay of more than five months (on 2021-07-09), while SKF-BAR detects it after only two (on 2021-03-23). For this anomaly, the SKF-AR model displays a constant trend in the estimations of the residual component (`res`.) and fails to quickly capture the changes in $x^{\texttt{L}}$ and $x^{\texttt{T}}$ after the anomaly develops, which is consistent with the findings from Section 3.4.1. The smallest anomaly magnitude that SKF-BAR can detect is -0.025 mm/year, as shown in Figure 3.11c, where the SKF-BAR model correctly estimates the $x^{\texttt{T}}$ hidden state after detecting the anomaly while maintaining a stationary residual process. In contrast, for the third case, the SKF-AR is unable to detect or identify the changes in the hidden states.

3.4.3 Dam Data

This section presents a probabilistic study of the `BAR`'s performance under different γ values (`BAR`-γ) on eight time series measured on a dam. When γ is too small, false alarms may occur due to the strict constraint; when it is too large, `BAR` is similar to `AR` and is capable of capturing trends. To estimate `BAR`'s performance, not only we must consider the anomaly detectability but also the false alarm rate. However, the ground truth of whether anomalies are present is typically unknown for real time series, making it impossible to categorize an alarm as true or false. To overcome this issue, we generate synthetic time series using a predefined SSM, on which synthetic acceleration, trend and level anomalies are overlaid. The performance of `BAR`-γ is then verified on these synthetic time series with simulated anomalies and finally validated on the real time series.

Overview of Datasets and Synthetic Time Series. The datasets used in this study consist of eight time series, each representing either automatic or manual measurements of crack openings or displacements on a dam. The first time series A01C spanning from 1998-01 to 2005-12 with a daily time interval measures the crack openings on a dam. It demonstrates an increasing trend and yearly periodicity. The second time series, A16D, measures the displacements, and spans from 2011-11 to 2022-12 with a half-day time interval. It shows a decreasing trend and yearly periodicity. The M08C time series measuring the crack openings has a time interval of 91 days. It shows a decreasing trend between 2013-12 and 2022-12. The crack openings, measured in M09C time series with a 49-day time interval, increase between 1996-04 and 2007-03. The displacements, measured in M11D time series with a 28-day time interval, increase between 2007-02 and 2022-12. The M12D displacement time series has a 63-day time interval and a decreasing trend from 2011-11 to 2022-12. The M13D displacement time series decreases from 2009-01 to 2020-07, having a 35-day time interval. The M14D displacement time series is measured every 41 days, and increases between 1991-06 and 1999-08. Yearly periodicity and missing data are observed throughout all the time series as depicted in Appendix B. Table 3.1 provides a summary of the measurement type, start and end times, acquisition period, number of observations, and SSM components employed for each time series. Among them, results for M08C are showcased in the following sections, while those for the other time series are documented in Appendices C, D, and E.

Table 3.1 Summary of eight time series measured on a dam with respect to time-series name, measurement type, start & end time, acquisition period, number of observations and SSM components. Crack openings and displacement are notated as C.O. and Disp. Stationary/ non-stationary regimes are referred to as S./ N.S. The abbreviation of res. includes `AR` or `BAR` residual component. The *highlighted row* is the time series example showcased in the following sections.

Name	Type	Start time	End time	Period [day]	#obs.	Components (S./ N.S.)
A01C	C.O.	1998-01-29	2005-12-29	1	2789	`LT`/`LA`+ `KR` (10)+ res.
A16D	Disp.	2011-11-17	2022-12-07	0.5	8154	`LT`/`LA`+ `KR` (10)+ res.
M08C	*C.O.*	*2013-12-09*	*2022-12-06*	*91*	*39*	`LT`/`LA`+ `PD`+ *res.*
M09C	C.O.	1996-04-11	2007-03-06	49	56	`LT`/`LA`+ `PD`+ res.
M11D	Disp.	2007-02-14	2022-12-06	28	149	`LT`/`LA`+ `PD`+ res.
M12D	Disp.	2011-11-17	2022-12-07	63	95	`LT`/`LA`+ `PD`+ res.
M13D	Disp.	2009-01-14	2020-07-28	35	122	`LT`/`LA`+ `PD`+ res.
M14D	Disp.	1991-06-10	1999-08-17	41	74	`LT`/`LA`+ `PD`+ res.

Figure 3.12 illustrates the time series M08C in black, along with 100 similar synthetic time series in colors. The real observations range between -3.4 mm and -2.2 mm exhibiting

an annual periodic pattern and a decreasing trend. The synthetic time series are generated using a SSM with optimized parameters, whose matrices are listed in Appendix A.3. In the next part where we use the synthetic time series to verify the performance of BAR-γ, we use a unified time span of 10 years instead of the original ones.

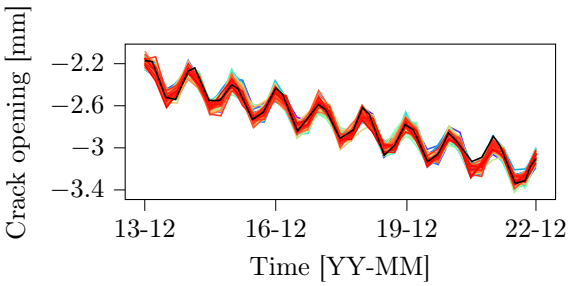


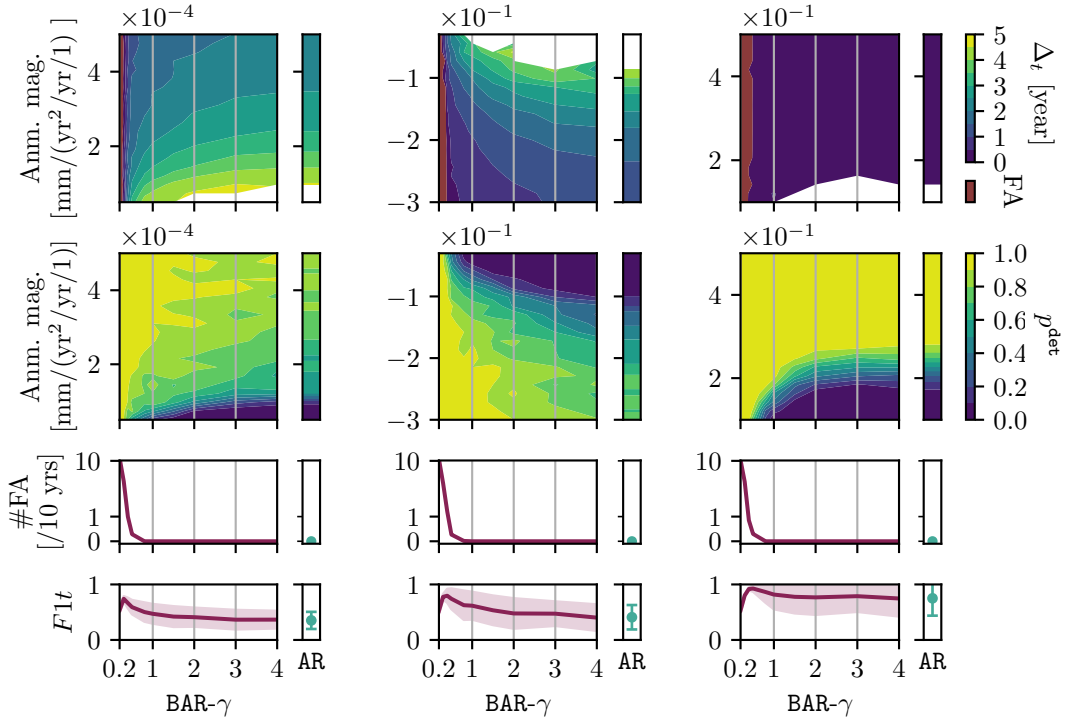
Figure 3.12 100 synthetic time series (colored) similar to a real time series (black) measured on a dam.

Verification of BAR-γ Performance. The time to detection of anomalies may vary according to their starting times, even when they have identical magnitudes. To account for this factor, for each anomaly magnitude and γ value, the BAR-γ model is run on 100 synthetic time series with anomalies introduced at random timestamps within the first five years. An anomaly is considered detected when the probability of the non-stationary regime is higher than 0.5 within a detection window $l_{dw} = 5$ years, so that even if an anomaly is imposed at the end of the first five years, there is enough time to detect it in the second half. An alarm triggered on a time series where there is no simulated anomaly is considered as a false alarm. Whenever a false alarm is triggered, the analysis on this time series is terminated so that the maximum number of false alarms for each time series is one.

Figure 3.13 evaluates BAR-γ and AR using four metrics: (1) the average time to detection Δ_t over the 100 synthetic time series; (2) the detection probability ($p^{\texttt{det}}$) which quantifies, out of the 100 synthetic time series, how many have correctly triggered alarms; (3) the number of false alarms per 10 years; (4) the $F1t$ metric which aggregates the anomaly detectability, time to detection, as well as false alarm rate as described in Section 3.3. In Figure 3.13, the BAR and AR methods are verified on the acceleration, trend and level anomalies presented in Section 2.1.2. The optimal parameter γ^* for detecting the acceleration anomalies using BAR is found to be 0.3 according to its highest $F1t$ score of 0.74, where the expected number of false alarms is 3.5 per ten years. Compared to AR, BAR-0.3 exhibits a better performance with a shorter time to detection and higher detection probability, as shown in the first and the second rows of plots by dark blue and yellow colors, respectively. The white regions in the

Δ_t subfigures indicate non-detectable anomalies and the dark red regions mark the negative time to detection caused by false alarms (FA). By increasing γ, i.e., relaxing the constraints imposed on the `AR` hidden state, the `BAR` eventually converges to an `AR` component with both methods having a similar performance in the four tested metrics.

Similar convergences are observed in the detection of trend and level anomalies, where `BAR`-0.4 and `BAR`-0.5 perform best with expected $F1t$ scores of 0.80 and 0.93, respectively. When a level anomaly occurs, both SKF-BAR and SKF-AR models either detect it with little delay or fail to trigger an alarm altogether. This is because level anomalies only shift the hidden state $x^{\mathtt{L}}$ at one timestamp after which they switch back to the stationary regime, in contrast to the accumulated changes with the trend or acceleration anomalies. The additional verification on the other seven time series is provided in Appendix C.



(a) Acceleration anomaly (b) Trend anomaly (c) Level anomaly

Figure 3.13 Performance of the `BAR` method under different γ on synthetic time series with three types of anomalies in comparison with the `AR`. Evaluation metrics are time to detection (Δ_t), detection probability ($p^{\mathtt{det}}$), false alarm number (#FA) per ten years and $F1t$ score. The white regions on the top three subfigures indicate the undetectable anomalies and their dark red regions represent all the negative time to detection caused by the false alarms (FA). False alarm rate is drawn in logarithmic scale. A higher value of $F1t$ indicates better performance. Repetition $\mathtt{N} = 100$.

The expected values and standard deviations of the $F1t$ scores for BAR-γ^* and AR verified on the synthetic datasets with acceleration anomalies are summarized in Figure 3.14 for the eight real time series. Similar plots for trend and level anomalies are included in Appendix D. The $F1t$ scores for the three types of anomalies, as well as the optimal parameter γ^* and the false alarm rate under each γ^*, are listed in Table 3.2. Through a better tuning of the γ parameter, BAR significantly outperforms AR with higher expected $F1t$ scores and lower standard deviations. $F1t$ scores of BAR that are 80% higher than those of AR are highlighted in Table 3.2.

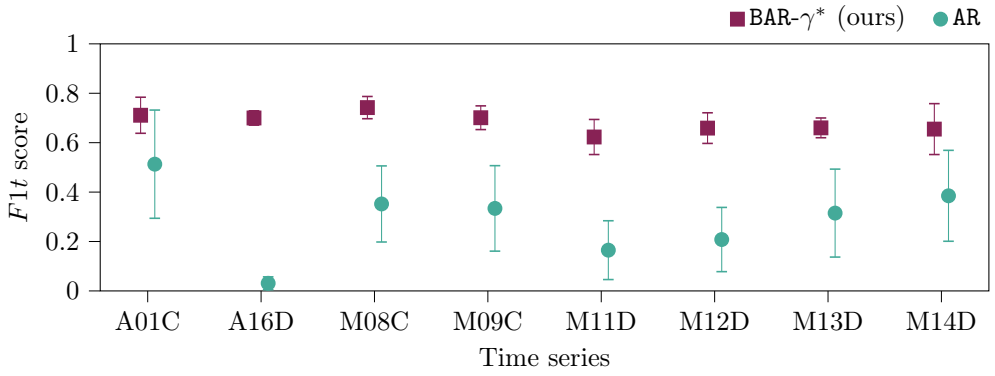


Figure 3.14 The $F1t$ scores for BAR-γ^* and AR verified on the synthetic datasets similar to the eight real time series measured on a dam. Acceleration anomalies are overlaid on the synthetic time series. Additional $F1t$ scores with respect to the trend and level anomalies are shown in Appendix D.

Validation on Real Data with Synthetic Anomalies. A validation experiment on the real time series M08C with synthetic acceleration anomalies is shown here to confirm the verification results obtained above. Additional validation results for other time series and anomaly types are included in Appendix E. The detection window is set to be half the number of data points. Acceleration anomalies of magnitudes varying between 10^{-5} and 10^{-3} mm/year^2 are randomly added to the first half of the time series, while the second half is reserved as the detection window. In order to validate the entire time series, the same process is also repeated on a copy of the time series for which the time direction is flipped. Here, γ is set to 0.4 instead of the optimal value of 0.3 as suggested by the highest $F1t$ score in order to reduce the false alarm rate to the target value of one per ten years.

Figure 3.15 shows the time to detection and the probability of regime switch achieved by the BAR-0.4 model in comparison with the AR model. In the second plot of Figure 3.15, the faded circles provide an overview of the time to detection for each real time series with anomalies introduced at a random location, while the solid lines and the shaded regions

Table 3.2 Summary of AR vs. BAR-γ^* performance. γ^* is the optimal parameter corresponding to the highest $F1t$ in the BAR method. **Bolded** numbers indicate the $F1t$ scores improved by more than 80% in BAR compared to AR.

		AR		BAR-γ^* (ours)			
Anm.	TS	$\mathbb{E}(F1t)$	$\sigma(F1t)$	$\mathbb{E}(F1t)$	$\sigma(F1t)$	γ^*	#FA/10 yrs
Acceleration	A01C	0.513	0.219	0.711	0.073	2.0	1.533
	A16D	0.031	0.026	**0.700**	0.029	1.2	6.749
	M08C	0.352	0.154	**0.742**	0.045	0.3	3.466
	M09C	0.334	0.173	**0.701**	0.048	0.4	3.009
	M11D	0.165	0.119	**0.623**	0.071	0.5	1.853
	M12D	0.208	0.130	**0.659**	0.062	0.4	1.865
	M13D	0.315	0.178	**0.660**	0.040	0.4	5.635
	M14D	0.385	0.184	0.655	0.103	0.5	1.032
Trend	A01C	0.302	0.211	**0.659**	0.140	2.2	0.661
	A16D	N/A	N/A	**0.686**	0.044	1.2	6.083
	M08C	0.409	0.220	**0.796**	0.122	0.4	1.207
	M09C	0.194	0.137	**0.643**	0.038	0.4	3.739
	M11D	0.380	0.202	**0.740**	0.102	0.5	1.782
	M12D	0.429	0.224	0.771	0.076	0.4	1.585
	M13D	0.443	0.255	0.772	0.108	0.5	1.382
	M14D	0.402	0.244	0.712	0.127	0.5	1.376
Level	A01C	0.466	0.291	0.536	0.226	2.4	0.250
	A16D	0.901	0.213	0.898	0.214	4.0	0.000
	M08C	0.750	0.312	0.927	0.055	0.5	0.298
	M09C	0.792	0.304	0.886	0.137	0.5	0.801
	M11D	0.692	0.358	0.932	0.103	0.8	0.038
	M12D	0.795	0.299	0.897	0.137	0.5	0.494
	M13D	0.791	0.318	0.862	0.075	0.5	1.403
	M14D	0.694	0.355	0.859	0.063	0.5	1.226

represent their means and standard deviations for each anomaly magnitude. The BAR-0.4 consistently exhibits a higher detection probability than the AR until they both converge to one, at which point it has the potential to detect anomalies twice as fast as the AR. The smallest detectable anomaly magnitude of BAR-0.4 is around 2×10^{-5} mm/year^2 compared to 7×10^{-5} mm/year^2 for the AR. These results are consistent with those in Figure 3.13. Take the range of anomaly magnitude between 2×10^{-4} and 4×10^{-4} as an example, the results in Figure 3.13 indicate that the time to detection of BAR-0.4 lies in the interval between 1 and 2 years with a probability close to one, while the AR has a time to detection ranging from 2 to 3.5 years with a probability between 0.6 and 0.8. Similar ranges are observed in Figure 3.15 as well, indicating that the model evaluation on the synthetic time series is an effective representation of the models' performance on the real time series.

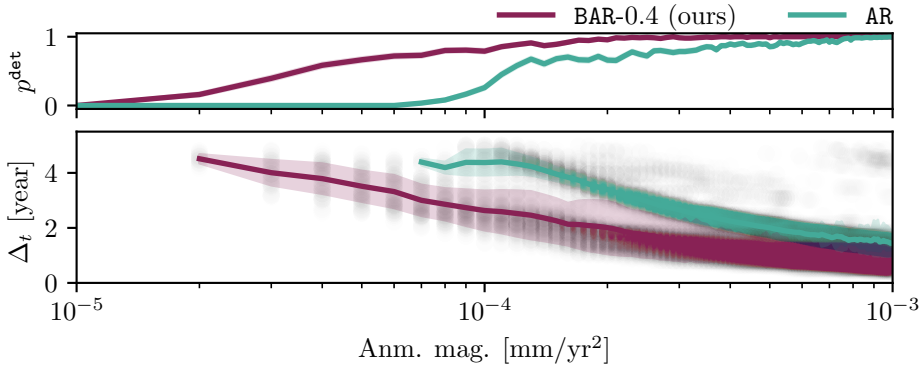


Figure 3.15 Validation results with regard to the anomaly detection probability (p^{det}) and time to detection (Δ_t) on the real time series M08C with acceleration anomalies of different magnitudes by using BAR-0.4 and AR components. The hyperparameter γ is set to 0.4 in BAR. Additional validation results for other time series and anomaly types are reported in Appendix E.

3.5 Conclusions

This study presents a new approach within the existing SKF framework for improving the detection of anomaly for SHM time series. The method proposes a novel residual component consisting of a constrained autoregressive hidden state. The performance of this new approach is evaluated using synthetic as well as bridge and dam datasets using a stochastic performance evaluation scheme involving three common types of anomalies encountered in infrastructure measurements. The experimental results demonstrate that the bounded autoregressive model surpasses the performance of the existing autoregressive model with (1) an improved accuracy in estimating hidden states, (2) an early detection of anomalies, (3) a capacity to detect

smaller anomaly magnitudes, and (4) the ability to control the tradeoff between the anomaly detectability and the false alarm rate. The anomaly detectability, the time to detection, and the false alarm rate are all integrated into the new $F1t$ evaluation metric. Overall, these findings highlight the effectiveness of the proposed method in enhancing the detection of anomalies in infrastructure.

CHAPTER 4 Anomaly Detection via Residual Stationarity Intervention

4.1 Introduction

As demonstrated in Chapter 3, imposing constraints on the autoregressive (`AR`) component improves the SKF's anomaly detectability. Nevertheless, a key limitation of the SKF remains: as reviewed in Section 2.3.1, after a detection, the SKF still requires a period to adapt its baseline to the new stationary regime. This chapter presents a new method for applying interventions on a SSM in order to accelerate its adaptation to a new regime, while maintaining the stationarity of its residuals. In this approach, referred to as *Residual Stationarity Intervention* (RSI), interventions are predicted by a Bayesian neural network (BNN) and are only applied to SSMs if the probability of anomaly exceeds a threshold. This probability is quantified jointly from the data likelihood and the probability that the residual hidden state remains stationary. The interventions reassign the irreversible changes induced by anomalies to the baseline components, preventing them from being captured by the residuals. In addition to the interventions, the BNN predicts the amount of time that has elapsed between the start of the anomaly and the time when the intervention is triggered. This allows accounting for the inevitable delays in the detection by intervening in the SSM retrospectively, enabling SSMs to adapt more quickly to new regimes.

This chapter is structured as follows. Section 4.2 presents the details of the RSI method. Section 4.3 presents a comparative performance evaluation of RSI against existing anomaly detection approaches using both synthetic and real time series. The main contributions of this chapter include:

- a two-step detection process for changepoints between stationary regimes, which first detects anomalies and then intervenes in SSMs to adapt to the new stationary regime;
- a BNN-based intervention model to predict both the magnitude and the timing of the intervention;
- an evaluation method to verify the proposed method against other approaches on synthetic time series and to validate it on real ones.

4.2 Methodology

As already depicted in Figure 2.14, the autoregressive (`AR`) residual component in a SSM is typically more flexible than the baseline components and often captures part of the drift

caused by anomalies in time series. In the RSI method proposed in this section, we leverage such drifts in the residuals in order to detect anomalies. The overall procedure of the RSI method is illustrated in Figure 4.1. Following the regular Kalman-filter step, an auxiliary drift SSM is employed to further decompose the residual estimates, which allows extracting the irreversible drift in the trend of the residuals, denoted as the $x^{\texttt{Td}}$ hidden state. RSI detects anomalies by monitoring the stationarity of this hidden state in the quantification of the probability of anomaly. When the probability of anomaly exceeds a threshold, an intervention model predicts the changes required for the hidden states to adapt to a new stationary regime. Section 4.2.1 presents the procedure to isolate the non-stationary drift from the stationary residuals. Section 4.2.2 presents how to quantify the probability of anomaly by combining the likelihood of the data with the probability that the residual remains stationary, and Section 4.2.3 describes the intervention model relying on a BNN.

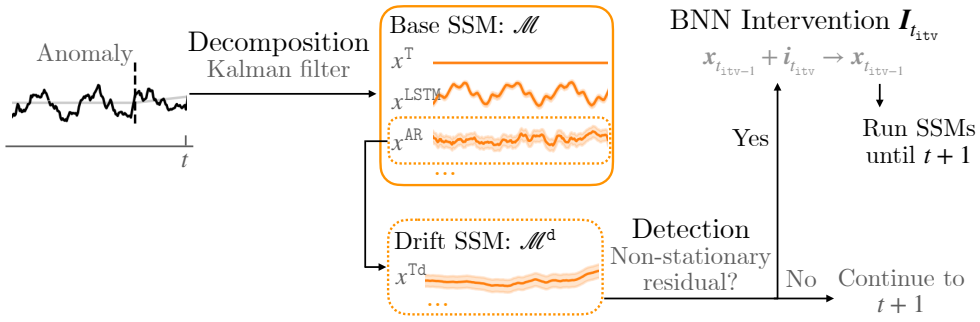


Figure 4.1 Overview of the RSI procedure from time t to $t+1$, illustrating the key methodologies covered in this section: residual's drift modeling with an auxiliary SSM, residual stationarity monitoring, and BNN-driven intervention.

4.2.1 Residual's Drift Model

As presented in Section 2.4.2, the AR component captures stationary residuals as well as drifts caused by anomalies. The residual's drift can be isolated from its stationary part using an auxiliary SSM consisting of a local trend and an autoregressive component ($\mathcal{C}$), referred to as the *drift model* (d), $\mathcal{M}^{\texttt{d}} = \{\mathcal{C}^{\texttt{LTd}}, \mathcal{C}^{\texttt{ARd}}\}$. The drift model further decomposes the base SSM AR's posterior hidden state $X^{\texttt{AR}}_{t|t} \sim \mathcal{N}(\mu^{\texttt{AR}}_{t|t}, (\sigma^{\texttt{AR}}_{t|t})^2)$ into the state vector $\boldsymbol{x}^{\texttt{d}} = [x^{\texttt{Ld}}, x^{\texttt{Td}}, x^{\texttt{ARd}}]^\intercal$. The transition and observation models of the drift model are

$$\boldsymbol{x}^{\texttt{d}}_t = \mathbf{A}^{\texttt{d}}\boldsymbol{x}^{\texttt{d}}_{t-1} + \boldsymbol{w}^{\texttt{d}}_t, \quad \mathbf{W}^{\texttt{d}} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{Q}^{\texttt{d}}\right),$$
$$\hat{x}^{\texttt{AR}}_t = \mathbf{F}^{\texttt{d}}\boldsymbol{x}^{\texttt{d}}_t + v^{\texttt{d}}_t, \quad V^{\texttt{d}} \sim \mathcal{N}\left(0, \left(\sigma^{\texttt{d}}_V\right)^2\right),$$

where $\boldsymbol{x}_t^{\mathtt{d}}$ is a realization of the prior hidden-states $\boldsymbol{X}_{t|t-1}^{\mathtt{d}} \sim \mathcal{N}(\boldsymbol{\mu}_{t|t-1}^{\boldsymbol{X}^{\mathtt{d}}}, \boldsymbol{\Sigma}_{t|t-1}^{\boldsymbol{X}^{\mathtt{d}}})$, and $\hat{x}_t^{\mathtt{AR}}$ is a realization of the prior predictive for the AR of the base SSM, i.e., $\hat{X}_{t|t-1}^{\mathtt{AR}} \sim \mathcal{N}(\hat{\mu}_{t|t-1}^{\mathtt{AR}}, (\hat{\sigma}_{t|t-1}^{\mathtt{AR}})^2)$. The $\mathbf{A}^{\mathtt{d}}$, $\mathbf{Q}^{\mathtt{d}}$, $\mathbf{F}^{\mathtt{d}}$ matrices are detailed in Appendix F.1. The posterior hidden states in the drift model are obtained using the RTS smoother equations [88] following

$$
\begin{aligned}
\boldsymbol{\mu}_{t|t}^{\boldsymbol{X}^{\mathtt{d}}} &= \boldsymbol{\mu}_{t|t-1}^{\boldsymbol{X}^{\mathtt{d}}} + \mathbf{J}_t(\mu_{t|t}^{\mathtt{AR}} - \hat{\mu}_{t|t-1}^{\mathtt{AR}}), \\
\boldsymbol{\Sigma}_{t|t}^{\boldsymbol{X}^{\mathtt{d}}} &= \boldsymbol{\Sigma}_{t|t-1}^{\boldsymbol{X}^{\mathtt{d}}} + \mathbf{J}_t \left((\sigma_{t|t}^{\mathtt{AR}})^2 - (\hat{\sigma}_{t|t-1}^{\mathtt{AR}})^2 \right) (\mathbf{J}_t)^{\intercal}, \\
\mathbf{J}_t &= \boldsymbol{\Sigma}_{t|t-1}^{\boldsymbol{X}^{\mathtt{d}}} \left(\mathbf{F}^{\mathtt{d}} \right)^{\intercal} \left(\hat{\sigma}_{t|t-1}^{\mathtt{AR}} \right)^{-2}.
\end{aligned}
$$

Returning to the example shown in Figure 2.14, the drift can now be isolated by the $x^{\mathtt{Td}}$ hidden state as shown in Figure 4.2, where the posterior hidden states from the drift model are plotted in orange. The $x^{\mathtt{Td}}$ hidden state provide a clean state representation while remaining sensitive to drifts caused by anomalies, offering a more distinguishable separation from the stationary regime than the base SSM residuals $x_{t|t}^{\mathtt{AR}}$. In Section 4.2.2, we will rely on the $x_{t|t}^{\mathtt{Td}}$ to quantify the probability that the residual remains stationary, and in Section 4.2.3 to learn the intervention required on the hidden states of the base SSM.

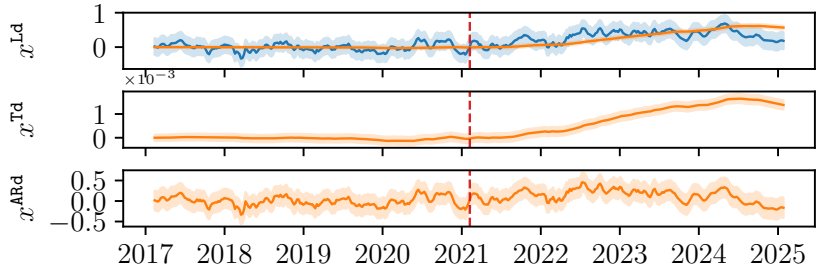


Figure 4.2 Residual decomposition with the drift model. The posterior hidden states from the drift model are plotted in orange. The residuals plotted with $x^{\mathtt{Ld}}$ in the first subfigure are estimated using the base SSM in the example shown in Figure 2.14.

4.2.2 Probability of Anomaly

In this section, we present how to identify anomalies by combining the probability that the residual remains stationary along with the data likelihood. It concludes with a discussion explaining why alternative methods, such as supervised- or reinforcement-based methods, have difficulties coping with the detection of slowly-developing anomalies.

The probability density describing stationary residuals is quantified using a set of synthetic realizations of $\mu_{t|t}^{\mathtt{Td}}$, the expected value for the trend of the drift model, generated under a stationary condition. We rely on synthetic data because typically the real training and

validation sets only consist of short stationary segments which limits their ability to generalize to unforeseen cases. To address this issue, we use the base SSM to generate $\mathtt{N}$ synthetic time series of length $\mathtt{L}$, $\{\boldsymbol{y}^1_{1:\mathtt{L}}, \boldsymbol{y}^2_{1:\mathtt{L}}, ..., \boldsymbol{y}^{\mathtt{N}}_{1:\mathtt{L}}\}$. Both the base and drift SSMs are then applied on these time series, along with the real training and validation sets, to collect realizations of $\mu^{\mathtt{Td}}_{t|t}$. The expected values for the trend hidden state of the drift model are then used to estimate the parameters of a Gaussian distribution, $\mathcal{N}(\mu^{\mathtt{SR}}, (\sigma^{\mathtt{SR}})^2)$, describing the *stationary residual* ($\mathtt{SR}$).

We consider two possible control actions, $\mathcal{A} = \{a_0, a_1\}$, applied to the posterior state, $\boldsymbol{X}_{t-1|t-1}$, of the base model before applying the Kalman filter at each time step t. The action a_0 corresponds to "*do nothing*", leaving $\boldsymbol{X}_{t-1|t-1}$ unchanged, while a_1 corresponds to "*intervene*" on the base and drift models so that an unexpected drift wrongfully captured by the $\mathtt{AR}$ is reassigned to the baseline component. Under a_1, the posterior states $\boldsymbol{X}_{t-1|t-1}$ are modified by transferring the local level and trend components from the drift model to the base model, according to

$$X^{\mathtt{L}}_{t-1|t-1} \leftarrow X^{\mathtt{L}}_{t-1|t-1} + X^{\mathtt{Ld}}_{t-1|t-1}, \tag{4.1}$$

$$X^{\mathtt{T}}_{t-1|t-1} \leftarrow X^{\mathtt{T}}_{t-1|t-1} + X^{\mathtt{Td}}_{t-1|t-1}, \tag{4.2}$$

and resets the expected value of the latent states associated with the baseline in the drift SSM following

$$\mu^{\mathtt{Ld}}_{t-1|t-1} \leftarrow 0,$$
$$\mu^{\mathtt{Td}}_{t-1|t-1} \leftarrow \mu^{\mathtt{SR}}.$$

Depending on the action selected, $a \in \{a_0, a_1\}$, the Kalman filter begins with one of two possible hidden states $\boldsymbol{X}^a_{t-1|t-1}$, resulting in different observation predictions $Y^a_{t|t-1}$ and posterior drift estimates $X^{\mathtt{Td},a}_{t|t}$. The data likelihood is

$$f_{Y^a_{t|t-1}}(y_t) = \mathcal{N}\left(y_t; \mu^{Y,a}_{t|t-1}, \left(\sigma^{Y,a}_{t|t-1}\right)^2\right),$$

and the likelihood that the expected value for the trend of the drift model is a realization from the $\mathtt{SR}$ distribution is

$$f_{\mathtt{SR}}(\mu^{\mathtt{Td},a}_{t|t}) = \mathcal{N}\left(\mu^{\mathtt{Td},a}_{t|t}; \mu^{\mathtt{SR}}, \left(\sigma^{\mathtt{SR}}\right)^2\right).$$

By assuming that the observation and the stationary residual prior predictives are indepen-

dent, and having the action's prior probability $p(a)$, the joint likelihood of observing data y_t, having a stationary residual, and taking action a_i is computed following

$$f(y_t, \mu_{t|t}^{\texttt{Td}}, a|\boldsymbol{y}_{1:t-1}) = f_{Y_{t|t-1}^{a}}(y_t) \cdot f_{\texttt{SR}}(\mu_{t|t}^{\texttt{Td},a}) \cdot p(a).$$

The probability of anomaly is quantified by the probability that a_1 is required to maintain a stationary residual along with the predictive likelihood of y_t, normalized over the two possible actions,

$$\begin{aligned} p_t^{\texttt{anm}} = p(a_1|\boldsymbol{y}_{1:t}, \mu_{t|t}^{\texttt{Td}}) &= \frac{f(y_t, \mu_{t|t}^{\texttt{Td}}, a_1|\boldsymbol{y}_{1:t-1})}{f(y_t, \mu_{t|t}^{\texttt{Td}}|\boldsymbol{y}_{1:t-1})} \\ &= \frac{f(y_t, \mu_{t|t}^{\texttt{Td}}, a_1|\boldsymbol{y}_{1:t-1})}{\sum_i f(y_t, \mu_{t|t}^{\texttt{Td}}, a_i|\boldsymbol{y}_{1:t-1})}. \end{aligned}$$

A $p_t^{\texttt{anm}}$ value above a specified threshold, $\hat{p}^{\texttt{anm}}$, triggers an alarm for a potential anomaly. The method for estimating that threshold is detailed in Appendix G.

Alternative anomaly detection methods, such as supervised classification and reinforcement learning, could also be included as a decision layer for SSMs. However, any method relying on a residual's drift is affected by the delayed adaptation described in Section 2.4.2, where residuals falsely appear to be stationary when the magnitude of the anomaly is small relative to the residuals. In a supervised classification setting, such residuals labeled as abnormal are mixed in the training set with residuals from stationary regimes that are labeled as normal. Assigning two different labels to similar stationary residual inputs leads to ambiguous predictions. In a reinforcement learning framework, agents encounter stationary states both during stationary regimes and at the early stage of an anomaly development when the magnitude is still lower than the residuals. This leads them to associate two conflicting optimal actions with the same state. Therefore, both approaches are prone to false alarms and unreliable detection during the adaptation period. In contrast, using the anomaly probability proposed in this section provides a simple tunable threshold that can distinguish normal from abnormal states.

4.2.3 Intervention Model

When an anomaly is detected at time $t_{\texttt{det}}$, where $p^{\texttt{anm}} \geq \hat{p}^{\texttt{anm}}$, an intervention is applied permanently to correct the hidden states of the base SSM so that it adapts to the data from the new regime. Formally, the intervention applied at time $t_{\texttt{itv}}$ is modelled as a Gaussian additive correction, $\boldsymbol{i}_{t_{\texttt{itv}}} : \boldsymbol{I}_{t_{\texttt{itv}}} \sim \mathcal{N}(\boldsymbol{i}_{t_{\texttt{itv}}}; \boldsymbol{\mu}_{t_{\texttt{itv}}}^{\boldsymbol{I}}, (\boldsymbol{\Sigma}_{t_{\texttt{itv}}}^{\boldsymbol{I}})^2)$, introduced into the transition step of

the Kalman filter, modifying the state vector according to

$$\boldsymbol{x}_{t_{\texttt{itv}}} = \mathbf{A}(\boldsymbol{x}_{t_{\texttt{itv}-1}} + \boldsymbol{i}_{t_{\texttt{itv}}}) + \boldsymbol{w}_{t_{\texttt{itv}}}. \tag{4.3}$$

The intervention defined in Eqs. 4.1-4.2 assigns the drift wrongfully captured by the residual to the baseline component of the base SSM. However, this intervention is typically insufficient for allowing the base SSM to fully adapt to a new trend, because the residual only captures a part of the drift caused by anomalies, as reviewed in Section 2.4.2.

In this section, we propose using a Bayesian neural network (BNN) to predict a target intervention that compensates for the baseline's drift caused by an anomaly. The BNN is trained using the tractable approximate Gaussian inference (TAGI) method introduced in Section 2.5.2. The training set, $\mathcal{D} = \{\tau_1, \tau_2, ...\}$, is collected through the simulation of synthetic anomalies, where we use the base SSM to generate a set of $\texttt{N}$ synthetic stationary time series of length $\texttt{L}$, and apply each time an anomaly with random magnitude and starting time, yielding $\{\boldsymbol{y}_{1:\texttt{L}}^{\texttt{anm}1}, \boldsymbol{y}_{1:\texttt{L}}^{\texttt{anm}2}, ..., \boldsymbol{y}_{1:\texttt{L}}^{\texttt{anm N}}\}$. These time series are then processed with the base and drift SSMs in order to generate the input-target training pairs for the BNN.

With $q \in \mathbb{Z}_{\geq 0}$ being a non-negative integer, we start to collect training samples from the time step $2^q + 1$ in all synthetic time series with anomalies. For each time step t in the n^{th} time series, $\boldsymbol{y}_{2^q+1:\texttt{L}}^{\texttt{anm}\ n}$, one training sample τ_j is collected as an input-target pair. The input consists of the current expected value for $X_{t|t}^{\texttt{Td}}$ together with its estimates over a logarithmically spaced time history, according to

$$\text{input}_j = \mu_{\boldsymbol{\mathcal{H}}}^{\texttt{Td}} = \{\mu_{t|t}^{\texttt{Td}},\ \mu_{t-1|t-1}^{\texttt{Td}},\ \mu_{t-2|t-2}^{\texttt{Td}},\ \mu_{t-4|t-4}^{\texttt{Td}},\ ...,\ \mu_{t-2^q|t-2^q}^{\texttt{Td}}\}. \tag{4.4}$$

The target that needs to be learned by the BNN is defined as

$$\text{target}_j = \begin{cases} \{0, 0, 0\}, & \text{if } t < t_0^{\texttt{anm}} \\ \{x_t^{\texttt{L,anm}}, x_t^{\texttt{T,anm}}, d_t^{\texttt{anm}}\}, & \text{otherwise} \end{cases},$$

where $x_t^{\texttt{L,anm}}$ and $x_t^{\texttt{T,anm}}$ are the level and trend of the synthetic anomaly ($\texttt{anm}\ n$) at the time t, $t_0^{\texttt{anm}}$ denotes the anomaly starting time, and $d_t^{\texttt{anm}} = t - t_0^{\texttt{anm}}$ is the anomaly duration between the current time and its starting time.

The BNN trained on the dataset $\mathcal{D}$ can generalize to unforeseen anomalies. For future data, it takes as input the historical expected values ($\mu_{\boldsymbol{\mathcal{H}}}^{\texttt{Td}}$) for the trend of the drift model, and predicts Gaussian distributions for three anomaly features at the intervention time, representing the predicted corrections to the level $\hat{X}_{t_{\texttt{itv}}}^{\texttt{L,anm}}$, the trend $\hat{X}_{t_{\texttt{itv}}}^{\texttt{T,anm}}$, and the duration since the anomaly

started $\hat{D}^{\texttt{anm}}_{t_{\texttt{itv}}}$,

$$\hat{X}^{\texttt{L,anm}}_{t_{\texttt{itv}}} \sim \mathcal{N}\left(\hat{\mu}^{\texttt{L,anm}}_{t_{\texttt{itv}}}, (\hat{\sigma}^{\texttt{L,anm}}_{t_{\texttt{itv}}})^2\right),$$
$$\hat{X}^{\texttt{T,anm}}_{t_{\texttt{itv}}} \sim \mathcal{N}\left(\hat{\mu}^{\texttt{T,anm}}_{t_{\texttt{itv}}}, (\hat{\sigma}^{\texttt{T,anm}}_{t_{\texttt{itv}}})^2\right),$$
$$\hat{D}^{\texttt{anm}}_{t_{\texttt{itv}}} \sim \mathcal{N}\left(\hat{\mu}^{D,\texttt{anm}}_{t_{\texttt{itv}}}, (\hat{\sigma}^{D,\texttt{anm}}_{t_{\texttt{itv}}})^2\right).$$

These predicted distributions are then used to construct the intervention $\boldsymbol{I}_{t_{\texttt{itv}}}$ correcting the level and trend at the intervention time $t_{\texttt{itv}}$, following

$$t_{\texttt{itv}} = t_{\texttt{det}} - \lfloor \hat{\mu}^{D,\texttt{anm}}_{t_{\texttt{itv}}} \rceil, \tag{4.5}$$
$$I^{\texttt{T}}_{t_{\texttt{itv}}} \sim \mathcal{N}\left(\hat{\mu}^{\texttt{T,anm}}_{t_{\texttt{itv}}}, (\hat{\sigma}^{\texttt{T,anm}}_{t_{\texttt{itv}}})^2\right), \tag{4.6}$$
$$I^{\texttt{L}}_{t_{\texttt{itv}}} \sim \mathcal{N}\left(0, (\hat{\sigma}^{\texttt{L,anm}}_{t_{\texttt{itv}}})^2\right), \tag{4.7}$$

where $\lfloor \cdot \rceil$ denotes rounding to the nearest integer. The equations 4.5–4.7 instantiate the general form of Eq. 4.3: the intervention time $t_{\texttt{itv}}$ is sent back from the detection time $t_{\texttt{det}}$ by the predicted anomaly duration $\hat{D}^{\texttt{anm}}$ to compensate for the detection delay, while $I^{\texttt{T}}_{t_{\texttt{itv}}}$ and $I^{\texttt{L}}_{t_{\texttt{itv}}}$ fill the trend and level entries of $\boldsymbol{\mu}^{I}_{t_{\texttt{itv}}}$ and $\boldsymbol{\Sigma}^{I}_{t_{\texttt{itv}}}$, with all other entries set to zero. In parallel with introducing $\boldsymbol{i}_{t_{\texttt{itv}}}$ into the transition step of the base SSM following Eq. 4.3, the expected values of the hidden states associated with the baseline in the drift SSM are reset following

$$\mu^{\texttt{Ld}}_{t_{\texttt{itv}}-1|t_{\texttt{itv}}-1} \leftarrow 0, \tag{4.8}$$
$$\mu^{\texttt{Td}}_{t_{\texttt{itv}}-1|t_{\texttt{itv}}-1} \leftarrow \mu^{\texttt{SR}}, \tag{4.9}$$

so that the drift components no longer carry the non-stationary pattern that has been reassigned to the baseline of the base SSM. The two SSMs are then run forward from $t_{\texttt{itv}}$ to the current time step. Fig. 4.3 illustrates this retrospective correction on both SSMs.

With the methods proposed in this section, the Kalman filter at time step t is processed as follows: First, we quantify the probability of anomaly $p^{\texttt{anm}}_t$ as described in Section 4.2.2. If $p^{\texttt{anm}}_t$ is below a specified threshold, the base and drift SSMs proceed with the Kalman filter following Sections 2.3.1 and 4.2.1. Otherwise, we apply the intervention $\hat{\mu}^{D,\texttt{anm}}_t$ steps prior to the detection time, according to the BNN predictions obtained following Section 4.2.3. The SSMs processing is then sent back to the time step $t_{\texttt{itv}}$ with the intervention applied to the posteriors from the time step $t_{\texttt{itv}} - 1$, onto which the Kalman filter is reapplied recursively until the current time t corresponding to the anomaly detection time.

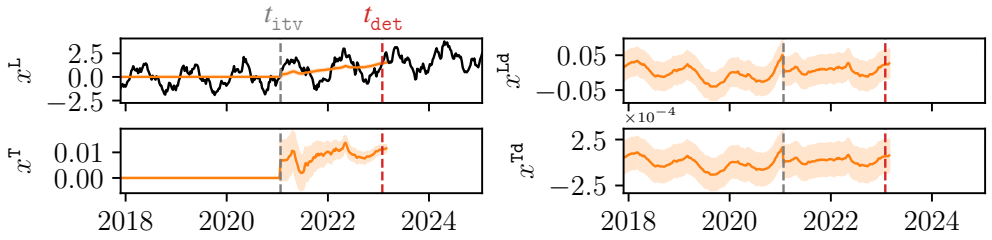


Figure 4.3 Illustration of the retrospective intervention. The intervention is applied at the intervention time $t_{\texttt{itv}}$ (gray dashed line), sent back from the detection time $t_{\texttt{det}}$ (red dashed line) by the predicted anomaly duration (Eq. 4.5). *Left two panels (base SSM):* the level $x^{\texttt{L}}$ and trend $x^{\texttt{T}}$ are corrected at $t_{\texttt{itv}}$ by the interventions $I^{\texttt{L}}_{t_{\texttt{itv}}}$ and $I^{\texttt{T}}_{t_{\texttt{itv}}}$ (Eqs. 4.6–4.7), reassigning the non-stationary pattern to the baseline of the base SSM. *Right two panels (drift SSM):* the baseline hidden states of the drift model, $x^{\texttt{Ld}}$ and $x^{\texttt{Td}}$, are reset at $t_{\texttt{itv}}$ following Eqs. 4.8–4.9. Both SSMs are then run forward from $t_{\texttt{itv}}$ to the current step.

The RSI method is well suited for large-scale infrastructure monitoring thanks to its low computational demand. Compared to a standard Kalman filter, the additional cost introduced by RSI is primarily associated with the execution of an auxiliary Kalman filter corresponding to the residual's drift model. The dominant time complexity of the base Kalman filter is $\mathcal{O}(n^3)$ per time step, where n denotes the number of hidden states, compared to a constant time complexity required by the auxiliary Kalman filter with 3 hidden states. When an intervention is triggered, the forward pass of the TAGI neural network has a time complexity of $\mathcal{O}(w)$, where w is the number of network parameters, and reprocessing the Kalman filters for k time steps in the past incurs a cost of $\mathcal{O}(kn^3)$. Consequently, the maximum time complexity of the RSI method occurs at time steps where an intervention is required and is given by $\mathcal{O}\left((k+1)n^3 + w\right)$; such events are rare in practice.

For a single time series in our experiments, the average computational time of RSI is approximately 0.01 s per time step, resulting in a total running time that scales linearly with the length of the series. The memory requirement remains constant over time, with a usage of approximately 385 MB. Once the base SSM-LSTM and the intervention model are pretrained, RSI can update more than 200 K hourly time series in an online setting on an CPU with 32 GB of memory.

4.3 Experiments

In this section, we compare the Residual Stationarity Intervention (RSI) method, which builds upon the open-source library Canari [`https://github.com/Bayes-Works/canari`], with three classical anomaly detection approaches, namely the SKF (reviewed in Section 2.3.1),

Matrix Profile (Section 2.2.1), and Prophet (Section 2.2.2), as well as two deep-learning-based methods reviewed in Section 2.2.3: LSTMED and TranAD. The implementation, together with a notebook example using synthetic time series, is available on GitHub: `https://github.com/zhanwen-xin/canari-rsi/tree/rsi-paper`. Section 4.3.1 presents the comparison on synthetic time series, and Section 4.3.2 evaluates the performance of the methods on real ones through a verification and validation process.

4.3.1 Synthetic Time Series

This section compares the performance of six detection methods on synthetic time series. It first presents two synthetic time series with different residual complexities, along with the anomaly generation procedure, and then probabilistically compares the anomaly detection performance of the six methods.

Description of Dataset. Using SSMs, we generate two stationary synthetic time series of length $\mathtt{T} = 12$ years with weekly intervals, as illustrated in Figure 4.4. Both SSMs share identical baseline and periodic components, while differing only in their residual components. The baseline component consists of a zero-valued level and trend throughout the entire time series. The periodic component is a superposition of two Fourier forms with periods of 52 and 13 weeks. Their residuals are generated with autoregressive ($\mathtt{AR}$) components having the same process error but different autocorrelation coefficients: $\phi^{\mathtt{AR}} = 0.5$ for the simple time series and $\phi^{\mathtt{AR}} = 0.9$ for the complex one. The autocorrelation coefficient determines the complexity of the time series, as larger values of $\phi^{\mathtt{AR}}$ amplify short-term fluctuations, which may be misinterpreted by detection models as baseline shifts, thereby causing false alarms. The model matrices for the two generative SSMs are provided in Appendix F.2.

The two SSMs are then used to generate different stationary time series realizations, to which synthetic anomalies are added for the probabilistic quantification of anomaly detectability. Each test series contains a single synthetic anomaly that shifts the zero-valued trend to a new constant magnitude, selected from {0.01, 0.02, 0.03, ..., 0.1, 0.2, 0.3, ..., 1} unit/year in its normalized space. For each magnitude, we test ten different anomalies, each at a randomly set starting time between 40% of the series length and three years before the end. The first 40% of the data is not involved in the evaluation and kept identical across all time series. Figure 4.5 illustrates one anomaly example for each magnitude, with its starting time marked by a dashed line.

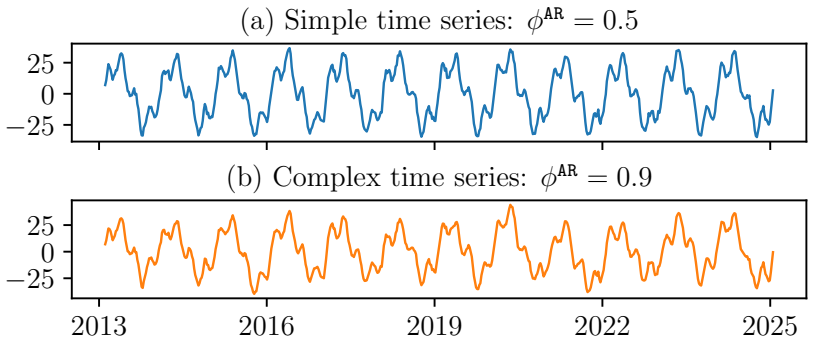


Figure 4.4 A simple and a complex synthetic time series generated using 2 different SSMs. The baselines and periodic components for them are the same. The residuals of the simple time series are generated with $\phi^{\mathtt{AR}} = 0.5$, and those of the complex one are generated with $\phi^{\mathtt{AR}} = 0.9$.

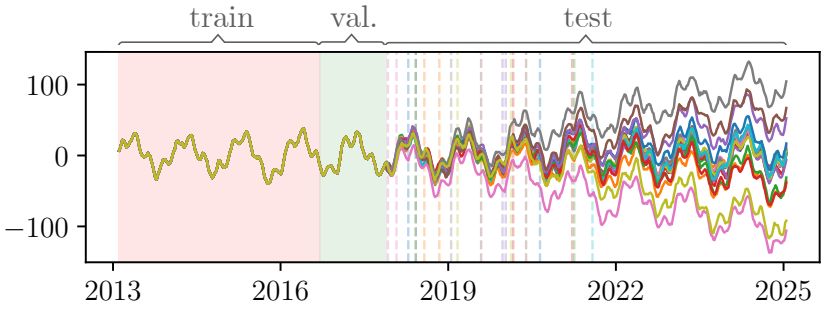


Figure 4.5 Examples of synthetic anomalies in different time series realizations generated by the SSM with $\phi^{\mathtt{AR}} = 0.9$. The starting times of anomalies with varying magnitudes are indicated by the vertical dashed lines. The training, validation (val.), and test sets are marked on top of the time series.

Results. The six detection models (i.e., RSI, SKF, Matrix Profile, Prophet, LSTMED, and TranAD) are configured following the experiment setup specified in Appendix G. RSI, SKF, LSTMED and TranAD rely on the training and validation sets as illustrated in Figure 4.5 to train models that describe the stationary part of the time series. For Matrix Profile, the stationary subsequence spaces are constructed using the combined training and validation sets. For Prophet, the model is retrained at each time step using all past observations.

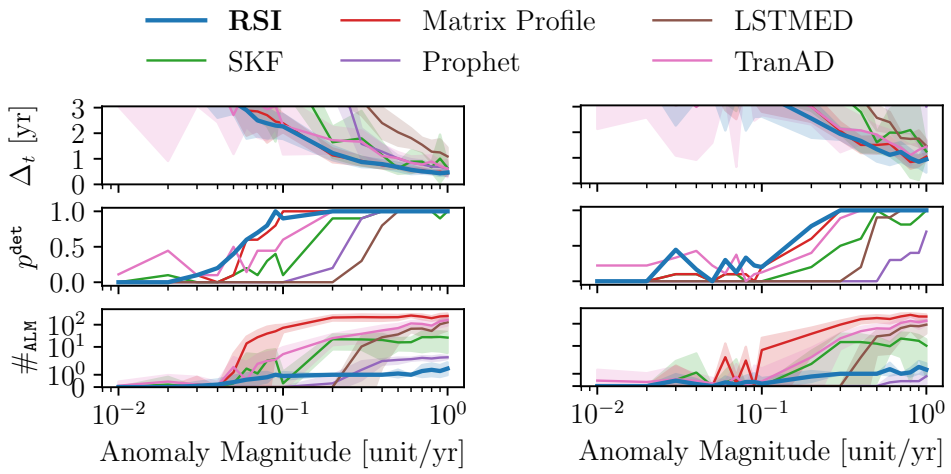
The anomaly thresholds for all six methods are calibrated on the entire synthetic stationary time series presented in Figure 4.4. The false alarm rates on the test set are reported in Table 4.1. All detection models except TranAD exhibit false alarm rates close to zero on the simple time series owing to the low variability of its residuals. In contrast, only Prophet and LSTMED maintain a zero false alarm rate on the complex time series, while RSI and SKF each trigger approximately one false alarm every 10 years. However, Matrix Profile produces more than eight false alarms/10 years and TranAD produces more than four due to their limited generalizability to unseen data outside the stationary subsequence space and their poor adaptation following detection.

Table 4.1 False alarm rate on the test set (#/10 years) of six detection models on two datasets.

Model	Simple time series	Complex time series
RSI	0.02	1.26
SKF	0.17	1.03
Matrix Profile	0.00	8.65
Prophet	0.00	0.00
LSTMED	0.00	0.00
TranAD	3.56	4.44

The time series with false alarms are excluded from the comparison of time to detection (Δ_t), detection probability ($p^{\texttt{det}}$), and the number of alarms ($\#_{\texttt{ALM}}$) that are presented in Figure 4.6. The time to detection is defined as the difference between the first time when an alarm is triggered and the anomaly starting time, with a maximum time to detection set to three years. The expected value, along with the standard deviation for the time to detection for each anomaly magnitude, is plotted in the first row of Figure 4.6. The detection probability reported in the second row is computed as the ratio between the number of anomalies detected within three years and the total number of test series under the same anomaly magnitude. We also count the number of alarms triggered on each time series to evaluate how well each model adapts to the new trend induced by anomalies. Because each time series in the test sets contains only one anomaly, all the alarms after the first are regarded as redundant ones required for the adaptation to new regimes. The expected values and standard deviations

for the number of alarms are reported in the third row.



(a) Simple time series (b) Complex time series

Figure 4.6 Comparison of six detection models on simple and complex time series with anomalies with respect to time to detection (Δ_t), detection probability ($p^{\texttt{det}}$), and number of alarms ($\#_{\texttt{ALM}}$) across a range of anomaly magnitudes. The anomaly magnitude and the number of alarms ($\#_{\texttt{ALM}}$) are plotted on a logarithmic scale. Solid lines correspond to the expected value, and the shaded regions indicate the $\pm 1\sigma$ confidence region.

All six models detect small-magnitude anomalies with shorter time to detection in the simple time series compared to the complex one. Our method and Matrix Profile achieve the best performance in both time to detection and detection probability across both simple and complex time series, followed by TranAD, SKF, Prophet and LSTMED. Furthermore, RSI exhibits low sensitivity to hyperparameter choices, with its performance remaining stable across a wide range of configurations on the simple synthetic validation set, as demonstrated by the hyperparameter analysis provided in Appendix H.

Although Matrix Profile shows a similar detectability to RSI, it adapts poorly to new regimes, continuously triggering alarms after the first detection. This occurs because the level of all the subsequences $\boldsymbol{s}_t$ deviates from the zero-valued level in the stationary subsequence space $\mathcal{S}^{\texttt{MP}}$ once an anomaly is introduced. While normalizing subsequences can reduce repeated alarms in Matrix Profile, it severely degrades the performance in terms of time to detection and detection probability, performing even worse than Prophet. Compared to the other methods, RSI consistently triggers only around one alarm corresponding to the single anomaly in each test time series, indicating a superior adaptability to the new regimes. In non-stationary time series, the intervention in RSI corrects the baseline and prevents other components of the Kalman filter, such as the residual, from diverging. Appendix I presents an example

illustrating how the RSI intervention prevents the residual component from drifting away from its stationary distribution and maintains a stable residual variance and autocorrelation.

In terms of computational cost, Prophet is the most demanding, requiring approximately 4 hours to complete the test, followed by RSI at 18 minutes, LSTMED at around 3 minutes, while the remaining methods (SKF, Matrix Profile, and TranAD) each require approximately 1 minute, on an CPU with 32 GB of memory. Overall, our method demonstrates a superior anomaly detectability and adaptability to new regimes compared to the other methods on this synthetic dataset, with a moderate computational cost.

4.3.2 Real Time Series

This section compares the performance of six detection methods on real time series collected on dams in Canada. The real time series are first presented, along with the detrending process, in which baseline shifts are removed to obtain stationary time series for comparison. The six methods are then compared probabilistically by quantifying the detectability of anomalies on the detrended time series, onto which synthetic anomalies are added. Finally, the methods are compared on the original time series.

Description of Dataset and Detrending Process. Figure 4.7 presents ten normalized time series of crack opening and displacement from dams, with lengths ranging from 9 to 18 years, weekly data acquisition frequency, and containing missing data. All time series exhibit baseline shifts that may result from factors such as changes in dam behavior, interventions on the structure, or sensor malfunctions.

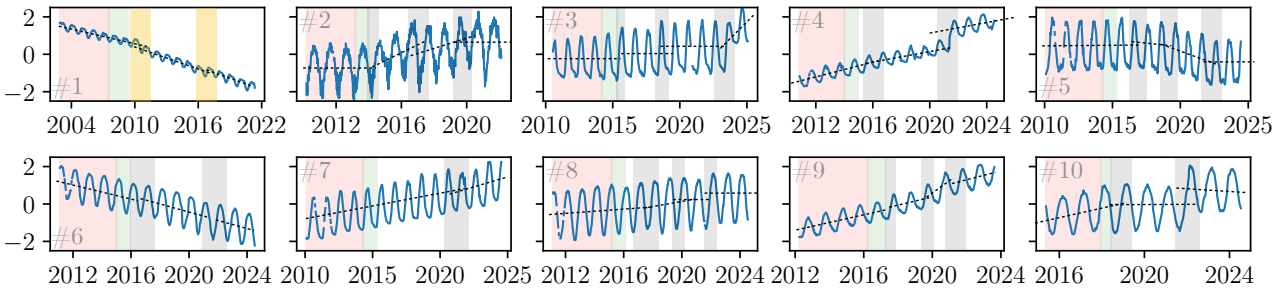


Figure 4.7 Normalized weekly time series of crack opening and displacement on dams, with lengths ranging from 9 to 18 years and containing missing data. The red region is the training set, the green one is the validation set, the yellow one corresponds to the anomaly labeled by dam engineers, and the grey one marks the potential anomalies corresponding to changepoints in the baseline responses.

A preliminary analysis was conducted by identifying shifts in baselines by connecting a linear

trend over the dominant peaks across multiple cycles within each stationary regime. The manually estimated baselines are shown as black dashed lines. In time series #1, for which true anomaly labels, marked by yellow shaded regions, are available, the baseline shifts identified manually coincide with the true anomaly labels. For time series #2-10 where no anomaly labels are available, we identified potential changepoint regions indicated by grey shaded areas, which correspond to the intersection of manually identified baselines. Although manually identifying changepoints is feasible when having decades of data at our disposal, it is a notoriously difficult task to do online as data is collected. Among them, time series #2, 3, 4, 5, 9, and 10 show clear baseline shifts, whereas time series #1, 6, 7, and 8 display more subtle changes.

We cannot systematically evaluate the anomaly detectability of the different methods using the raw data presented in Figure 4.7 because the nature, precise location, and exact number of anomalies present are uncertain. To address this limitation, we first compare the methods on detrended time series onto which we add synthetic anomalies. For that purpose, the baseline of each time series is removed using the Prophet library, thereby correcting the time series to a stationary range centered around zero, as shown in Figure 4.8. The detrended time series do not present clear changepoints like the raw ones, yet short-term reversible fluctuations and variations in periodic patterns remain because of the inherent complexity of real data. Note that because the detrending process relies on the Prophet method, the latter is possibly advantaged by the preprocessing.

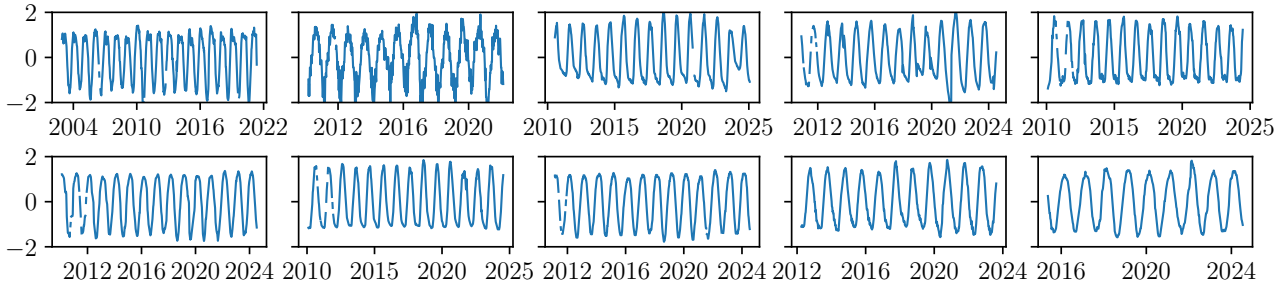


Figure 4.8 Detrended stationary time series obtained by removing the baselines estimated by Prophet from the real time series.

Verification on Detrended Time Series. The anomaly detection performance of the six models is first verified using synthetic anomalies generated according to the procedure described in Section 4.3.1, which are added to the detrended series. The different methods are configured following the procedure presented in Appendix G, with training and validation sets fixed to 30% and 10% of the detrended time series. The stationary subsequence spaces

in Matrix Profile are constructed using the combined training and validation sets. The anomaly thresholds for the different methods are calibrated on the entire detrended time series to prevent false alarms. This ensures a) that all methods operate under the shared assumption that the time series are stationary, and b) that they are maximally sensitive to detect deviations introduced by the synthetic anomalies. Without enforcing the absence of false alarms, it would be difficult, if not impossible, to determine whether a detection corresponds to an anomaly or to short-term variations inherent to the real time series.

Figure 4.9 presents the evaluation metrics (time to detection, detection probability, and number of alarms) averaged across all ten detrended series. Our method, RSI, achieves the highest detection probability with shorter detection delay for all anomaly magnitudes, followed by Matrix Profile, while Prophet, TranAD, and SKF exhibit a similar detectability that is inferior to that of RSI and Matrix Profile. Consistent with the results detailed on synthetic time series, LSTMED displays the weakest detection performance. Although in Section 4.3.1, Matrix Profile matched the detection performance of RSI on synthetic time series, it now underperforms RSI on real data in both detection and adaptation to new regimes, showing longer detection delays, lower detection probabilities, and a large number of redundant alarms. This limitation arises because real time series contain more complex periodic patterns that are difficult to capture with a simple stationary subsequence space. Even though Prophet, TranAD, and SKF exhibit a similar performance in terms of time to detection and detection probability, Prophet is the most computationally demanding among the six methods. The relative running times of the six methods are consistent with those observed on synthetic time series, as reported in Section 4.3.1.

Validation on Real Time Series. We compare the anomaly detection performance of the six methods on the ten raw time series using the manually defined anomaly labels. The proportions of the training and validation sets, which are used by all six methods, are represented by red and green shaded areas plotted in Figure 4.7, so that they end before the first anomaly occurs.

Following the experiment setup presented in Appendix G, RSI, SKF, LSTMED, and TranAD rely on the training and validation sets as illustrated in Figure 4.7 to train models that describe the stationary part of the time series. For Matrix Profile, the stationary subsequence spaces are constructed using the training sets. All subsequences in Matrix Profile are normalized in this experiment to handle multiple anomalies in real data and to prevent anomaly scores from continuously increasing due to deviations in the levels. For Prophet, the model is retrained at each time step using all past observations. The anomaly thresholds for all six methods are calibrated on the combined training and validation sets.

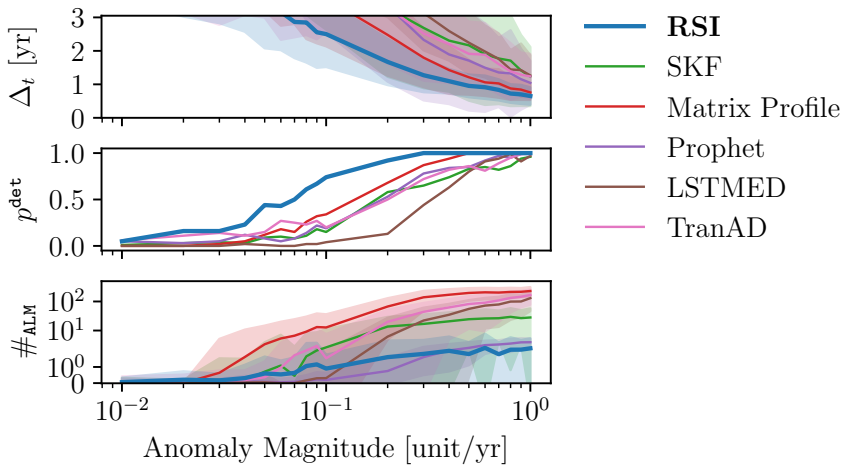


Figure 4.9 Comparison of six detection models on detrended time series with anomalies, evaluated in terms of time to detection (Δ_t), detection probability ($p^{\mathtt{det}}$), and number of alarms ($\#_{\mathtt{ALM}}$). The results are averaged across all the abnormal time series derived from the 10 detrended time series.

Figure 4.10 compares the detections for the six methods within the test sets against the anomaly labels indicated by yellow (true anomalies) and grey (potential anomalies) regions. For time series #1, where true anomaly labels are available, RSI is the fastest to detect the first anomaly, followed by Prophet, TranAD, LSTMED, and SKF. Matrix Profile fails to detect the first anomaly. For each method, the grey number indicates its ranking for its first valid detection of the first anomaly. Although TranAD and LSTMED detect this anomaly at an early stage, they subsequently trigger repeated alarms in the new regime, and therefore fail to provide informative detection thereafter. Following the adaptation to the new regime, the second anomaly provides an additional basis for comparison. In this case, SKF fails to detect the second anomaly, while again RSI detects it earlier than Prophet, yielding the best overall performance on this time series.

For the other time series without true anomaly labels, the manually identified potential anomalies serve only as a reference baseline and may have been defined differently by other annotators. The subjectivity in manual labeling may influence the evaluation results on real data. Consequently, an objective comparison is only feasible with experiments on detrended time series. Despite the limitation of potentially subjective labels, the performance of the six detection methods on these time series can be categorized qualitatively into three groups:

1) *Repeated detections*, as observed for LSTMED and TranAD. Although these methods can rapidly flag the first anomaly, their repeated alarms in the subsequent regimes prevent them from providing informative long-term detection without retraining.

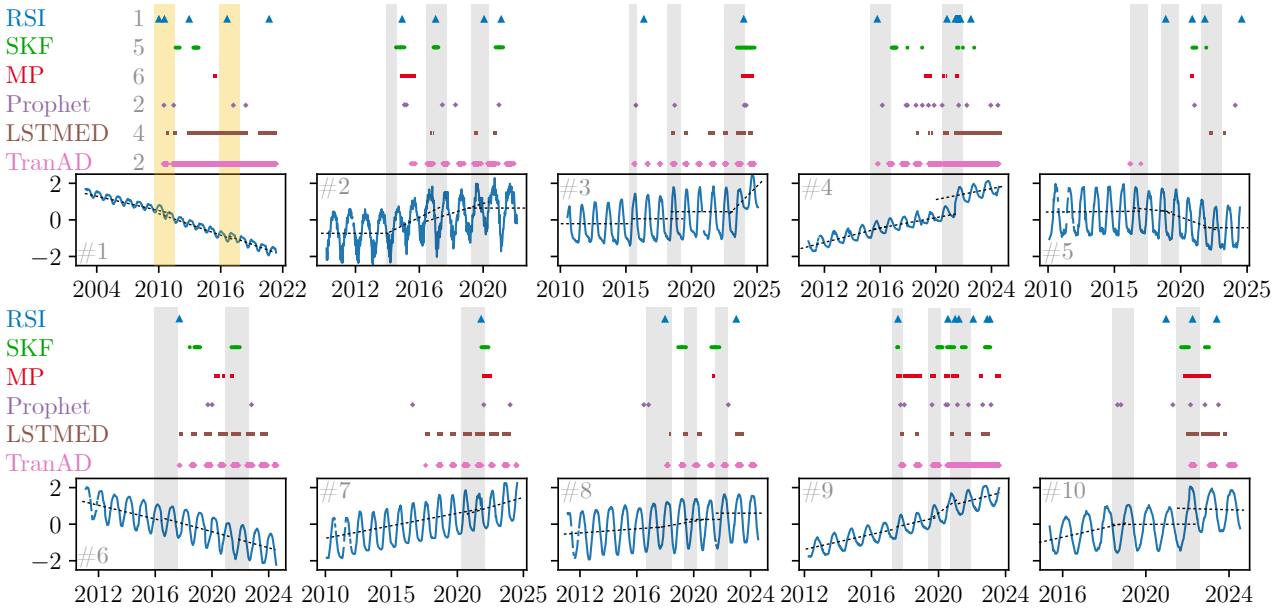


Figure 4.10 Comparison of six methods against confirmed (yellow areas) and potential (grey areas) anomalies on real time series. Grey numbers indicate the ranking of their first valid detection on time series #1. The markers above the time series indicate the alarms triggered by different methods.

2) *Missed or substantially delayed detections*, such as Matrix Profile failing to trigger within the second half of time series #2, and SKF and Matrix Profile triggering only near the end of time series #3.

3) *Valid detections.* RSI exhibits a dominant detectability on time series #4 and #5, and shares the best overall detection performance with other methods in several cases, for example with SKF on time series #2, with both SKF and Matrix Profile on time series #7, and with both SKF and Prophet on time series #8 and #9. Although Prophet is a strong baseline and achieves the best performance on time series #3 and #10, its high computational demand in online operation limits its applicability for large-scale structural health monitoring.

In general, deep-learning-based methods exhibit a worse adaptability to new regimes than the classical methods, as demonstrated by the results presented here as well as in the preceding experiments on synthetic data. This limitation arises for two main reasons: 1) the amount of training data available within a single time series is often insufficient to train deep networks with sophisticated architectures, and 2) these models struggle to generalize to new regimes on which they have not been trained. These issues are further amplified when anomaly detection relies directly on predictions from deep networks, as in TranAD, where anomaly scores are

quantified based on reconstruction errors. Consequently, the development of an adaptive framework is required before such methods can be reliably used in SHM online detection settings.

4.4 Conclusions

This study introduces the RSI method that relies on interventions for SSMs applied to non-stationary time series, enabling the detection of anomalies that shift the data from one stationary regime to another, and facilitating the adaptation to new regimes. The inevitable delay in detecting gradual drifts in SHM data can be mitigated by applying the intervention at the anomaly's starting time predicted by the intervention model. Employing a BNN as the intervention model within the Kalman filtering framework allows both the intervention time and magnitude to be predicted, without the need to predefine or explicitly evaluate them through optimization, as required by typical intervention models. The robustness of the proposed BNN-based intervention model is demonstrated empirically, while its flexibility points to several promising avenues for future research. These include the incorporation of additional inputs into the BNN to further enhance adaptation to new regimes and the extension of the framework to predict different types of interventions. The proposed method outperforms the existing anomaly detection approaches it was compared with on both synthetic and real time series, showing higher detection probabilities and shorter times to detection across anomalies of varying magnitudes. Overall, the results demonstrate that RSI is an effective intervention model for detecting small-magnitude and slow-developing anomalies in infrastructure.

CHAPTER 5 Classification-Guided Residual Stationarity Intervention

5.1 Introduction

The methods developed in the two previous chapters each address part of the SKF's limitations reviewed in Section 2.3.1. On the one hand, the SKF enhanced with the bounded autoregressive (BAR) component presented in Chapter 3 enforces the stationarity of the residual component, preventing anomaly-induced drift from being absorbed into the residuals and thereby enabling faster detection of anomalies of smaller magnitudes. On the other hand, the RSI method presented in Chapter 4 applies interventions directly to the SSM's hidden states, allowing it to adapt quickly to the new regime once an anomaly is detected. Despite these improvements, both methods assume that when an anomaly switches the data to a new regime, the nature of that regime is known in advance: the SKF-BAR still requires the analyst to specify a non-stationary model describing the data from the new regimes, while RSI implicitly assumes that any future anomaly takes the form of a change from one constant trend to another. In practice, the prior knowledge required by the SKF-BAR is rarely available, and the anomaly type assumed by RSI may not match the anomalies actually encountered.

This assumption becomes problematic in long-term SHM, where a structure is likely to experience several anomalies of different natures over its service life. As already reviewed in Section 2.1.2, structural anomalies typically manifest themselves as irreversible baseline changes, either as a shift in the level or as a slow-developing change in the trend. Because RSI was designed exclusively for the latter type, encountering a level anomaly leads it to incorrectly apply a trend intervention; the baseline estimate then shifts away from the data, and repeated false alarms are triggered as the method attempts to compensate for its erroneous trend correction. A method intended for long-term SHM must therefore not only detect anomalies in an online manner, but also identify their types in order to apply appropriate corrections.

This chapter presents the Residual Stationarity Intervention and Classification (RSIC) method to address these limitations. RSIC retains the core advantage of RSI, i.e., the ability to leverage residual's drift as an early indicator of anomaly, while extending it in two important directions. First, RSIC incorporates a level anomaly type in addition to the trend anomaly, enabling it to respond appropriately to both kinds of structural change. Second, RSIC introduces a probabilistic classification mechanism, based on a Beta-Bernoulli formulation with fractional pseudo-count updates, that identifies the anomaly type from the observations accu-

mulated after a detection and applies the corresponding targeted intervention retrospectively at the detection time. By matching the intervention to the inferred anomaly type, RSIC improves the adaptability to the new regime and substantially reduces the false alarm rate that limits the use of RSI in the presence of level anomalies.

This chapter is organized as follows. Section 5.2 presents the RSIC methodology. Section 5.3 evaluates RSIC on both synthetic and real time series, covering a verification study on detrended data and a validation on real observations. The main contributions of this chapter include:

- a classification-guided intervention framework, RSIC, that extends the RSI method to handle both level and trend anomalies;
- a probabilistic anomaly-type classification mechanism, which infers the intervention type online from the observations accumulated after a detection;
- an extensive comparative evaluation of RSIC against six anomaly detection methods on synthetic and real time series, where each test time series contains two sequential anomalies of different types, reflecting the methods' performance under subsequent anomalies in long-term SHM.

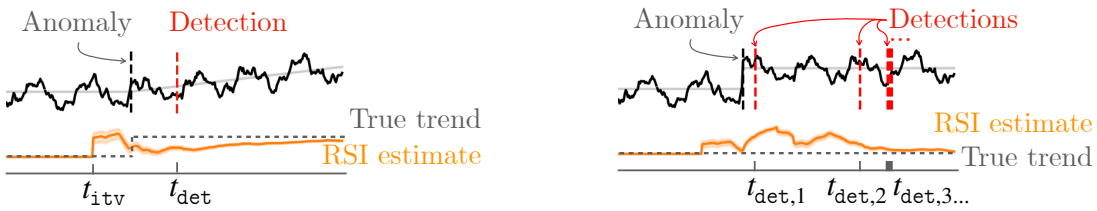
5.2 Methodology

This section presents the RSIC methodology. Section 5.2.1 first reviews the two limitations of the RSI method presented in Chapter 4. An overview of the RSIC framework is provided in Section 5.2.2. Section 5.2.3 then presents how to infer the intervention at the detection time. Section 5.2.4 describes the estimation of the tentative interventions, which have a more accurate mean and no uncertainty attached to them, while Section 5.2.5 describes how the intervention type is classified using the tentative interventions.

5.2.1 Limitations of RSI

As presented in Chapter 4, the RSI method detects anomalies by monitoring the stationarity of the residuals extracted from a SSM, and applies a BNN-driven intervention to the model's latent states when an anomaly is detected. The BNN predicts three anomaly features: the corrections to the level and trend hidden states, as well as the duration elapsed since the anomaly started, from which the intervention time $t_{\texttt{itv}}$ is estimated retrospectively. This approach has been shown to be effective in detecting anomalies in time series transitioning

from one trend-stationary regime to another. However, it presents two limitations, illustrated in Figure 5.1.



(a) Limitation 1 (b) Limitation 2

Figure 5.1 Illustration of the two limitations of the RSI method: (a) early intervention gradually diminishing before the true anomaly starting time, and (b) repeated false alarms triggered by an anomaly only shifting the level.

First, when the predicted intervention time precedes the true anomaly starting time, the uncertainty associated with the intervention applied to the base SSM may gradually shrink, as the model continues to process observations from the pre-anomaly regime. By the time the anomaly actually begins, the intervention may have been largely absorbed, hindering the model's adaptation to the new regime, as illustrated in Figure 5.1a.

Second, the interventions in RSI are inherently designed to correct changes in the trend, such that anomalies involving a shift in the level alone may trigger repeated false alarms as RSI attempts to repeatedly correct a trend that it has erroneously modified, as illustrated in Figure 5.1b.

5.2.2 Overview of RSIC

The RSI's limitations motivate the two key developments introduced in RSIC: first, rather than applying the intervention at an intervention time predicted by a BNN, RSIC infers the intervention from the observations $\boldsymbol{y}_{t_{\texttt{det}}:t}$ collected between the detection time $t_{\texttt{det}}$ and a later time $t > t_{\texttt{det}}$, and applies the intervention retrospectively at $t_{\texttt{det}}$. Second, RSIC is extended to account for two types of anomalies: a shift in the level or in the trend.

Figure 5.2 summarizes the RSIC procedure. After an anomaly is detected at time $t_{\texttt{det}}$, the SSMs continue to process incoming observations without any model intervention, until enough confidence is gained to distinguish the anomaly type at a later time $t > t_{\texttt{det}}$. At each time step, two BNNs predict the interventions $\boldsymbol{I}_t^\tau$, with $\tau \in \{\texttt{L}, \texttt{T}\}$, corresponding to level and trend interventions. They are mapped back to the detection time to obtain $\boldsymbol{I}_{t_{\texttt{det}}}^\tau$. Tentative interventions are then constructed from $\boldsymbol{I}_{t_{\texttt{det}}}^\tau$, and separately applied to their own copy of the

base SSM, producing the prior states $\boldsymbol{X}^{\tau}_{t_{\texttt{det}}|t_{\texttt{det}}-1}$. A model selection procedure determines the intervention type τ that best explains the observed data, and the corresponding intervention is permanently applied to $\boldsymbol{X}_{t_{\texttt{det}}-1|t_{\texttt{det}}-1}$. The Kalman filter then resumes from $t_{\texttt{det}}$ to update the hidden states up to t.

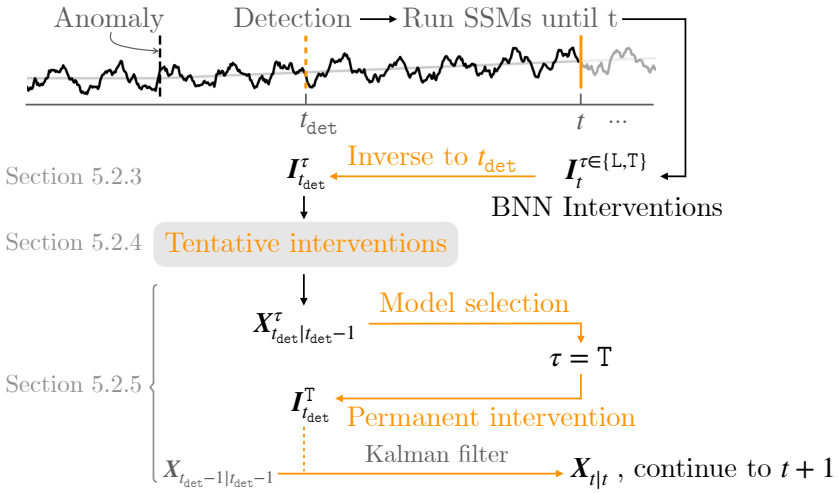


Figure 5.2 Overview of the RSIC procedure using an example where the anomaly is detected at time $t_{\texttt{det}}$, and a trend ($\texttt{T}$) intervention is identified at time t.

5.2.3 Intervention Types and Their Inference

In contrast to RSI which applies an intervention designed to correct the shift in the trend (Eqs. 4.6 and 4.7), RSIC is extended to account for two types of anomalies through the corresponding level ($\boldsymbol{I}^{\texttt{L}}$) and trend ($\boldsymbol{I}^{\texttt{T}}$) interventions. $\boldsymbol{I}^{\texttt{L}} = [I^{\texttt{L}}]$ is added only to the level hidden state, while $\boldsymbol{I}^{\texttt{T}} = [I^{\texttt{LcT}}, I^{\texttt{T}}]^{\intercal}$ modifies both the level and trend hidden states, where $I^{\texttt{LcT}}$ corresponds to the correction to the level caused by the change in trend that has accumulated from the anomaly starting time up to the intervention time. Once an intervention type is identified with sufficient confidence, the corresponding intervention is applied to the posterior state $\boldsymbol{X}_{t_{\texttt{det}}-1|t_{\texttt{det}}-1}$ according to Eq. 4.3 with $t_{\texttt{itv}} = t_{\texttt{det}}$.

The two interventions at the detection time, $\boldsymbol{I}^{\texttt{L}}_{t_{\texttt{det}}}$ and $\boldsymbol{I}^{\texttt{T}}_{t_{\texttt{det}}}$, are inferred as follows: at each time step $t > t_{\texttt{det}}$, two BNNs trained following the procedure in Appendix J separately predict the two interventions, $\boldsymbol{I}^{\texttt{L}}_{t}$ and $\boldsymbol{I}^{\texttt{T}}_{t}$. They take as input the posterior means $\mu^{\texttt{Td}}_{i|i}$ from the drift SSM over a logarithmically spaced time history $\forall i \in \{t_{\texttt{det}}, t_{\texttt{det}}-1, t_{\texttt{det}}-2, t_{\texttt{det}}-4, \ldots, t_{\texttt{det}}-2^{q}\}$,

with $q \in \mathbb{Z}_{\geq 0}$, and output the interventions according to

$$\begin{aligned} \boldsymbol{I}_t^{\mathtt{L}} &= [I_t^{\mathtt{L}}] = \mathrm{BNN}(\mu_{i|i}^{\mathtt{Td}}; \boldsymbol{\theta}^{\mathtt{L}}), \\ \boldsymbol{I}_t^{\mathtt{T}} &= [I_t^{\mathtt{LcT}}, I_t^{\mathtt{T}}]^\intercal = \mathrm{BNN}(\mu_{i|i}^{\mathtt{Td}}; \boldsymbol{\theta}^{\mathtt{T}}). \end{aligned}$$

where $\boldsymbol{\theta}^{\tau}$ represents the network parameters for the intervention type $\tau \in \{\mathtt{L}, \mathtt{T}\}$.

From the BNN predictions obtained for the current time t, the corresponding interventions for the detection time $t_{\mathtt{det}}$ are obtained by reversing the kinematic transition model following

$$\begin{aligned} \boldsymbol{I}_{t_{\mathtt{det}}}^{\tau} &= \tilde{\mathbf{A}}^{-1} \boldsymbol{I}_t^{\tau}, \\ \boldsymbol{\mu}_{t_{\mathtt{det}}}^{\boldsymbol{I},\tau} &= \tilde{\mathbf{A}}^{-1} \boldsymbol{\mu}_t^{\boldsymbol{I},\tau}, \\ \boldsymbol{\Sigma}_{t_{\mathtt{det}}}^{\boldsymbol{I},\tau} &= \tilde{\mathbf{A}}^{-1} \boldsymbol{\Sigma}_t^{\boldsymbol{I},\tau} (\tilde{\mathbf{A}}^{-1})^\intercal, \end{aligned}$$

$$\text{where } \tilde{\mathbf{A}} = \begin{cases} \begin{bmatrix} 1 \end{bmatrix} & \text{for } \tau = \mathtt{L}, \\ \begin{bmatrix} 1 & t - t_{\mathtt{det}} \\ 0 & 1 \end{bmatrix} & \text{for } \tau = \mathtt{T}. \end{cases}$$

The intervention $\boldsymbol{I}_{t_{\mathtt{det}}}^{\tau}$ serves two purposes: 1) it is used to estimate a *tentative intervention* for each of the two intervention types following Section 5.2.4. Each tentative intervention is separately applied to its own copy of the SSM, and the resulting predictions are compared to determine which intervention type best explains the observed data, as detailed in Section 5.2.5; 2) once the intervention type has been identified, the corresponding intervention, $\boldsymbol{I}_{t_{\mathtt{det}}}^{\mathtt{L}}$ or $\boldsymbol{I}_{t_{\mathtt{det}}}^{\mathtt{T}}$, is applied to the model as the *permanent intervention*.

5.2.4 Tentative Interventions

We quantify the probabilities of the two intervention types by separately applying a *tentative intervention* to the base SSM to compare their predictions over the window between the detection time $t_{\mathtt{det}}$ and the current time step t. The intervention $\boldsymbol{I}_{t_{\mathtt{det}}}^{\tau}$ cannot be used directly as the tentative intervention because its uncertainty can hinder the distinguishability by allowing the baseline hidden states to adapt to the new data regardless of the intervention type. For example, a level intervention with sufficient uncertainty can incorrectly cause the baseline to drift gradually, effectively behaving like a trend intervention even it is only intended to model a step shift in the level at a single time step. In order to preserve the structure intended for each intervention type during the classification phase, we instead rely on a tentative intervention that only corrects the expected values of the baseline hidden states. Figure 5.3 illustrates how the tentative intervention is derived from the BNN intervention $\boldsymbol{I}_{t_{\mathtt{det}}}^{\tau}$.

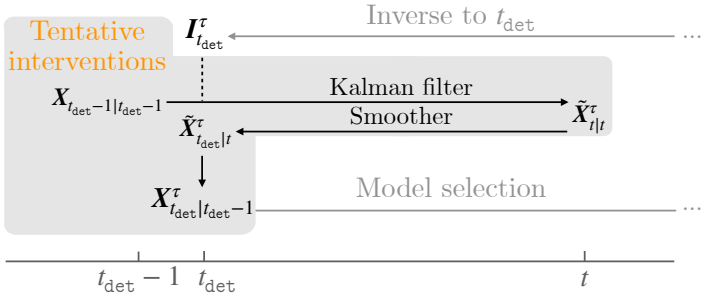


Figure 5.3 Tentative intervention estimation, where the BNN intervention $\boldsymbol{I}^{\tau}_{t_{\text{det}}}$ is refined via Kalman filtering and smoothing to obtain the prior state $\boldsymbol{X}^{\tau}_{t_{\text{det}}|t_{\text{det}}-1}$ used for model selection.

Specifically, we follow Eq. 4.3 to apply the intervention $\boldsymbol{I}^{\tau}_{t_{\text{det}}}$ on $\boldsymbol{X}_{t_{\text{det}}-1|t_{\text{det}}-1}$. The Kalman filter is then run from t_{det} until t to obtain $\tilde{\boldsymbol{X}}^{\tau}_{t|t}$, and the Kalman smoother is applied from t back to t_{det} to obtain $\tilde{\boldsymbol{X}}^{\tau}_{t_{\text{det}}|t} \sim \mathcal{N}\left(\tilde{\boldsymbol{\mu}}^{\tau}_{t_{\text{det}}|t}, (\tilde{\boldsymbol{\Sigma}}^{\tau}_{t_{\text{det}}|t})^2\right)$. The tentative intervention leverages the smoothed state and resets the corresponding expected values of the prior state in the base SSM, while the covariance of the prior state is kept as that of the prediction without intervention. For $\tau \in \{\mathtt{L}, \mathtt{T}\}$, the new prior state at t_{det} is denoted as $\boldsymbol{X}^{\tau}_{t_{\text{det}}|t_{\text{det}}-1}$, and its moments are set according to

$$\begin{aligned}
\mu^{\mathtt{L},\tau}_{t_{\text{det}}|t_{\text{det}}-1} &\leftarrow \tilde{\mu}^{\mathtt{L},\tau}_{t_{\text{det}}|t}, \\
\mu^{\mathtt{T},\tau}_{t_{\text{det}}|t_{\text{det}}-1} &\leftarrow \tilde{\mu}^{\mathtt{T},\tau}_{t_{\text{det}}|t}, \text{ if } \tau = \mathtt{T}.
\end{aligned}$$

The drift model is reset following Eqs. 4.8 and 4.9 along with the tentative interventions.

5.2.5 Classification of Intervention Types

Starting from the two prior states associated with the tentative interventions, $\boldsymbol{X}^{\tau}_{t_{\text{det}}|t_{\text{det}}-1}$, in parallel, two separate Kalman filters process the observations $\boldsymbol{y}_{t_{\text{det}}:t}$, collect the observation prediction Y^{τ}_i, and the posterior means of the residual trend $\mu^{\mathtt{Td},\tau}_{i|i}$ for each time step $i \in \{t_{\text{det}} : t\}$ under each intervention type $\tau \in \{\mathtt{L}, \mathtt{T}\}$. These quantities are then used in a probabilistic model selection framework to classify the intervention type, as illustrated in Figure 5.4. Once the intervention type is identified, the corresponding intervention, either $\boldsymbol{I}^{\mathtt{L}}_{t_{\text{det}}}$ or $\boldsymbol{I}^{\mathtt{T}}_{t_{\text{det}}}$, is permanently applied to $\boldsymbol{X}_{t_{\text{det}}-1|t_{\text{det}}-1}$, and the Kalman filter resumes from t_{det} to t. If no type is identified with sufficient confidence, the SSMs continue to $t+1$.

The prior distribution over the probability of the level intervention, $p^{\mathtt{L}}$, at the detection time

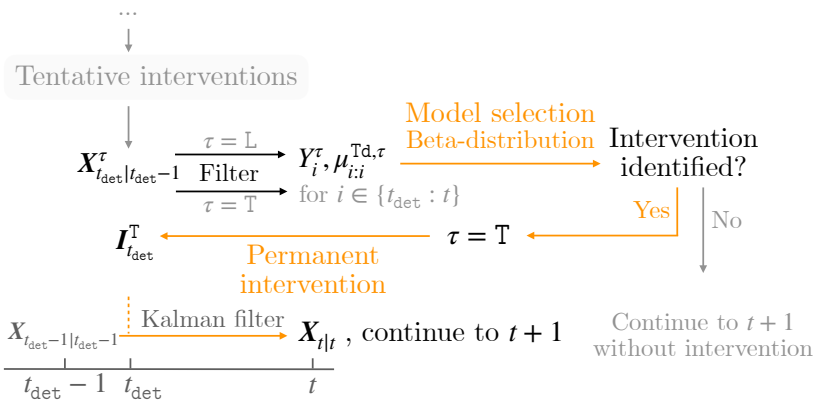


Figure 5.4 Procedure for classifying intervention type and applying a permanent intervention once a trend (T) intervention is identified at time t.

$t_{\mathtt{det}}$, is modeled by a Beta distribution,

$$f_{t_{\mathtt{det}}}(p^{\mathrm{L}}) = \mathcal{B}(p^{\mathrm{L}}; \alpha_0, \beta_0),$$

with prior hyperparameters $\{\alpha_0, \beta_0\} > 0$. The prior distribution for the trend (T) intervention is obtained through the complement following $p^{\mathrm{T}} = 1 - p^{\mathrm{L}}$.

According to the Beta-Bernoulli conjugacy [132], if direct observation z^τ were available at each time step indicating which intervention of $\{\mathtt{L}, \mathtt{T}\}$ is the right one, the posterior hyperparameters $\{\alpha_t, \beta_t\}$ would be obtained following

$$\alpha_t = \alpha_0 + \sum_{i=t_{\mathtt{det}}}^{t} z_i^{\mathrm{L}},$$
$$\beta_t = \beta_0 + \sum_{i=t_{\mathtt{det}}}^{t} z_i^{\mathrm{T}},$$

$$\text{where } \begin{cases} z^{\mathrm{L}} = 1, z^{\mathrm{T}} = 0 \text{ if } \mathtt{L} \text{ is observed }, \\ z^{\mathrm{L}} = 0, z^{\mathrm{T}} = 1 \text{ if } \mathtt{T} \text{ is observed }. \end{cases}$$

However, in our case, the intervention type cannot be directly observed. We therefore approximate the update with a updating scheme in which each binary indicator z_i^τ is replaced by a fractional pseudo-count $\hat{z}_i^\tau \in [0, 1]$, the normalized likelihood of intervention τ given the observation at time step i,

$$\hat{z}_i^\tau = \frac{f_{Y_i^\tau}(y_i) \cdot f_{\mathtt{SR}}(\mu_{i|i}^{\mathtt{Td},\tau})}{\sum_{\tau \in \{\mathrm{L},\mathrm{T}\}} f_{Y_i^\tau}(y_i) \cdot f_{\mathtt{SR}}(\mu_{i|i}^{\mathtt{Td},\tau})},$$

where $f_{Y_i^\tau}(y_i) = \mathcal{N}\left(y_i; \mu_i^{Y,\tau}, \left(\sigma_i^{Y,\tau}\right)^2\right)$ is the likelihood of the observations, and $f_{\mathtt{SR}}(\mu_{i|i}^{\mathtt{Td},\tau}) = \mathcal{N}\left(\mu_{i|i}^{\mathtt{Td},\tau}; \mu^{\mathtt{SR}}, (\sigma^{\mathtt{SR}})^2\right)$ compares the $\mu_{i|i}^{\mathtt{Td},\tau}$ with the stationary residual distribution presented in Section 4.2.2. Assuming the pseudo-counts $\hat{z}_i^\tau$ constitute independent evidence across time steps, the posterior distribution of $p^{\mathtt{L}}$ at the current time step t is constrained to remain within the Beta family,

$$f_t(p^{\mathtt{L}}) = \mathcal{B}(p^{\mathtt{L}}; \alpha_t, \beta_t),$$

with hyperparameters updated using the fractional pseudo-counts

$$\alpha_t = \alpha_0 + \sum_{i=t_{\mathtt{det}}}^{t} \hat{z}_i^{\mathtt{L}},$$
$$\beta_t = \beta_0 + \sum_{i=t_{\mathtt{det}}}^{t} \hat{z}_i^{\mathtt{T}}.$$

This fractional update retains the recursive, closed-form structure of the conjugate case, which is essential for the online setting.

The posterior moments for the Beta distributions are given by

$$\mu_t^{P^{\mathtt{L}}} = \frac{\alpha_t}{\alpha_t + \beta_t},$$
$$(\sigma_t^{P^{\mathtt{L}}})^2 = \frac{\alpha_t\, \beta_t}{(\alpha_t + \beta_t)^2(\alpha_t + \beta_t + 1)},$$

whose p-quantile is then used to classify the intervention type.

Following a detection, let $P_t^\tau \sim \mathcal{B}(\alpha_t, \beta_t)$ at the current time t and let $F^{-1}_{P^\tau,t}$ denote its quantile function. Given two hyperparameters, $\tilde{p}$ for the p-quantile and $\tilde{t}$ for the maximum waiting time, the final intervention $\boldsymbol{I}_{t_{\mathtt{det}}}$ that is permanently applied to the base SSM is determined by

$$\boldsymbol{I}_{t_{\mathtt{det}}} = \begin{cases} \boldsymbol{I}^{\mathtt{L}}_{t_{\mathtt{det}}}, \text{ if } F^{-1}_{P^{\mathtt{L}},t}(\tilde{p}) > 0.5 \text{ and } t \le t_{\mathtt{det}} + \tilde{t}, \\ \boldsymbol{I}^{\mathtt{T}}_{t_{\mathtt{det}}}, \text{ if } F^{-1}_{P^{\mathtt{T}},t}(\tilde{p}) > 0.5 \text{ and } t \le t_{\mathtt{det}} + \tilde{t}, \\ \mathbf{0} \text{ and continue to } t+1, \text{ otherwise.} \end{cases}$$

The resulting intervention is then applied to the corresponding components of the posterior state vector $\boldsymbol{X}_{t_{\mathtt{det}}-1|t_{\mathtt{det}}-1}$ following Eq. 4.3, and the drift model is reset following Eqs. 4.8 and 4.9 if $\boldsymbol{I}_{t_{\mathtt{det}}} \neq 0$. The sensitivity of RSIC to these two hyperparameters and their tuning are provided in Appendix K.

Compared to RSI, the computational cost of RSIC is increased only within the classification

windows; outside them, RSIC incurs the same cost as RSI, i.e., one Kalman filter step for each of the base and drift SSMs per observation. During a classification window, each new time step additionally requires two parallel Kalman filter passes from $t_{\texttt{det}}$ to t and one Kalman smoother pass from t back to $t_{\texttt{det}}$, i.e., a per-step cost of $\mathcal{O}((t - t_{\texttt{det}})\, n^3)$ with state dimension n, and a cumulative cost per detection of $\mathcal{O}(\tilde{t}^2 n^3)$ in the worst case where no type is identified before the maximum waiting time. This cost is bounded by design through $\tilde{t}$, and in practice the window is much shorter: the mean time for classification ranges from 1.7 to 3.1 years across configurations (Table K.2), against $\tilde{t} = 5$ years. RSIC remains computationally tractable for online deployment in long-term SHM.

5.3 Experiments

This section compares the RSIC method against the original RSI method, and five other anomaly detection methods reviewed in Section 2.2: the Switching Kalman Filter (SKF), Discord Aware Matrix Profile (DAMP), Prophet, LSTMED, and TranAD. The experimental setup for every method is detailed in Appendix G, and the implementation of the RSIC is released on GitHub [`https://github.com/Bayes-Works/canari/tree/dev-anm-type%2Frsic-paper`]. Section 5.3.1 evaluates the seven methods on synthetic time series, and Section 5.3.2 compares their performance on real time series through a verification and validation process.

5.3.1 Synthetic Time Series

This section evaluates the performance of seven anomaly detection methods on synthetic time series. It first describes the generation process of the dataset, and then presents a probabilistic comparison of the different detection methods.

Description of Dataset. The synthetic time series are constructed by introducing anomalies on stationary base time series. The base time series are generated at a weekly sampling interval from a SSM combining three types of components: (i) a baseline component with zero level and zero trend, (ii) two periodic components specified in Fourier form with periods of 52 and 13 weeks, and (iii) an autoregressive component with a process-error standard deviation of 5 and an autocorrelation coefficient of 0.5.

All time series share the same training and validation sets, consisting of the first 250 observations (approximately five years). The test set comprises independent realizations of the base time series drawn from the same SSM.

Two anomalies are introduced sequentially in the test set of each time series. Two anomaly types are considered: a level (L) anomaly, defined as a shift from a constant level to another constant level, and a trend (T) anomaly, defined as a shift from a constant trend to another constant trend. Combining these two types yields four possible sequential configurations: L→L, L→T, T→L, and T→T.

The first anomaly, of fixed magnitude, is introduced at a random time step between one and two years after the end of the validation set. Six years of observations are then available for its detection. The second anomaly is subsequently introduced with varying magnitude, followed by another six years of observations for its detection. The detection performance of all methods is evaluated on the second anomaly, as this setting reflects the performance under sequential anomalies, a critical requirement for long-term structural monitoring as discussed in Section 2.1.3. For each magnitude of the second anomaly, 100 independent realizations of the test time series are generated, and the methods are evaluated across these 100 repetitions. In total, each synthetic time series spans 19 years: approximately 5 years of training and validation data, up to 2 years between the validation set and the starting time of the first anomaly, 6 years for detecting the first anomaly, and 6 years for detecting the second anomaly. Figure 5.5 illustrates representative time series containing an initial level anomaly followed by a trend anomaly.

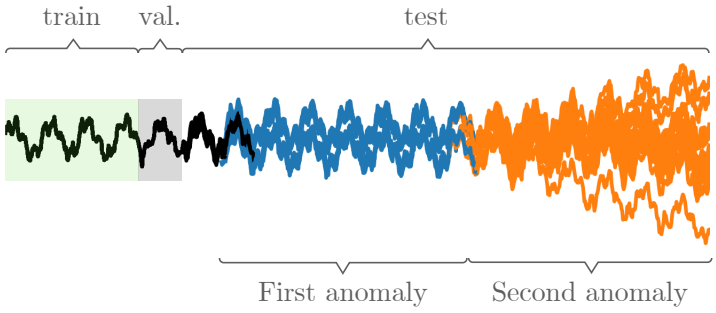


Figure 5.5 Examples of synthetic time series with two types of anomalies: (i) a level anomaly (blue), and (ii) a trend anomaly (orange). The training and validation (val.) sets are identical across all time series, while the test sets correspond to different realizations of the same SSMs with varying anomaly magnitudes and starting times.

Results. We assess the seven methods using four metrics: time to detection, detection probability, false alarm rate, and mean squared error (MSE) of the baseline estimation. All metrics are computed with respect to the detection of the second anomaly. The time to detection Δ_t is defined as the time difference between the starting time of the second anomaly and the first detection occurring thereafter. The detection probability $p^{\mathtt{det}}$ is computed as the

proportion of test time series in which the second anomaly is correctly identified. The false alarm rate is computed over the period spanning from the time step at which the first anomaly is detected to the end of each test time series; in this period, only the first detection after the starting time of the second anomaly is counted as a true detection, and all consecutive false detections across multiple time steps are collapsed into a single false alarm. Finally, the baseline MSE compares the estimated level and trend with their true values, both computed in the standardized space.

The detection performance for the first anomaly is not reported here, as the first detection occurs before any intervention is applied and RSIC is therefore equivalent to RSI, whose corresponding results have been analyzed in Chapter 4. When evaluating on the second anomaly, an alarm cannot be unambiguously attributed to it if the first anomaly was never detected. The evaluation is therefore restricted to a common subset of realizations: for each anomaly configuration and magnitude, only the realizations in which every method detected the first anomaly are retained, and all metrics are computed on this identical subset. In the $\mathtt{T} \to \mathtt{L}$ configuration, LSTMED triggers no alarms; it is therefore reported as producing no detections and is excluded from the definition of the common subset in that configuration.

Figure 5.6 reports the time to detection Δ_t and the detection probability $p^{\mathtt{det}}$ for the four anomaly configurations, and Table 5.1 summarizes the false alarm rate, the MSE for level and trend estimation, and the RSIC classification accuracy for both anomalies. Because DAMP, LSTMED, and TranAD do not produce explicit baseline estimates, they are evaluated only on time to detection, detection probability, and false alarm rate.

Overall, RSIC matches RSI on both time to detection and detection probability. However, RSI exhibits an increased false alarm rate whenever level anomalies are present, whereas RSIC maintains a stable false alarm rate below one per ten years across all anomaly configurations. LSTMED and TranAD achieve high detection probabilities for small anomaly magnitudes as displayed in Figure 5.6, but at the cost of substantially higher false alarm rates, indicating limited robustness in time series with multiple anomalies. Their detection probability also becomes unstable at larger magnitudes, as the anomaly shifts the time series far outside the training data distribution, where these methods fail to generalize. This sensitivity extends beyond anomaly magnitude: LSTMED fails to trigger any alarms in the $\mathtt{T} \to \mathtt{L}$ configuration, suggesting that it is also sensitive to the combination of anomaly types. Prophet likewise shows strong detectability, at times comparable to RSIC in detection probability, but its false alarm rate exceeds those of RSIC, RSI, and SKF, while remaining below LSTMED and TranAD. SKF, in contrast, achieves consistently low false alarm rates and accurate baseline estimation, yet requires a longer time to detection and yields a lower detection probability

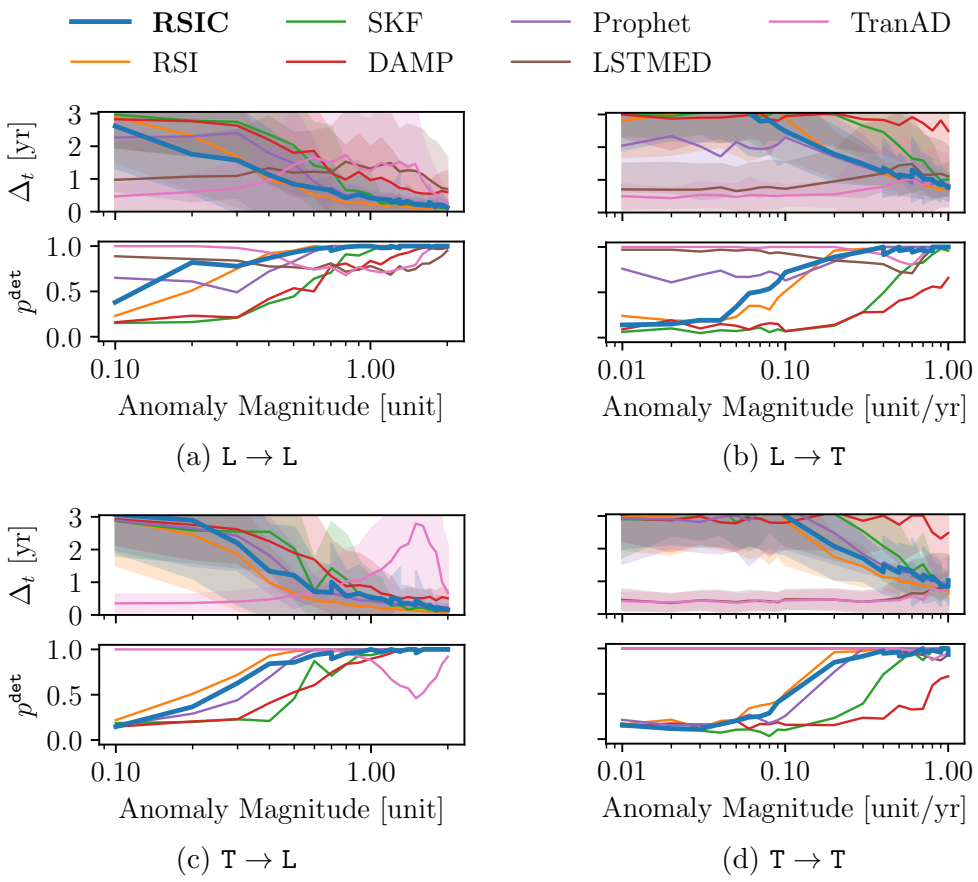


Figure 5.6 Time to detection (Δ_t) and detection probability ($p^{\mathtt{det}}$) of the tested methods on synthetic time series under different anomaly combinations: level to level (L→L), level to trend (L→T), trend to level (T→L), and trend to trend (T→T).

Table 5.1 Comparison of false alarm rate, baseline estimation errors (in standardized space) across different methods, and RSIC classification performance on synthetic time series with various anomaly combinations: level to level (L→L), level to trend (L→T), trend to level (T→L), and trend to trend (T→T).

Anomaly	Methods	# false alarm per 10 years	MSE_L	MSE_T	Classification accuracy	
					1st anm	2nd anm
L → L	RSIC	**0.97**	0.066	$\mathbf{1.38 \times 10^{-5}}$	0.79	0.48
	RSI	6.12	**0.050**	6.58×10^{-5}	-	-
	SKF	1.22	0.090	1.50×10^{-4}	-	-
	DAMP	0.13*	-	-	-	-
	Prophet	6.68	0.107	3.46×10^{-5}	-	-
	LSTMED	11.78	-	-	-	-
	TranAD	13.87	-	-	-	-
L → T	RSIC	0.88	0.037	$\mathbf{1.51 \times 10^{-5}}$	0.78	0.81
	RSI	2.66	**0.030**	3.19×10^{-5}	-	-
	SKF	**0.79**	0.056	7.33×10^{-5}	-	-
	DAMP	0.38*	-	-	-	-
	Prophet	5.29	0.058	2.47×10^{-5}	-	-
	LSTMED	12.42	-	-	-	-
	TranAD	14.38	-	-	-	-
T → L	RSIC	**0.77**	0.040	$\mathbf{1.17 \times 10^{-5}}$	0.99	0.26
	RSI	4.69	**0.037**	4.56×10^{-5}	-	-
	SKF	1.05	0.059	8.99×10^{-5}	-	-
	DAMP	0.12*	-	-	-	-
	Prophet	3.76	0.073	2.19×10^{-5}	-	-
	LSTMED	-	-	-	-	-
	TranAD	14.19	-	-	-	-
T → T	RSIC	**0.59**	0.021	1.15×10^{-5}	0.99	0.73
	RSI	1.07	**0.018**	1.26×10^{-5}	-	-
	SKF	0.61	0.028	1.76×10^{-5}	-	-
	DAMP	0.34*	-	-	-	-
	Prophet	1.95	0.026	$\mathbf{1.01 \times 10^{-5}}$	-	-
	LSTMED	15.47	-	-	-	-
	TranAD	15.11	-	-	-	-

* The false alarm rate is obtained by reporting the top-2 discords as detections, such that DAMP cannot produce more than two detections per time series.

than both RSI and RSIC. Among all evaluated methods, RSIC offers the best balance between detectability and false alarm rate.

DAMP reports the lowest false alarm rate across all anomaly configurations in Table 5.1, marked with $(*)$ to indicate that this result is not directly comparable to the other methods. Unlike the remaining methods, which operate in a fully online setting without prior knowledge of the number of anomalies, DAMP detects anomalies by relying on the top-2 discords [47]. Despite this structural advantage, DAMP does not achieve superior detectability: as shown in Figure 5.6, its detection probability is comparable to SKF in the $\texttt{L} \rightarrow \texttt{L}$ and $\texttt{T} \rightarrow \texttt{L}$ scenarios, but remains lower than RSI and RSIC in most scenarios, with a generally longer time to detection.

Beyond detection, RSIC also provides anomaly-type classification, a capability not offered by any of the competing methods. It attains high classification accuracy for the first anomaly regardless of type, but its accuracy degrades for the second anomaly when the latter is a level anomaly. This degradation stems from a mismatch between what classification accuracy measures and what RSIC actually identifies. Classification accuracy is evaluated against the true anomaly type, whereas RSIC identifies the intervention required to best reconcile the model with the incoming data. When the first intervention over- or under-estimates the true trend, the baseline carries a trend error into the second anomaly. Even if the second anomaly is a pure level shift, the model must also correct for this inherited trend discrepancy, causing RSIC to favor a trend intervention over a level one.

For baseline estimation, RSIC achieves the lowest MSE for trend in the $\texttt{L} \rightarrow \texttt{L}$, $\texttt{L} \rightarrow \texttt{T}$, and $\texttt{T} \rightarrow \texttt{L}$ scenarios, while RSI achieves the lowest MSE for level across all scenarios. The latter can be attributed to RSI's handling of the level transition: it keeps the mean of the level unchanged but introduces additional level uncertainty, enabling the filter to search for the new level gradually. RSIC, by contrast, applies a discrete one-step shift to the level's mean at the intervention time, which, while enabling rapid adaptation, could introduce a larger instantaneous estimation error and therefore yield a higher MSE in level estimation than RSI.

5.3.2 Real Time Series

In this section, the evaluation is extended to the real time series introduced in Section 4.3.2. It first recalls the dataset and the detrended time series used for a controlled comparison. In contrast with the experiments presented in Chapter 4, this section then verifies the detection performance by adding two synthetic anomalies instead of a single one to the detrended series, demonstrating the methods' performance on subsequent anomalies in long-term SHM.

Finally, it validates the methods against manually defined labels on the original time series.

Real and Detrended Datasets. The evaluation relies on the ten real time series presented in Section 4.3.2, i.e., normalized weekly measurements of crack opening and displacement collected on dams in Canada, spanning between 9 and 18 years and shown in Figure 4.7. As described in that section, the intersections between the manually estimated baseline segments indicate potential anomalies: those highlighted in yellow have been confirmed by dam engineers, while those highlighted in grey remain of unknown nature. Because the true anomaly labels are unknown for most time series, evaluating the methods directly on the raw data introduces subjectivity into the comparison. Following the same rationale as in Section 4.3.2, the controlled probabilistic evaluation is therefore first conducted on the detrended time series shown in Figure 4.8, obtained by removing the Prophet-fitted baseline from each series while preserving all other components such as seasonality, observation errors, and short-term fluctuations. As noted in Chapter 4, relying on Prophet for detrending provides that method with a possible preprocessing advantage in the subsequent comparison.

Verification on Detrended Time Series. The seven detection methods are first evaluated on the detrended time series onto which synthetic anomalies are introduced. The data partition follows Section 4.3.2, with the first 30% of observations forming the training set, the following 10% the validation set, and the remainder the test set. In contrast with Chapter 4, where a single trend anomaly was introduced, two synthetic anomalies of types level or trend are here introduced sequentially in the test set. The first anomaly has a fixed magnitude and starts at a random time step between six months and one year after the validation set. The second anomaly has a varying magnitude and starts at the midpoint between the first anomaly and the end of the time series. For each magnitude of the second anomaly, 100 repetitions are generated by randomly drawing the starting time of the first anomaly, which in turn determines the starting time of the second anomaly.

The seven methods are evaluated on the test set according to the experimental setup specified in Appendix G. The anomaly thresholds for all methods are calibrated on the stationary detrended time series. The metrics reported here — time to detection, detection probability, false alarm rate, baseline estimation errors in standardized space, and classification accuracy — are computed following the same definitions as in Section 5.3.1, and are averaged across the ten detrended time series.

Figure 5.7 presents the time to detection Δ_t and detection probability $p^{\mathtt{det}}$ across the four anomaly configurations. Considered in isolation, methods such as LSTMED, TranAD, and

Prophet appear competitive or superior to RSIC in terms of time to detection and probability, particularly for small anomaly magnitudes. However, Table 5.2 reveals that this apparent sensitivity comes at the cost of high false alarm rates, exceeding five per ten years in all configurations. Their high detectability is therefore driven by raising excessive false alarms rather than accurately identifying true anomalies, making them unreliable for long-term monitoring.

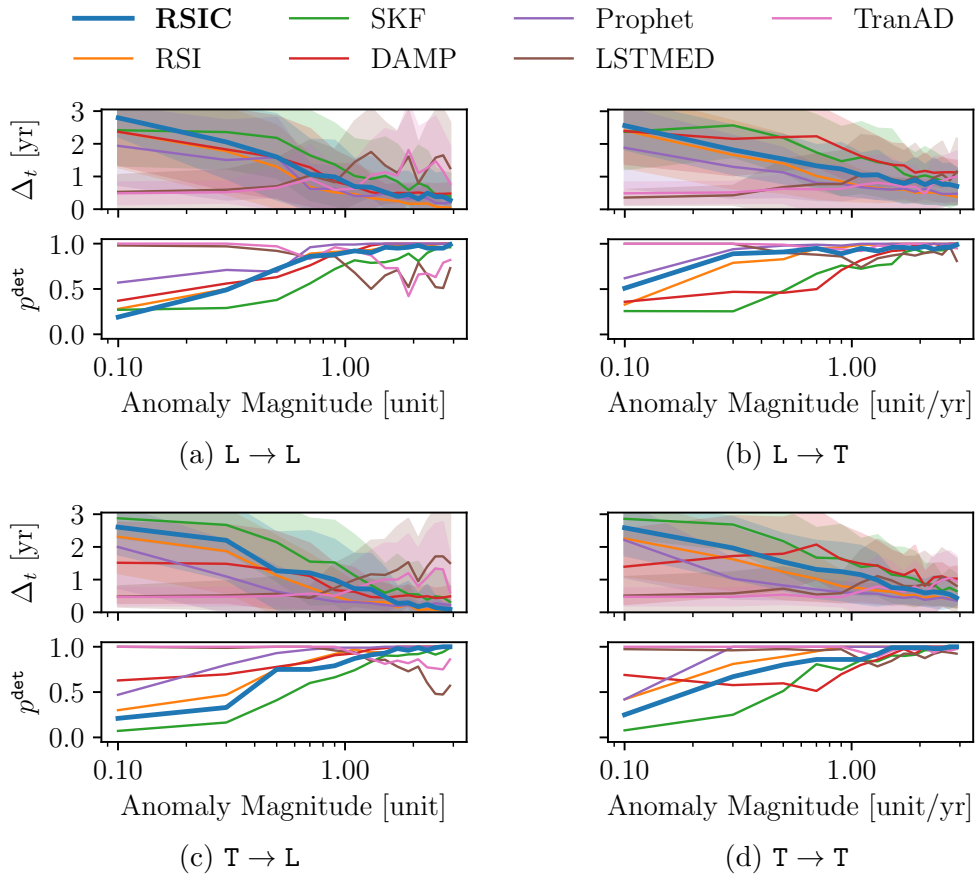


Figure 5.7 Time to detection (Δ_t) and detection probability ($p^{\mathtt{det}}$) of the tested methods on detrended time series under different anomaly combinations: level to level (L→L), level to trend (L→T), trend to level (T→L), and trend to trend (T→T).

In contrast, SKF achieves a lower false alarm rate than LSTMED, TranAD, Prophet, and RSI, but requires a longer time to detection and shows a lower detection probability across all configurations. Regarding DAMP, its false alarm rate is not directly comparable to the other methods, as it only reports the two principal discords. Even so, its detectability does not outperform the other methods; it is comparable to RSI and RSIC in the L→L and T→L configurations, but underperforms in the L→T and T→T configurations.

Table 5.2 Comparison of false alarm rate, baseline estimation errors (in standardized space) across different methods, and RSIC classification performance on ten detrended time series with different combinations of anomalies: level to level (L→L), level to trend (L→T), trend to level (T→L), and trend to trend (T→T).

Anomaly	Methods	# false alarm per 10 years	MSE_L	MSE_T	Classification accuracy	
					1st anm	2nd anm
L → L	RSIC	**1.34**	0.372	$\mathbf{6.77 \times 10^{-5}}$	0.82	0.27
	RSI	8.11	**0.202**	3.87×10^{-4}	-	-
	SKF	4.31	0.542	8.79×10^{-4}	-	-
	DAMP	0.13*	-	-	-	-
	Prophet	9.24	0.361	3.18×10^{-4}	-	-
	LSTMED	7.02	-	-	-	-
	TranAD	8.33	-	-	-	-
L → T	RSIC	**1.65**	0.421	$\mathbf{2.01 \times 10^{-4}}$	0.82	0.84
	RSI	6.75	**0.128**	2.63×10^{-4}	-	-
	SKF	3.65	0.479	6.54×10^{-4}	-	-
	DAMP	0.25*	-	-	-	-
	Prophet	9.19	0.254	3.48×10^{-4}	-	-
	LSTMED	8.66	-	-	-	-
	TranAD	7.85	-	-	-	-
T → L	RSIC	**0.96**	0.114	$\mathbf{9.36 \times 10^{-5}}$	0.99	0.07
	RSI	4.63	**0.110**	2.16×10^{-4}	-	-
	SKF	2.50	0.211	2.40×10^{-4}	-	-
	DAMP	0.13*	-	-	-	-
	Prophet	6.55	0.164	1.52×10^{-4}	-	-
	LSTMED	5.16	-	-	-	-
	TranAD	6.85	-	-	-	-
T → T	RSIC	**1.13**	0.085	1.52×10^{-4}	0.99	0.80
	RSI	3.09	**0.042**	$\mathbf{8.86 \times 10^{-5}}$	-	-
	SKF	1.73	0.145	1.75×10^{-4}	-	-
	DAMP	0.41*	-	-	-	-
	Prophet	6.11	0.056	1.29×10^{-4}	-	-
	LSTMED	5.99	-	-	-	-
	TranAD	6.42	-	-	-	-

* The false alarm rate is obtained by reporting the top-2 discords as detections, such that DAMP cannot produce more than two detections per time series.

RSI shows a moderate false alarm rate, lower than LSTMED, TranAD, and Prophet, but higher than SKF, while its detectability remains competitive across all configurations. However, its false alarm rate increases when a level anomaly is involved, as RSI cannot identify the anomaly type and systematically corrects the trend regardless, leading to more frequent false alarms. Beyond the false alarm rate, RSI achieves the lowest $\mathrm{MSE}_{\mathtt{L}}$ in all configurations, reflecting its better level adaptation.

RSIC addresses the false alarm limitation of RSI by further reducing its false alarm rate to a level below that of SKF, while preserving its detectability. Furthermore, RSIC attains the lowest $\mathrm{MSE}_{\mathtt{T}}$ in configurations involving level anomalies, as its classification-guided intervention avoids applying a trend correction when a level anomaly is identified, preventing unnecessary trend modifications that would otherwise increase the trend estimation error. RSIC thus emerges as the method that best balances detectability, baseline estimation, and false alarm rate across all four anomaly configurations.

Regarding RSIC's classification accuracy, the results for the first anomaly remain high in all configurations, reaching near-perfect values when the first anomaly is a trend one. For the second anomaly, accuracy degrades markedly when it is a level anomaly, dropping to 0.27 and 0.07 in the $\mathtt{L}\rightarrow\mathtt{L}$ and $\mathtt{T}\rightarrow\mathtt{L}$ configurations. This pattern is consistent with the behavior observed in Section 5.3.1, where imperfect trend estimation at the first intervention propagates a trend discrepancy that the second intervention must correct. On real detrended data, this effect is amplified due to a higher degree of residual complexity than in synthetic time series, which makes precise trend estimation harder and increases the likelihood that the first intervention over- or under-estimates the true trend. As a result, RSIC systematically favors a trend intervention at the second anomaly, for the reasons detailed in Section 5.3.1.

Validation on Real Time Series. The seven methods are finally compared on the ten raw time series using the manually identified anomaly labels. Following the same setup as specified in Appendix G, the training and validation sets correspond to the red and green shaded regions in Figure 4.7, each ending before the first anomaly, and the anomaly thresholds are calibrated on the combined training and validation sets. Figure 5.8 compares the alarms triggered by each method against the anomaly labels, which are confirmed by dam engineers for time series #1 (labeled in yellow) and of unknown nature for time series #2–10 (labeled in grey). The results are discussed accordingly in two parts.

Time series #1. For time series #1, RSIC and RSI are the first to detect the confirmed anomaly, followed by Prophet and TranAD in third place, LSTMED in fifth, DAMP in sixth,

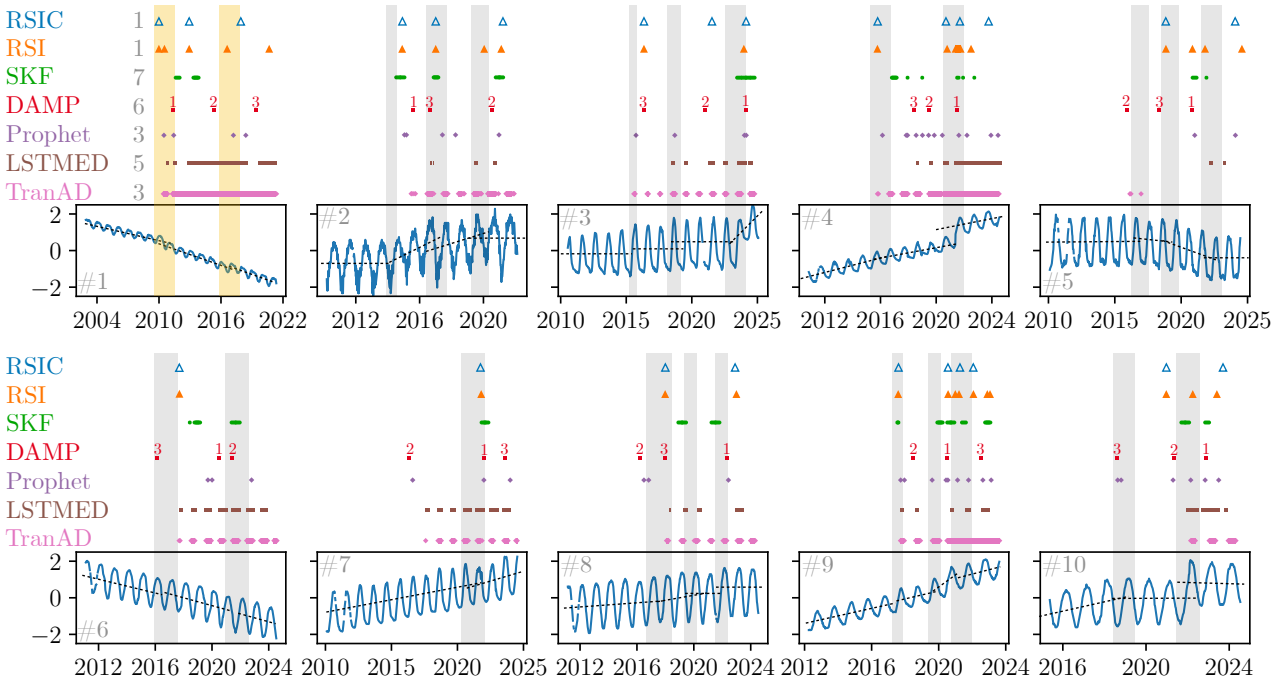


Figure 5.8 Comparison of seven methods against confirmed (yellow areas) and potential (grey areas) anomalies on real time series. Grey numbers indicate the ranking of their first valid detection on time series #1. The markers above the time series indicate the alarms triggered by different methods.

and SKF in seventh. However, ranking by time to detection alone does not fully reflect the quality of each method's output over the entire time series.

Despite sharing third place with Prophet, TranAD produces dense and continuous streams of alarms throughout the remainder of the time series after the first detection, making it uninformative for monitoring subsequent anomalies in the long term. LSTMED exhibits the same behavior, triggering continuous alarms and failing to adapt to the new regime after the first anomaly. In contrast, the remaining methods trigger sparse and clean alarms throughout the time series.

All methods except SKF indicate the second anomaly with a detection within or near the boundary of the second identified anomaly region. Among them, DAMP shows the earliest detection, followed by RSI, Prophet, and RSIC. Across the entire time series #1, RSI produces more false or repetitive alarms than RSIC and Prophet. Even though DAMP is the first method that detects the second anomaly, it operates differently from the other methods. Rather than detecting sequentially emerging regime changes, it identifies the top-3 globally most unusual subsequences in the time series, as the preliminary analysis of the real time series showed that most of them contain three anomalies. The number above each detection corresponds to its anomaly ranking. This global perspective makes its results difficult to compare directly with the sequential detection approach of the other methods.

Time series #2–10. For time series #2–10, the anomaly labels are not precisely known, and several global patterns emerge across the methods. LSTMED and TranAD consistently produce dense and continuous streams of alarms across all time series, confirming the behavior observed in the synthetic and detrended time series, and making them unsuitable for long-term monitoring. Prophet behaves inconsistently: it produces sparse alarms in some time series but switches to continuous triggering in others, such as time series #4 and #9.

SKF sometimes has a longer time to detection compared to other methods, but its alarms are well aligned with the grey anomaly regions in time series #2, #4, #6, #7, #9, and #10, confirming that it remains a strong baseline method for detecting regime changes in real time series.

As noted earlier, DAMP now identifies the top-3 most globally unusual subsequences rather than detecting sequentially emerging regime changes. As a result, while its top-ranked detections marked by 1 are close to the anomaly regions, they do not necessarily align with them, as seen in time series #2, #5, #6, and #10.

RSI and RSIC both produce sparse and clean alarms, with detections more consistently aligned with the grey anomaly regions than the other methods. RSIC tends to produce fewer

alarms than RSI across the ten time series, reflecting its lower false alarm rate observed in the synthetic and detrended time series. These observations highlight that jointly considering detectability and alarm sparsity is essential to assess method performance in a long-term monitoring context, as neither criterion alone is sufficient. It should be noted that, because the anomaly labels for time series #2–10 are of unknown nature, the comparison on the real time series remains subjective; only the verification on detrended data, where the anomaly positions are exactly known, provides an objective comparison of the methods. By the combined criterion of detectability and alarm sparsity, RSIC nevertheless shows the best balance among all methods, consistent with the results obtained on the objective comparison on detrended data.

5.4 Conclusions

This study proposed RSIC, an online anomaly detection method for state-space models that extends RSI in two directions. First, RSIC incorporates a level anomaly type in addition to the trend anomaly, enabling it to respond appropriately to both. Second, RSIC introduces a probabilistic classification procedure based on a Beta-Bernoulli formulation with fractional pseudo-count updates to infer the anomaly type and its corresponding intervention from observations accumulated after a detection.

Experiments on both synthetic and real dam monitoring time series demonstrated that RSIC matched RSI in time to detection and detection probability while reducing the false alarm rate, especially in configurations involving level anomalies. Competing methods each present limitations in long-term monitoring: 1) some methods require prior knowledge that is rarely available in practice, such as the number of anomalies in the time series, are difficult to deploy in real monitoring scenarios; 2) deep learning methods achieve high sensitivity at small anomaly magnitudes but produce excessive false alarms and fail to generalize to new regimes; 3) other methods offer lower false alarm rates at the cost of reduced detectability or inconsistent performance across anomaly configurations. Among all evaluated methods, RSIC consistently offered the best tradeoff between anomaly detectability, false alarm rate, and baseline estimation accuracy.

While RSIC attains a high classification accuracy for the first anomaly, its accuracy degrades for subsequent level anomalies due to the inherited error from the previous intervention. Future work could explore explicit uncertainty propagation between sequential interventions to account for this inherited error, which would improve classification accuracy on subsequent anomalies. Overall, RSIC extends the RSI framework with more targeted interventions and robust anomaly detection capabilities, suited for long-term SHM where multiple anomalies

of different types are likely to occur over the service life of a structure.

CHAPTER 6 Conclusion

6.1 Thesis Conclusion

Long-term monitoring of aging civil infrastructure requires anomaly detection methods to be 1) effective in detecting small baseline changes despite much larger reversible fluctuations, 2) reliable by achieving high detection probability while limiting false alarms, and 3) capable of adapting to new stationary regimes and continuously detecting subsequent anomalies. These capabilities need to be achieved while maintaining computational efficiency for scalable deployment across large monitoring networks of structures. As identified in Chapter 1, no existing method meets all of these demands at once. This thesis addressed this gap by developing online anomaly detection methods within the state-space model (SSM) framework.

The first contribution in Chapter 3 introduced the bounded autoregressive (BAR) component, which constrains the autoregressive hidden state within the bounds derived from its stationary process's standard deviation. This constraint is enforced through the mixture rectified linear unit (mReLU), a formulation that extends the ReLU activation function to a two-sided bounded probabilistic form. By enforcing the stationarity of the residuals, the BAR component redirects irreversible changes towards the baseline hidden states, where the switching Kalman filter (SKF) can detect them as regime switches. In addition, the *F1 temporal* (*F1t*) score was proposed as a time-series-wise evaluation metric that integrates the anomaly detectability, the time to detection, and the false alarm rate into a single quantity, providing evaluation criteria suited to the content of long-term SHM rather than to the overall accuracy alone. The experiments on synthetic time series, as well as on bridge and dam measurements, demonstrated that the SKF-BAR outperforms the existing SKF-AR with an improved accuracy in estimating hidden states, an earlier detection of anomalies of smaller magnitudes, and the ability to control, through a single hyperparameter, the tradeoff between the anomaly detectability and the false-alarm rate.

The second contribution in Chapter 4 presented the residual stationarity intervention (RSI) method, which addresses the SKF's slow adaptation to new regimes following a detection. RSI decomposes the residual estimates with an auxiliary drift model to isolate non-stationary drifts induced by anomalies. The probability of anomaly is quantified by combining the data likelihood with the probability that the residuals remain stationary. Once an anomaly is detected, a Bayesian neural network (BNN) predicts both the magnitude and the timing of an intervention applied retrospectively to the hidden states, compensating for the inevitable delays in detecting slowly-developing anomalies. The comparison with five existing detection

approaches on synthetic time series and on real ones recorded on dams showed that RSI achieves higher detection probabilities and shorter times to detection, and that it adapts faster to new regimes induced by trend anomalies with lower false alarm rates. Nevertheless, the intervention in RSI is designed exclusively for anomalies that alter the trend of the time series. When the structure experiences an anomaly of a different nature, such as a shift in the level alone, RSI could apply an erroneous trend correction and trigger false alarms.

The third contribution in Chapter 5 extended this framework with the residual stationarity intervention and classification (RSIC) method, motivated by a key challenge in long-term SHM: over its service life, a structure is likely to experience several anomalies that alter the baseline of the measured structural responses in different ways. A detection method must remain reliable after detecting a previous anomaly of unknown type, adapt to the new stationary regime that the anomaly induces, and continue to detect subsequent anomalies. RSIC addresses this challenge by incorporating a level anomaly type in addition to the trend one, and by quantifying the probability of each type using a new probabilistic classification mechanism based on a Beta-Bernoulli formulation with fractional pseudo-count updates. This mechanism identifies the anomaly type online from the observations accumulated after a detection and applies the corresponding targeted intervention retrospectively. To reflect the long-term SHM setting, the experiments on synthetic and real time series each included two sequential anomalies of different types, with the performance evaluated on the second anomaly, after the methods had detected the first one. Under these conditions, RSIC matches the detectability of RSI while substantially reducing its false alarm rate, especially when level anomalies are involved. RSIC offers the best balance between anomaly detectability, false alarm rate, and baseline estimation accuracy among all the evaluated methods.

Altogether, the methods developed in this thesis rest on a common principle: anomalies affect the stationarity of the residual component. The BAR component enforces this stationarity within the SKF framework, so that irreversible changes are redirected towards the baseline hidden states, where the SKF can detect them as regime switches. The RSI method follows the complementary route: it operates on SSMs alone, whose residuals drift when an anomaly occurs. RSI leverages this drift as an early indicator for detecting anomalies, and then applies corrections to the hidden states so that each component of the SSMs resumes capturing the data pattern it is designed for. The RSIC method completes the framework by identifying the nature of each detected anomaly, ensuring that the appropriate correction is applied for every anomaly occurring over the structure's service life. Each methodology was verified probabilistically on populations of synthetic time series and validated on real measurements collected on full-scale infrastructure, confirming that the proposed framework enables the early and reliable detection of anomalies at a computational cost compatible with monitoring

at scale.

6.2 Limitations

Despite the improvements demonstrated throughout this thesis, the proposed methods present limitations that should be acknowledged.

A first limitation concerns the types of anomalies considered. The interventions in RSIC are restricted to level and trend anomalies, both of which are two irreversible changes affecting the baseline of the structural responses. Although these two types cover the manifestations of structural deterioration motivating this thesis, other abnormal behaviors, such as changes in the amplitude or shape of the periodic patterns, changes in the residual variance, or sensor faults that do not translate into baseline changes, are not explicitly modeled and may be either missed or misclassified as one of the two available intervention types.

A second limitation lies in the reliance on the base SSM as a generative model. The stationary residual distribution used to quantify the probability of anomaly, as well as the training sets of the BNN intervention models, was constructed from synthetic time series generated by the base SSM. The performance of the proposed methods therefore depends on the ability of the base SSM to adequately describe the normal behavior of the structure; a misspecified model, or training and validation sets that are too short or that contain unknown anomalies, would degrade both the detection and the intervention performance.

A third limitation concerns the evaluation on real data. For most real datasets, the true nature, location, and number of anomalies are unknown, so that the validation on the raw time series relies on manually identified anomaly labels and remains partly subjective. An objective assessment of the detection performance is only feasible in a controlled setting, on detrended time series onto which anomalies of known type, magnitude, and starting time are simulated. Whether the performance measured in this controlled setting carries over to the raw measurements can therefore only be checked qualitatively, against the manually identified labels.

A fourth limitation is the inheritance of estimation errors between sequential interventions. As observed in Chapter 5, when the first intervention over- or underestimates the true trend, the resulting discrepancy is carried by the baseline into the subsequent regime, and RSIC then favors a trend intervention regardless of the true nature of the second anomaly. This behavior explains the degradation of the classification accuracy for subsequent level anomalies, and reflects the absence of an explicit mechanism propagating the uncertainty of one intervention to the next.

A fifth limitation is that both RSI and RSIC currently operate on a single local time series, which has two consequences. The first concerns explanatory variables. Neither method leverages the external factors, such as temperature and loading, that could enhance detection by explaining part of the variation observed in a target time series. When these factors are not accounted for, their effect remains mixed into the signal and can either mask a genuine anomaly or be mistaken for one. The second consequence is that the intervention is learned locally, from a single time series at a time. As a result, an intervention learned on one structure cannot be transferred to another, even when the structures share similar design, materials, and deterioration patterns.

Finally, the proposed methods involve hyperparameters whose selection affects the performance: the constraint parameter γ in the BAR component, whose optimal value was found to vary across time series and anomaly types, the anomaly probability threshold, the quantile and the maximum waiting time governing the classification in RSIC. Although default values and tuning procedures were proposed, these hyperparameters are currently calibrated per time series under the assumption that an anomaly-free stationary segment is available for training and calibration.

6.3 Future Research

The limitations identified above outline several directions for future research.

A first direction is the explicit propagation of uncertainty between sequential interventions. Retaining the posterior uncertainty associated with a past intervention, and accounting for it when classifying a subsequent anomaly, would improve the classification accuracy of RSIC for subsequent level anomalies. More generally, the classification mechanism could be extended beyond the binary level-versus-trend choice, towards a library of intervention types covering, for example, changes in periodic patterns, changes in the residual variance, or reversible sensor faults, with the BNN intervention models enriched with additional inputs to discriminate between them.

A second direction concerns the joint exploitation of multiple time series. Extending the residual stationarity framework to multivariate SSMs would allow the dependencies between sensors installed on the same structure to be leveraged, so that an anomaly captured by several sensors can be detected even earlier and distinguished from a single-sensor fault. At the network scale, the similarities between structures could be exploited to transfer intervention models learned on well-instrumented structures to those with shorter monitoring histories, reducing the reliance of the BNN training on synthetic data generated from a single base

SSM.

A third direction is the development of a benchmark for anomaly detection in long-term SHM, addressing the evaluation limitation identified above. The evaluation process developed for this thesis could serve as the template for such a public benchmark made available to the SHM community. This process combines probabilistic verification, using detrended time series onto which synthetic anomalies of known type, magnitude, and starting time are introduced; and qualitative validation, using raw time series with anomalies confirmed by dam engineers. Extending this protocol to a larger number of structures, instruments, and confirmed anomaly instances would provide a shared basis for comparing detection methods, and would reduce the subjectivity that currently limits the validation on real data.

A fourth direction is the reduction of the remaining manual effort in deploying the methods. The constraint parameter of the BAR component, the anomaly probability threshold, and the classification hyperparameters of RSIC could be estimated online from the data, for instance by treating them as additional latent variables or by optimizing them with respect to the $F1t$ score on continuously updated synthetic time series. Combining the BAR component with the RSIC framework within a single architecture, where the residuals are simultaneously constrained and monitored, is also a natural integration step that was not explored in this thesis. In the longer term, the outputs of the framework could further be linked to the decision-making processes that motivate SHM, relating the detected anomaly types and estimated baseline changes to maintenance planning and reliability assessment.

Beyond civil engineering, the principle of residual stationarity underlying the BAR, RSI, and RSIC methods is generic, and its application to other domains involving long-term monitoring of slowly evolving systems constitutes another promising avenue.

REFERENCES

[1] American Society of Civil Engineers, "2025 report card for America's infrastructure: A comprehensive assessment of America's infrastructure," American Society of Civil Engineers, Tech. Rep., 2025.

[2] J. Paparrizos, Y. Kang, P. Boniol, R. S. Tsay, T. Palpanas, and M. J. Franklin, "TSB-UAD: An end-to-end benchmark suite for univariate time-series anomaly detection," in *Proceedings of the VLDB Endowment*, 2022.

[3] M. Goswami, C. Challu, L. Callot, L. Minorics, and A. Kan, "Unsupervised model selection for time-series anomaly detection," in *Proceedings of ICLR*, 2023.

[4] K. Murphy, "Switching Kalman filters," U.C. Berkeley, Tech. Rep., 1998.

[5] S. Khazaeli, "Damage detection for structural health monitoring using reinforcement and imitation learning," Ph.D. dissertation, Polytechnique Montréal, 2022.

[6] Statistics Canada, "Canada's core public infrastructure survey: Replacement values, 2022," *The Daily*, Oct. 2024.

[7] The Associated Press, "Canada overpass collapse kills five," *The Denver Post*, Oct. 2006.

[8] Commission d'enquête sur le viaduc de la Concorde, "Report of the commission of inquiry into the collapse of a portion of the de la Concorde overpass," Gouvernement du Québec, Tech. Rep., 2007.

[9] L. H. Silva Rotta, E. Alcântara, E. Park, R. G. Negri, Y. N. Lin, N. Bernardo, T. S. G. Mendes, and C. R. Souza Filho, "The 2019 Brumadinho tailings dam collapse: Possible cause and impacts of the worst human and environmental disaster in Brazil," *International Journal of Applied Earth Observation and Geoinformation*, vol. 90, p. 102119, 2020.

[10] S. Gamse, W.-H. Zhou, F. Tan, K.-V. Yuen, and M. Oberguggenberger, "Hydrostatic-season-time model updating using Bayesian model class selection," *Reliability Engineering & System Safety*, vol. 169, pp. 40–50, 2018.

[11] P. Léger, P. Côté, and R. Tinawi, "Finite element analysis of concrete swelling due to alkali-aggregate reactions in dams," *Computers & Structures*, vol. 60, no. 4, pp. 601–611, 1996.

[12] B. T. Svendsen, G. T. Frøseth, O. Øiseth, and A. Rønnquist, "A data-based structural health monitoring approach for damage detection in steel bridges using experimental data," *Journal of Civil Structural Health Monitoring*, vol. 12, no. 1, pp. 101–115, 2022.

[13] L. H. Nguyen and J.-A. Goulet, "Anomaly detection with the switching Kalman filter for structural health monitoring," *Structural Control and Health Monitoring*, vol. 25, no. 4, p. e2136, 2018.

[14] European Committee for Standardization, "EN 1990: Eurocode – basis of structural design," CEN, Belgium, 2002.

[15] Canadian Standards Association, "CSA S6:19 – Canadian highway bridge design code," CSA Group, Canada, 2019.

[16] M. Basseville and I. Nikiforov, *Detection of Abrupt Changes: Theory and Application*. Prentice Hall Englewood Cliffs, 1993, vol. 104.

[17] J. Durbin and S. J. Koopman, *Time Series Analysis by State Space Methods*. Oxford University Press, 05 2012.

[18] J. Brownjohn, "Structural health monitoring of civil infrastructure," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 365, no. 1851, pp. 589–622, 2007.

[19] V. Gattulli and L. Chiaramonte, "Condition assessment by visual inspection for a bridge management system," *Computer-Aided Civil and Infrastructure Engineering*, vol. 20, no. 2, pp. 95–107, 2005.

[20] J. Mottershead and M. Friswell, "Model updating in structural dynamics: A survey," *Journal of Sound and Vibration*, vol. 167, no. 2, pp. 347–375, 1993.

[21] Z. He, W. Li, H. Salehi, H. Zhang, H. Zhou, and P. Jiao, "Integrated structural health monitoring in bridge engineering," *Automation in Construction*, vol. 136, p. 104168, 2022.

[22] F. Pozo, D. A. Tibaduiza, and Y. Vidal, "Sensors for structural health monitoring and condition monitoring," *Sensors*, vol. 21, no. 5, 2021.

[23] R. Ceravolo, G. Coletta, G. Miraglia, and F. Palma, "Statistical correlation between environmental time series and data from long-term monitoring of buildings," *Mechanical Systems and Signal Processing*, vol. 152, p. 107460, 2021.

[24] L. C. K. Reuben and D. Mba, "Diagnostics and prognostics using switching Kalman filters," *Structural Health Monitoring*, vol. 13, no. 3, pp. 296–306, 2014.

[25] M. Azimi, A. D. Eslamlou, and G. Pekcan, "Data-driven structural health monitoring and damage detection through deep learning: State-of-the-art review," *Sensors*, vol. 20, no. 10, 2020.

[26] H. Al-Nasser, T. Al-Zuriqat, K. Dragos, C. C. Geck, and K. Smarsly, "Identification of combined sensor faults in structural health monitoring systems," *Smart Materials and Structures*, vol. 33, no. 8, p. 085026, 2024.

[27] P. Côté, V. Roy, A. Meynadier, and B. Miquel, "Applicability of the Bayesian Dynamic Linear Model method in dam behaviour analysis and comparison with the HST method," in *Dam Safety 2022 — 39th Annual Conference*. Association of State Dam Safety Officials, Sep. 2022.

[28] S. Aminikhanghahi and D. J. Cook, "A survey of methods for time series change point detection," *Knowledge and Information Systems*, vol. 51, no. 2, pp. 339–367, 2017.

[29] S. Hu, J. Yang, J. Huang, D. Li, and C. Li, "An improved kernel entropy component analysis for damage detection under environmental and operational variations," *Sensors*, vol. 25, no. 5, 2025.

[30] H. Sohn, "Effects of environmental and operational variability on structural health monitoring," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 365, no. 1851, pp. 539–560, Feb. 2007.

[31] Q. Han, Q. Ma, J. Xu, and M. Liu, "Structural health monitoring research under varying temperature condition: A review," *Journal of Civil Structural Health Monitoring*, vol. 11, no. 1, pp. 149–173, 2021.

[32] A. Moallemi, A. Burrello, D. Brunelli, and L. Benini, "Exploring scalable, distributed real-time anomaly detection for bridge health monitoring," *IEEE Internet of Things Journal*, vol. 9, no. 18, pp. 17 660–17 674, 2022.

[33] A. Mahadevan and M. Mathioudakis, "Cost-aware retraining for machine learning," *Knowledge-Based Systems*, vol. 293, p. 111610, 2024.

[34] D. M. Frangopol, M.-J. Kallen, and J. M. v. Noortwijk, "Probabilistic models for life-cycle performance of deteriorating structures: review and future directions," *Progress in Structural Engineering and Materials*, vol. 6, no. 4, pp. 197–212, 2004.

[35] S.-S. Jin and H.-J. Jung, "Vibration-based damage detection using online learning algorithm for output-only structural health monitoring," *Structural Health Monitoring*, vol. 17, no. 4, pp. 727–746, 2018.

[36] Z. Xu and J. H. Saleh, "Machine learning for reliability engineering and safety applications: Review of current status and future opportunities," *Reliability Engineering & System Safety*, vol. 211, p. 107530, 2021.

[37] N. Aswal, S. Sen, and L. Mevel, "Switching Kalman filter for damage estimation in the presence of sensor faults," *Mechanical Systems and Signal Processing*, vol. 175, p. 109116, 2022.

[38] S. Zhong, H. Langseth, and T. D. Nielsen, "A classification-based approach to monitoring the safety of dynamic systems," *Reliability Engineering & System Safety*, vol. 121, pp. 61–71, 2014.

[39] S. Schmidl, P. Wenig, and T. Papenbrock, "Anomaly detection in time series: A comprehensive evaluation," in *Proceedings of the VLDB Endowment*, 2022.

[40] E. Keogh, J. Lin, and A. Fu, "HOT SAX: efficiently finding the most unusual time series subsequence," in *Fifth IEEE International Conference on Data Mining (ICDM)*, 2005.

[41] F. Angiulli and C. Pizzuti, "Fast outlier detection in high dimensional spaces," in *European conference on principles of data mining and knowledge discovery*. Springer, 2002, pp. 15–27.

[42] H. Sarmadi and A. Karamodin, "A novel anomaly detection method based on adaptive Mahalanobis-squared distance and one-class kNN rule for structural health monitoring under environmental effects," *Mechanical Systems and Signal Processing*, vol. 140, p. 106495, 2020.

[43] W.-J. Yan, D. Chronopoulos, K.-V. Yuen, and Y.-C. Zhu, "Structural anomaly detection based on probabilistic distance measures of transmissibility function and statistical threshold selection scheme," *Mechanical Systems and Signal Processing*, vol. 162, p. 108009, 2022.

[44] M. M. Alamdari, T. Rakotoarivelo, and N. L. D. Khoa, "A spectral-based clustering for structural health monitoring of the Sydney harbour bridge," *Mechanical Systems and Signal Processing*, vol. 87, pp. 384–400, 2017.

[45] C.-C. M. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, H. A. Dau, D. F. Silva, A. Mueen, and E. Keogh, "Matrix profile I: All pairs similarity joins for time series: A unifying view that includes motifs, discords and shapelets," in *16th IEEE International Conference on Data Mining (ICDM)*, 2016, pp. 1317–1322.

[46] R. Zarei, G. Huang, and J. Wu, "Graphts: Graph-represented time series for subsequence anomaly detection," *PLOS one*, vol. 18, no. 8, p. e0290092, 2023.

[47] Y. Lu, R. Wu, A. Mueen, M. A. Zuluaga, and E. Keogh, "Matrix profile XXIV: Scaling time series anomaly detection to trillions of datapoints and ultra-fast arriving data streams," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, 2022, pp. 1173–1182.

[48] Q. Liu, P. Boniol, T. Palpanas, and J. Paparrizos, "Time-series anomaly detection: Overview and new trends," in *Proceedings of the VLDB Endowment*, vol. 17, no. 12, 2024, pp. 4229–4232.

[49] Y. Wang and Y. Ni, "Bayesian dynamic forecasting of structural strain response using structural health monitoring data," *Structural Control and Health Monitoring*, vol. 27, no. 8, p. e2575, 2020.

[50] M. Gul and F. Necati Catbas, "Statistical pattern recognition for structural health monitoring using time series modeling: Theory and experimental verifications," *Mechanical Systems and Signal Processing*, vol. 23, no. 7, pp. 2192–2204, 2009.

[51] T. Buishand and A. Klein Tank, "Regression model for generating time series of daily precipitation amounts for climate change impact studies," *Stochastic Hydrology and Hydraulics*, vol. 10, no. 2, pp. 87–106, 1996.

[52] F. Salazar, M. Á. Toledo, J. M. González, and E. Oñate, "Early detection of anomalies in dam performance: A methodology based on boosted regression trees," *Structural Control and Health Monitoring*, vol. 24, no. 11, 2017.

[53] G. Willm and N. Beaujoint, "Les méthodes de surveillance des barrages au service de la production hydraulique d'Electricité de France-problèmes ancients et solutions nouvelles," in *Proceedings of 9th International Commission on Large Dams*, 1967.

[54] S. Gamse, M. J. Henriques, M. Oberguggenberger, and J. T. Mata, "Analysis of periodicities in long-term displacement time series in concrete dams," *Structural Control and Health Monitoring*, vol. 27, no. 3, p. e2477, 2020.

[55] S. Turrisi, A. Cigada, and E. Zappa, "A cointegration-based approach for automatic anomalies detection in large-scale structures," *Mechanical Systems and Signal Processing*, vol. 166, p. 108483, 2022.

[56] S. J. Taylor and B. Letham, "Forecasting at scale," *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018.

[57] E. Žunić, K. Korjenić, K. Hodžić, and D. Đonko, "Application of facebook's Prophet algorithm for successful sales forecasting based on real-world data," *International Journal of Computer Science and Information Technology*, vol. 12, no. 2, p. 23–36, Apr. 2020.

[58] N. Amin, Y.-W. Kim, C. Kang, and Y.-C. Byun, "Electricity load forecasting using hybrid prophet–tft model incorporating weather data," *Energy Reports*, vol. 15, p. 108846, 2026.

[59] A. Shakeel, D. Chong, and J. Wang, "Load forecasting of district heating system based on improved FB-Prophet model," *Energy*, vol. 278, p. 127637, 2023.

[60] D. Zhao, R. Zhang, H. Zhang, and S. He, "Prediction of global omicron pandemic using ARIMA, MLR, and Prophet models," *Scientific Reports*, vol. 12, no. 1, p. 18138, 2022.

[61] Y. Wang, Z. Yan, D. Wang, M. Yang, Z. Li, X. Gong, D. Wu, L. Zhai, W. Zhang, and Y. Wang, "Prediction and analysis of COVID-19 daily new cases and cumulative cases: times series forecasting and machine learning models," *BMC Infectious Diseases*, vol. 22, no. 1, p. 495, 2022.

[62] R. P. Adams and D. J. C. MacKay, "Bayesian online changepoint detection," *arXiv preprint:0710.3742*, 2007.

[63] Y. Saatçi, R. D. Turner, and C. E. Rasmussen, "Gaussian process change point models," in *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010, pp. 927–934.

[64] Z. Zamanzadeh Darban, G. I. Webb, S. Pan, C. Aggarwal, and M. Salehi, "Deep learning for time series anomaly detection: A survey," *ACM Computing Surveys*, vol. 57, no. 1, pp. 1–42, Oct. 2024.

[65] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[66] J. Miao, H. Tao, H. Xie, J. Sun, and J. Cao, "Reconstruction-based anomaly detection for multivariate time series using contrastive generative adversarial networks," *Information Processing & Management*, vol. 61, no. 1, p. 103569, 2024.

[67] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *2017 International joint conference on neural networks (IJCNN)*. IEEE, 2017, pp. 1578–1585.

[68] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection: A review," *ACM Computing Surveys*, vol. 54, no. 2, Mar. 2021.

[69] F. Luleci, F. Necati Catbas, and O. Avci, "CycleGAN for undamaged-to-damaged domain translation for structural health monitoring and damage detection," *Mechanical Systems and Signal Processing*, vol. 197, p. 110370, 2023.

[70] M. H. Soleimani-Babakamali, R. Sepasdar, K. Nasrollahzadeh, and R. Sarlo, "A system reliability approach to real-time unsupervised structural health monitoring without prior information," *Mechanical Systems and Signal Processing*, vol. 171, p. 108913, 2022.

[71] K. Yang, S. Kim, and J. B. Harley, "Unsupervised long-term damage detection in an uncontrolled environment through optimal autoencoder," *Mechanical Systems and Signal Processing*, vol. 199, p. 110473, 2023.

[72] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[73] E. Sousa Tomé, M. Pimentel, and J. Figueiras, "Online early damage detection and localisation using multivariate data analysis: Application to a cable-stayed bridge," *Structural Control and Health Monitoring*, vol. 26, no. 11, p. e2434, 2019.

[74] Z. Ma, C.-B. Yun, H.-P. Wan, Y. Shen, F. Yu, and Y. Luo, "Probabilistic principal component analysis-based anomaly detection for structures with missing data," *Structural Control and Health Monitoring*, vol. 28, no. 5, p. e2698, 2021.

[75] K. Yang, Y. Ding, H. Jiang, H. Zhao, and G. Luo, "A two-stage data cleansing method for bridge global positioning system monitoring data based on bi-direction long and short term memory anomaly identification and conditional generative adversarial networks data repair," *Structural Control and Health Monitoring*, vol. 29, no. 9, p. e2993, 2022.

[76] W. Deng, Y. Li, K. Huang, D. Wu, C. Yang, and W. Gui, "LSTMED: An uneven dynamic process monitoring method based on LSTM and autoencoder neural network," *Neural Networks*, vol. 158, pp. 30–41, 2023.

[77] S. Tuli, G. Casale, and N. R. Jennings, "TranAD: deep transformer networks for anomaly detection in multivariate time series data," in *Proceedings of the VLDB Endowment*, vol. 15, no. 6, Feb. 2022, pp. 1201–1214.

[78] Z. Tang, Z. Chen, Y. Bao, and H. Li, "Convolutional neural network-based data anomaly detection method using multiple information for structural health monitoring," *Structural Control and Health Monitoring*, vol. 26, no. 1, p. e2296, 2019.

[79] Y. Zhang, X. Wang, Z. Ding, Y. Du, and Y. Xia, "Anomaly detection of sensor faults and extreme events based on support vector data description," *Structural Control and Health Monitoring*, vol. 29, no. 10, p. e3047, 2022.

[80] X. Jian, H. Zhong, Y. Xia, and L. Sun, "Faulty data detection and classification for bridge structural health monitoring via statistical and deep-learning approach," *Structural Control and Health Monitoring*, vol. 28, no. 11, p. e2824, 2021.

[81] J. Audibert, P. Michiardi, F. Guyard, S. Marti, and M. A. Zuluaga, "Do deep neural networks contribute to multivariate time series anomaly detection?" *Pattern Recognition*, vol. 132, p. 108945, 2022.

[82] X. Zhang, P. Cui, R. Xu, L. Zhou, Y. He, and Z. Shen, "Deep stable learning for out-of-distribution generalization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5372–5382.

[83] K. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[84] J.-A. Goulet, *Probabilistic Machine Learning for Civil Engineers*. MIT Press, 2020.

[85] Q.-A. Wang, C.-B. Wang, Z.-G. Ma, W. Chen, Y.-Q. Ni, C.-F. Wang, B.-G. Yan, and P.-X. Guan, "Bayesian dynamic linear model framework for structural health monitoring data forecasting and missing data imputation during typhoon events," *Structural Health Monitoring*, vol. 21, no. 6, pp. 2933–2950, 2022.

[86] L. A. Overbey, C. C. Olson, and M. D. Todd, "A parametric investigation of state-space-based prediction error methods with stochastic excitation for structural health monitoring," *Smart Materials and Structures*, vol. 16, no. 5, p. 1621, Aug. 2007.

[87] J.-A. Goulet, "Bayesian dynamic linear models for structural health monitoring," *Structural Control and Health Monitoring*, vol. 24, no. 12, 2017.

[88] H. E. Rauch, C. T. Striebel, and F. Tung, "Maximum likelihood estimates of linear dynamic systems," *AIAA journal*, vol. 3, no. 8, pp. 1445–1450, 1965.

[89] S. Khazaeli, L. H. Nguyen, and J.-A. Goulet, "Anomaly detection using state-space models and reinforcement learning," *Structural Control and Health Monitoring*, vol. 28, no. 6, p. e2720, 2021.

[90] S. Khazaeli and J.-A. Goulet, "Damage detection for structural health monitoring using reinforcement and imitation learning," *Structure and Infrastructure Engineering*, pp. 1–18, 2024.

[91] L. H. Nguyen, I. Gaudot, S. Khazaeli, and J.-A. Goulet, "A kernel-based method for modeling non-harmonic periodic phenomena in Bayesian dynamic linear models," *Frontiers in Built Environment*, vol. 5, p. 8, 2019.

[92] V. D. Vuong, L. H. Nguyen, and J.-A. Goulet, "Coupling LSTM neural networks and state-space models through analytically tractable inference," *International Journal of Forecasting*, vol. 41, no. 1, pp. 128–140, 2025.

[93] B. Deka and J.-A. Goulet, "Approximate Gaussian variance inference for state-space models," *International Journal of Adaptive Control and Signal Processing*, vol. 37, no. 11, pp. 2934–2962, 2023.

[94] L. H. Nguyen and J.-A. Goulet, "Real-time anomaly detection with Bayesian dynamic linear models," *Structural Control and Health Monitoring*, vol. 26, no. 9, p. e2404, 2019.

[95] B. Deka, L. H. Nguyen, S. Amiri, and J.-A. Goulet, "The Gaussian multiplicative approximation for state-space models," *Structural Control and Health Monitoring*, vol. 29, no. 3, p. e2904, 2022.

[96] M. West and J. Harrison, *Bayesian forecasting and dynamic models*. Springer, 1997.

[97] G. E. P. Box and G. C. Tiao, "Intervention analysis with applications to economic and environmental problems," *Journal of the American Statistical Association*, vol. 70, no. 349, pp. 70–79, 1975.

[98] J. E. Jarrett and E. Kyper, "ARIMA modeling with intervention to forecast and analyze chinese stock prices," *International Journal of Engineering Business Management*, vol. 3, p. 17, 2011.

[99] L. Li, M. S. Cuerden, B. Liu, S. Shariff, A. K. Jain, and M. Mazumdar, "Three statistical approaches for assessment of intervention effects: A primer for practitioners," *Risk Management and Healthcare Policy*, vol. 14, pp. 757–770, 2021.

[100] A. Kahn-Lang and K. Lang, "The promise and pitfalls of differences-in-differences: Reflections on 16 and pregnant and other applications," *Journal of Business & Economic Statistics*, vol. 38, no. 3, pp. 613–620, 2020.

[101] A. Wagner, S. Soumerai, F. Zhang, and D. Ross-Degnan, "Segmented regression analysis of interrupted time series studies in medication use research," *Journal of Clinical Pharmacy and Therapeutics*, vol. 27, pp. 299 – 309, Aug. 2002.

[102] P. Toman, N. Ravishanker, N. Lally, and S. Rajasekaran, "Forecasting robust gaussian process state space models for assessing intervention impact in internet of things time series," *Forecasting*, vol. 7, no. 2, 2025.

[103] D. K. Blough, "Intervention analysis in multivariate time series via the Kalman filter," in *Estimation and Analysis of Insect Populations: Proceedings of a Conference held in Laramie, Wyoming, January 25–29, 1988*. Springer, 1989, pp. 389–403.

[104] M. Song, R. Astroza, H. Ebrahimian, B. Moaveni, and C. Papadimitriou, "Adaptive Kalman filters for nonlinear finite element model updating," *Mechanical Systems and Signal Processing*, vol. 143, p. 106837, 2020.

[105] O. Khan, M. El Mistiri, D. E. Rivera, C. A. Martin, and E. Hekler, "A Kalman filter-based hybrid model predictive control algorithm for mixed logical dynamical systems: Application to optimized interventions for physical activity," in *2022 IEEE 61st Conference on Decision and Control (CDC)*, 2022, pp. 2586–2593.

[106] R. G. Krishnan, U. Shalit, and D. Sontag, "Deep Kalman filters," *arXiv preprint:1511.05121*, 2015.

[107] J. Liu, D. J. Spakowicz, G. I. Ash, R. Hoyd, R. Ahluwalia, A. Zhang, S. Lou, D. Lee, J. Zhang, C. Presley *et al.*, "Bayesian structural time series for biomedical sensor data: A flexible modeling framework for evaluating interventions," *PLOS computational biology*, vol. 17, no. 8, p. e1009303, 2021.

[108] J. Li, P. Fearnhead, P. Fryzlewicz, and T. Wang, "Automatic change-point detection in time series via deep learning," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 86, no. 2, pp. 273–285, Jan. 2024.

[109] H. Suresh, N. Hunt, A. Johnson, L. A. Celi, P. Szolovits, and M. Ghassemi, "Clinical intervention prediction and understanding with deep neural networks," in *Proceedings of the 2nd Machine Learning for Healthcare Conference*, F. Doshi-Velez, J. Fackler, D. Kale, R. Ranganath, B. Wallace, and J. Wiens, Eds., vol. 68, 2017, pp. 322–337.

[110] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher *et al.*, "A survey of uncertainty in deep neural networks," *Artificial intelligence review*, vol. 56, no. Suppl 1, pp. 1513–1589, 2023.

[111] M. Alfaro-Ponce, "Dynamic neural networks in biomedical applications: A review from embedded systems to ai-driven interventions," *Neurocomputing*, vol. 673, p. 132825, 2026.

[112] H. Zhang, M. Gül, and B. Kostić, "Eliminating temperature effects in damage detection for civil infrastructure using time series analysis and autoassociative neural networks," *Journal of Aerospace Engineering*, vol. 32, no. 2, p. 04019001, 2019.

[113] M. Forgione and D. Piga, "Neural state-space models: Empirical evaluation of uncertainty quantification," *IFAC-PapersOnLine*, vol. 56, no. 2, pp. 4082–4087, 2023, 22nd IFAC World Congress.

[114] E. Goan and C. Fookes, "Bayesian neural networks: An introduction and survey," in *Case Studies in Applied Bayesian Data Science*. Springer International Publishing, 2020, pp. 45–87.

[115] J.-A. Goulet, L. H. Nguyen, and S. Amiri, "Tractable approximate Gaussian inference for Bayesian neural networks," *Journal of Machine Learning Research*, vol. 22, no. 251, pp. 1–23, 2021.

[116] B. Deka, L. H. Nguyen, and J.-A. Goulet, "Analytically tractable heteroscedastic uncertainty quantification in Bayesian neural networks for regression tasks," *Neurocomputing*, vol. 572, p. 127183, 2024.

[117] S. A. K. Fakhri, Z. Hamida, and J.-A. Goulet, "Scalable probabilistic deterioration model based on visual inspections and structural attributes from large networks of bridges," *Advanced Engineering Informatics*, vol. 64, p. 103035, 2025.

[118] L. H. Nguyen and J.-A. Goulet, "Analytically tractable hidden-states inference in Bayesian neural networks," *Journal of Machine Learning Research*, vol. 23, pp. 1–33, 2022.

[119] Q. Liu and J. Paparrizos, "The elephant in the room: Towards a reliable time-series anomaly detection benchmark," in *Conference on Neural Information Processing Systems*, vol. 37, 2024, pp. 108 231–108 261.

[120] C. J. Van Rijsbergen, *Information Retrieval*, 2nd ed. Butterworth-Heinemann, 1979.

[121] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 2006, p. 233–240.

[122] N. Tatbul, T. J. Lee, S. Zdonik, M. Alam, and J. Gottschlich, "Precision and recall for time series," *Conference on Neural Information Processing Systems*, vol. 31, 2018.

[123] A. Huet, J. M. Navarro, and D. Rossi, "Local evaluation of time series anomaly detection algorithms," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 635–645.

[124] J. Paparrizos, P. Boniol, T. Palpanas, R. S. Tsay, A. Elmore, and M. J. Franklin, "Volume under the surface: a new accuracy evaluation measure for time-series anomaly detection," in *Proceedings of the VLDB Endowment*, vol. 15, no. 11, Jul. 2022, p. 2774–2787.

[125] A. Tartakovsky, I. Nikiforov, and M. Basseville, *Sequential Analysis: Hypothesis Testing and Changepoint Detection*, 1st ed. Chapman & Hall/CRC, 2014.

[126] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Conference on Neural Information Processing Systems*, 2017.

[127] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint:1411.1784*, 2014.

[128] A. Krizhevsky and G. Hinton, "Convolutional deep belief networks on CIFAR-10," *unpublished*, 2010.

[129] A. Runnalls, "Kullback-Leibler approach to Gaussian mixture reduction," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 43, no. 3, pp. 989–999, 2007.

[130] D. Simon and D. L. Simon, "Constrained Kalman filtering via density function truncation for turbofan engine health estimation," *International Journal of Systems Science*, vol. 41, no. 2, pp. 159–171, 2010.

[131] A. Bhatnagar, P. Kassianik, C. Liu, T. Lan, W. Yang, R. Cassius, D. Sahoo, D. Arpit, S. Subramanian, G. Woo, A. Saha, A. K. Jagota, G. Gopalakrishnan, M. Singh, K. C. Krithika, S. Maddineni, D. Cho, B. Zong, Y. Zhou, C. Xiong, S. Savarese, S. Hoi, and H. Wang, "Merlion: End-to-end machine learning for time series," *Journal of Machine Learning Research*, vol. 24, no. 226, pp. 1–6, 2023.

[132] H. Raiffa and R. Schlaifer, *Applied Statistical Decision Theory*. John Wiley & Sons, 2000.

[133] Z. Wu, E. Liang, M. Luo, S. Mika, J. E. Gonzalez, and I. Stoica, "RLlib flow: Distributed reinforcement learning is a dataflow problem," in *Conference on Neural Information Processing Systems*, 2021.

APPENDIX A SSM MATRICES AND PARAMETERS FOR BAR

This appendix provides all the model matrices and their parameters used in the three experiments shown in Section 3.4.

A.1 Synthetic Data

This appendix provides a summary of the model matrices and parameters used in the synthetic time series generation and the SKF analysis.

A.1.1 Synthetic Data Generation

In the synthetic experiment, the data generation employs `LL` and `AR` components. After the synthetic time series are generated, a trend anomaly is applied on top of each of them. Model matrices to generate synthetic data (GS) are

$$\begin{aligned}
\mathbf{A}^{\mathrm{GS}} &= \text{blockdiag}\left(\mathbf{A}^{\mathtt{LL}}, \mathbf{A}^{\mathtt{AR}}\right) \\
\mathbf{F}^{\mathrm{GS}} &= \left[\mathbf{F}^{\mathtt{LL}}, \mathbf{F}^{\mathtt{AR}}\right] \\
\mathbf{Q}^{\mathrm{GS}} &= \text{blockdiag}\left(\mathbf{Q}^{\mathtt{LL}}, \mathbf{Q}^{\mathtt{AR}}\right).
\end{aligned}$$

Parameter values for synthetic data generation are

$$\mathcal{P}^{\mathrm{GS}} = \{\phi^{\mathtt{AR}}, \sigma^{\mathtt{AR}}, \sigma_V\} = \{0.9, 0.2, 0.001\}.$$

A.1.2 Stationary Model Matrices in SKF

The stationary model matrices for the synthetic time series, referred to as SS, are

$$\begin{aligned}
\mathbf{A}^{\mathrm{SS}} &= \text{blockdiag}\left(\mathbf{A}^{\mathtt{LcT}}, \mathbf{A}^{\mathtt{AR/BAR}}\right) \\
\mathbf{F}^{\mathrm{SS}} &= \left[\mathbf{F}^{\mathtt{LcT}}, \mathbf{F}^{\mathtt{AR/BAR}}\right] \\
\mathbf{Q}^{\mathrm{SS},\mathtt{LL(LL)}} &= \text{blockdiag}\left(\mathbf{Q}^{\mathtt{LL(LL)}}, \mathbf{Q}^{\mathtt{AR/BAR}}\right) \\
\mathbf{Q}^{\mathrm{SS},\mathtt{LT(LL)}} &= \text{blockdiag}\left(\mathbf{Q}^{\mathtt{LT(LL)}}, \mathbf{Q}^{\mathtt{AR/BAR}}\right),
\end{aligned}$$

where `LcT` denotes the local level component compatible with trend.

A.1.3 Non-Stationary Model Matrices in SKF

The non-stationary model matrices for the synthetic time series, referred to as NS, are

$$\begin{aligned}
\mathbf{A}^{\mathrm{NS}} &= \text{blockdiag}\left(\mathbf{A}^{\mathtt{LT}}, \mathbf{A}^{\mathtt{AR/BAR}}\right) \\
\mathbf{F}^{\mathrm{NS}} &= \left[\mathbf{F}^{\mathtt{LT}}, \mathbf{F}^{\mathtt{AR/BAR}}\right] \\
\mathbf{Q}^{\mathrm{NS},\mathtt{LT(LT)}} &= \text{blockdiag}\left(\mathbf{Q}^{\mathtt{LT(LT)}}, \mathbf{Q}^{\mathtt{AR/BAR}}\right) \\
\mathbf{Q}^{\mathrm{NS},\mathtt{LL(LT)}} &= \text{blockdiag}\left(\mathbf{Q}^{\mathtt{LL(LT)}}, \mathbf{Q}^{\mathtt{AR/BAR}}\right).
\end{aligned}$$

Parameter values for both regimes in SKF on synthetic time series are

$$\mathcal{P}^{\mathrm{S}} = \{\phi^{\mathtt{AR}}, \sigma^{\mathtt{AR}}, \sigma_V, z^{11}, z^{22}\} = \{0.9, 0.2, 0.001, 0.999999, 0.999999\}.$$

A.2 Bridge Data

This appendix presents the model matrices and their parameters used in the SKF analysis for the bridge dataset including measurements of air temperature (T) and elongation (E).

A.2.1 Stationary Model Matrices in SKF

The stationary model matrices for the bridge time series, referred to as SB, are

$$\begin{aligned}
\mathbf{A}^{\mathrm{SB}} &= \text{blockdiag}\left(\overbrace{\mathbf{A}^{\mathtt{LL,T}}, \mathbf{A}^{\mathtt{PD,T}}\mathbf{A}^{\mathtt{AR,T}}}^{\text{Temperature}}, \overbrace{\mathbf{A}^{\mathtt{LcT,E}}, \mathbf{A}^{\mathtt{AR/BAR,E}}}^{\text{Elongation}}\right) \\
\mathbf{Q}^{\mathrm{SB},\mathtt{LL(LL)}} &= \text{blockdiag}\left(\overbrace{\mathbf{Q}^{\mathtt{LL,T}}, \mathbf{Q}^{\mathtt{PD,T}}\mathbf{Q}^{\mathtt{AR,T}}}^{\text{Temperature}}, \overbrace{\mathbf{Q}^{\mathtt{LL(LL),E}}, \mathbf{Q}^{\mathtt{AR/BAR,E}}}^{\text{Elongation}}\right) \\
\mathbf{Q}^{\mathrm{SB},\mathtt{LT(LL)}} &= \text{blockdiag}\left(\overbrace{\mathbf{Q}^{\mathtt{LL,T}}, \mathbf{Q}^{\mathtt{PD,T}}\mathbf{Q}^{\mathtt{AR,T}}}^{\text{Temperature}}, \overbrace{\mathbf{Q}^{\mathtt{LT(LL),E}}, \mathbf{Q}^{\mathtt{AR/BAR,E}}}^{\text{Elongation}}\right).
\end{aligned}$$

The observation matrix $\mathbf{F}$ models the dependency between temperature and elongation. For the standard `AR` model, the observation matrix is

$$\mathbf{F}^{\mathrm{SB},\mathtt{AR}} = \begin{bmatrix}
\overbrace{1}^{\mathtt{LL,T}} & \overbrace{1 \quad 0}^{\mathtt{PD,T}} & \overbrace{1}^{\mathtt{AR,T}} & \overbrace{0 \quad 0}^{\mathtt{LcT,E}} & \overbrace{0}^{\mathtt{AR,E}} \\
0 & \beta^{\mathtt{PD,E|T}} \quad 0 & \beta^{\mathtt{AR,E|T}} & 1 \quad 0 & 1
\end{bmatrix}.$$

For BAR, the observation matrix is

$$\mathbf{F}^{\mathrm{SB,BAR}} = \begin{bmatrix} \overbrace{1}^{\mathrm{LL,T}} & \overbrace{1 \quad 0}^{\mathrm{PD,T}} & \overbrace{1}^{\mathrm{AR,T}} & \overbrace{0 \quad 0}^{\mathrm{LcT,E}} & \overbrace{0 \quad 0}^{\mathrm{BAR,E}} \\ 0 & \beta^{\mathrm{PD,E|T}} \quad 0 & \beta^{\mathrm{AR,E|T}} & 1 \quad 0 & 0 \quad 1 \end{bmatrix},$$

where the first row corresponds to the temperature observation and the second corresponds to the elongation. β is the dependency ratio between the elongation hidden states and the temperature hidden states.

A.2.2 Non-Stationary Model Matrices in SKF

The non-stationary model matrices for the bridge time series, referred to as NB, are

$$\begin{aligned} \mathbf{A}^{\mathrm{NB}} &= \text{blockdiag}\left(\overbrace{\mathbf{A}^{\mathrm{LL,T}}, \mathbf{A}^{\mathrm{PD,T}}\mathbf{A}^{\mathrm{AR,T}}}^{\text{Temperature}}, \overbrace{\mathbf{A}^{\mathrm{LT,E}}, \mathbf{A}^{\mathrm{AR/BAR,E}}}^{\text{Elongation}} \right) \\ \mathbf{Q}^{\mathrm{NB,LT(LT)}} &= \text{blockdiag}\left(\overbrace{\mathbf{Q}^{\mathrm{LL,T}}, \mathbf{Q}^{\mathrm{PD,T}}\mathbf{Q}^{\mathrm{AR,T}}}^{\text{Temperature}}, \overbrace{\mathbf{Q}^{\mathrm{LT(LT),E}}, \mathbf{Q}^{\mathrm{AR/BAR,E}}}^{\text{Elongation}} \right) \\ \mathbf{Q}^{\mathrm{NB,LL(LT)}} &= \text{blockdiag}\left(\overbrace{\mathbf{Q}^{\mathrm{LL,T}}, \mathbf{Q}^{\mathrm{PD,T}}\mathbf{Q}^{\mathrm{AR,T}}}^{\text{Temperature}}, \overbrace{\mathbf{Q}^{\mathrm{LL(LT),E}}, \mathbf{Q}^{\mathrm{AR/BAR,E}}}^{\text{Elongation}} \right). \end{aligned}$$

The observation matrix $\mathbf{F}^{\mathrm{NB}}$ is identical as $\mathbf{F}^{\mathrm{SB}}$ for both methods.
Parameter values for both regimes in SKF on bridge time series are

$$\begin{aligned} \mathcal{P}^{\mathrm{B}} &= \left\{ \overbrace{\phi^{\mathrm{AR,T}}, \sigma^{\mathrm{AR,T}}, \sigma_V^{\mathrm{T}}}^{\text{Temperature}}, \overbrace{\phi^{\mathrm{AR,E}}, \sigma^{\mathrm{AR,E}}, \sigma_V^{\mathrm{E}}, \beta^{\mathrm{PD,E|T}}, \beta^{\mathrm{AR,E|T}}, z^{11}, z^{22}}^{\text{Elongation}} \right\} \\ &= \{0.36981, 3.3065, 0.1, 0.74033, 0.0040317, 0.001, 0.019987, \\ &\quad 0.019689, 0.999999999, 0.999999999\}. \end{aligned}$$

A.3 Dam Data

This appendix presents the model matrices and their parameters used in the SKF analysis for the dam dataset consisting of eight time series.

A.3.1 Stationary Model Matrices in SKF

For automatically collected dam time series, the stationary model matrices, referred to as SDA, are

$$\begin{aligned}
\mathbf{A}^{\mathrm{SDA}} &= \text{blockdiag}\left(\mathbf{A}^{\mathtt{TcA}}, \mathbf{A}^{\mathtt{KR}}, \mathbf{A}^{\mathtt{AR/BAR}}\right) \\
\mathbf{F}^{\mathrm{SDA}} &= \left[\mathbf{F}^{\mathtt{TcA}}, \mathbf{F}^{\mathtt{KR}}, \mathbf{F}^{\mathtt{AR/BAR}}\right] \\
\mathbf{Q}^{\mathrm{SDA},\mathtt{LT(LT)}} &= \text{blockdiag}\left(\mathbf{Q}^{\mathtt{LT(LT)}}, \mathbf{Q}^{\mathtt{KR}}, \mathbf{Q}^{\mathtt{AR/BAR}}\right) \\
\mathbf{Q}^{\mathrm{SDA},\mathtt{LA(LT)}} &= \text{blockdiag}\left(\mathbf{Q}^{\mathtt{LA(LT)}}, \mathbf{Q}^{\mathtt{KR}}, \mathbf{Q}^{\mathtt{AR/BAR}}\right).
\end{aligned}$$

For manually collected dam time series, the stationary model matrices, referred to as SDM, are

$$\begin{aligned}
\mathbf{A}^{\mathrm{SDM}} &= \text{blockdiag}\left(\mathbf{A}^{\mathtt{TcA}}, \mathbf{A}^{\mathtt{PD}}, \mathbf{A}^{\mathtt{AR/BAR}}\right) \\
\mathbf{F}^{\mathrm{SDM}} &= \left[\mathbf{F}^{\mathtt{TcA}}, \mathbf{F}^{\mathtt{PD}}, \mathbf{F}^{\mathtt{AR/BAR}}\right] \\
\mathbf{Q}^{\mathrm{SDM},\mathtt{LT(LT)}} &= \text{blockdiag}\left(\mathbf{Q}^{\mathtt{LT(LT)}}, \mathbf{Q}^{\mathtt{PD}}, \mathbf{Q}^{\mathtt{AR/BAR}}\right) \\
\mathbf{Q}^{\mathrm{SDM},\mathtt{LA(LT)}} &= \text{blockdiag}\left(\mathbf{Q}^{\mathtt{LA(LT)}}, \mathbf{Q}^{\mathtt{PD}}, \mathbf{Q}^{\mathtt{AR/BAR}}\right).
\end{aligned}$$

A.3.2 Non-Stationary Model Matrices in SKF

For automatically collected dam time series, the non-stationary model matrices, referred to as NDA, are

$$\begin{aligned}
\mathbf{A}^{\mathrm{NDA}} &= \text{blockdiag}\left(\mathbf{A}^{\mathtt{LA}}, \mathbf{A}^{\mathtt{KR}}, \mathbf{A}^{\mathtt{AR/BAR}}\right) \\
\mathbf{F}^{\mathrm{NDA}} &= \left[\mathbf{F}^{\mathtt{LA}}, \mathbf{F}^{\mathtt{KR}}, \mathbf{F}^{\mathtt{AR/BAR}}\right] \\
\mathbf{Q}^{\mathrm{NDA},\mathtt{LA(LA)}} &= \text{blockdiag}\left(\mathbf{Q}^{\mathtt{LA(LA)}}, \mathbf{Q}^{\mathtt{KR}}, \mathbf{Q}^{\mathtt{AR/BAR}}\right) \\
\mathbf{Q}^{\mathrm{NDA},\mathtt{LT(LA)}} &= \text{blockdiag}\left(\mathbf{Q}^{\mathtt{LT(LA)}}, \mathbf{Q}^{\mathtt{KR}}, \mathbf{Q}^{\mathtt{AR/BAR}}\right).
\end{aligned}$$

For manually collected dam time series, the non-stationary model matrices (NDM) are

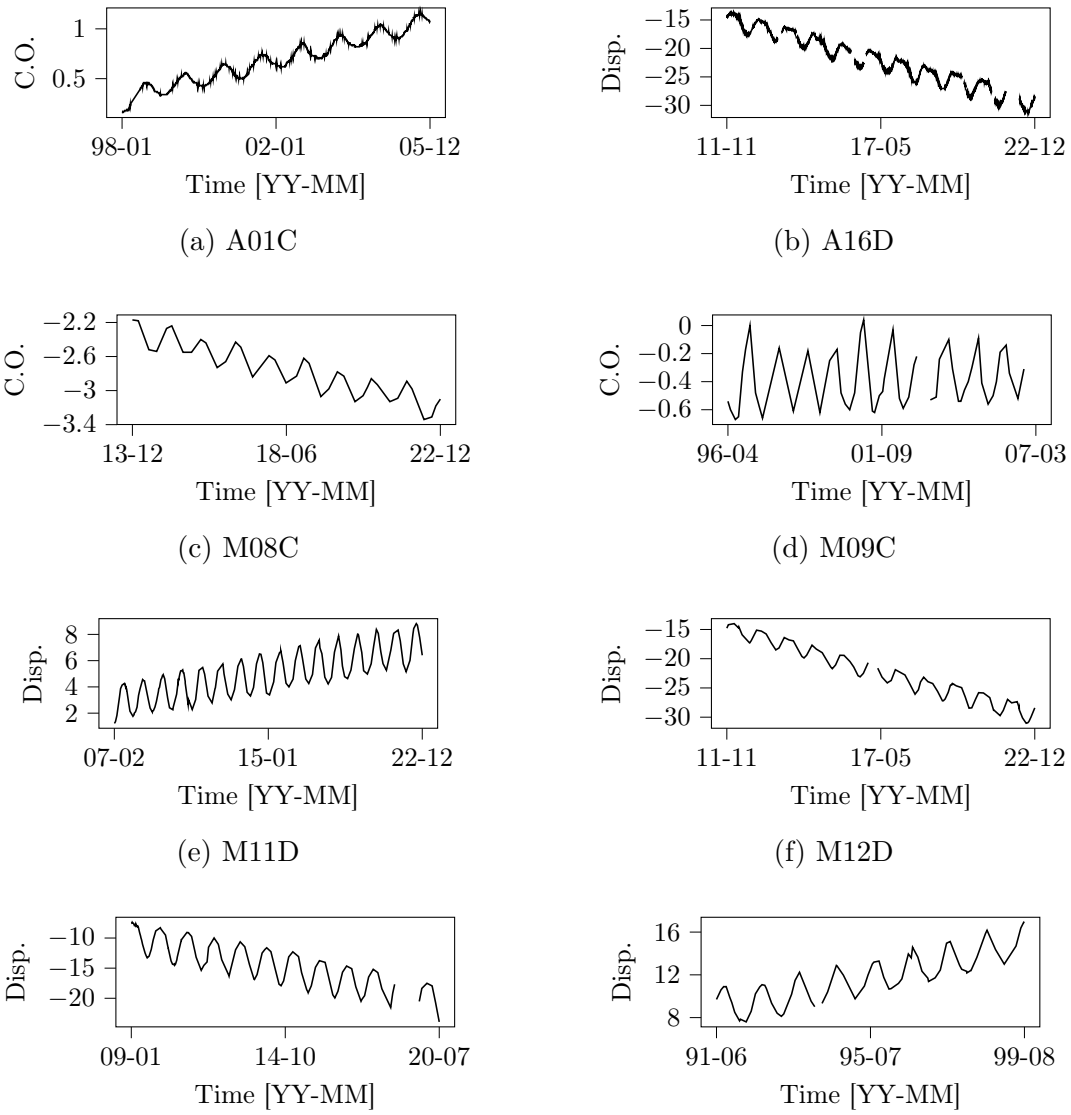
$$\begin{aligned}
\mathbf{A}^{\mathrm{NDM}} &= \text{blockdiag}\left(\mathbf{A}^{\mathtt{LA}}, \mathbf{A}^{\mathtt{PD}}, \mathbf{A}^{\mathtt{AR/BAR}}\right) \\
\mathbf{F}^{\mathrm{NDM}} &= \left[\mathbf{F}^{\mathtt{LA}}, \mathbf{F}^{\mathtt{PD}}, \mathbf{F}^{\mathtt{AR/BAR}}\right] \\
\mathbf{Q}^{\mathrm{NDM},\mathtt{LA(LA)}} &= \text{blockdiag}\left(\mathbf{Q}^{\mathtt{LA(LA)}}, \mathbf{Q}^{\mathtt{PD}}, \mathbf{Q}^{\mathtt{AR/BAR}}\right) \\
\mathbf{Q}^{\mathrm{NDM},\mathtt{LT(LA)}} &= \text{blockdiag}\left(\mathbf{Q}^{\mathtt{LT(LA)}}, \mathbf{Q}^{\mathtt{PD}}, \mathbf{Q}^{\mathtt{AR/BAR}}\right).
\end{aligned}$$

Parameter values for both regimes in SKF on dam time series series are

$$\mathcal{P}^{\text{A01C}} = \{\ell^{\texttt{KR}}, \phi^{\texttt{AR}}, \sigma^{\texttt{AR}}, \sigma_V, z^{11}, z^{22}\} = \{0.98461, 0.91225, 0.0060635, 0.001, 0.99999, 0.99999\}$$
$$\mathcal{P}^{\text{A16D}} = \{\ell^{\texttt{KR}}, \phi^{\texttt{AR}}, \sigma^{\texttt{AR}}, \sigma_V, z^{11}, z^{22}\} = \{0.85519, 0.99412, 0.033654, 0.001, 0.999999, 0.999999\}$$
$$\mathcal{P}^{\text{M08C}} = \{\phi^{\texttt{AR}}, \sigma^{\texttt{AR}}, \sigma_V, z^{11}, z^{22}\} = \{0.34319, 0.036042, 0.001, 0.99999, 0.99999\}$$
$$\mathcal{P}^{\text{M09C}} = \{\phi^{\texttt{AR}}, \sigma^{\texttt{AR}}, \sigma_V, z^{11}, z^{22}\} = \{0.088333, 0.091778, 0.001, 0.99999, 0.99999\}$$
$$\mathcal{P}^{\text{M11D}} = \{\phi^{\texttt{AR}}, \sigma^{\texttt{AR}}, \sigma_V, z^{11}, z^{22}\} = \{0.39252, 0.23484, 0.001, 0.999999, 0.999999\}$$
$$\mathcal{P}^{\text{M12D}} = \{\phi^{\texttt{AR}}, \sigma^{\texttt{AR}}, \sigma_V, z^{11}, z^{22}\} = \{0.084549, 0.4741, 0.001, 0.999999, 0.999999\}$$
$$\mathcal{P}^{\text{M13D}} = \{\phi^{\texttt{AR}}, \sigma^{\texttt{AR}}, \sigma_V, z^{11}, z^{22}\} = \{0.32141, 0.39865, 0.001, 0.99999, 0.99999\}$$
$$\mathcal{P}^{\text{M14D}} = \{\phi^{\texttt{AR}}, \sigma^{\texttt{AR}}, \sigma_V, z^{11}, z^{22}\} = \{0.16912, 0.34072, 0.001, 0.99999, 0.99999\}.$$

APPENDIX B OVERVIEW OF THE DAM DATASETS FOR EVALUATING BAR PERFORMANCE

This appendix provides the plots of the eight dam datasets used in the experiments.



(a) A01C (b) A16D

(c) M08C (d) M09C

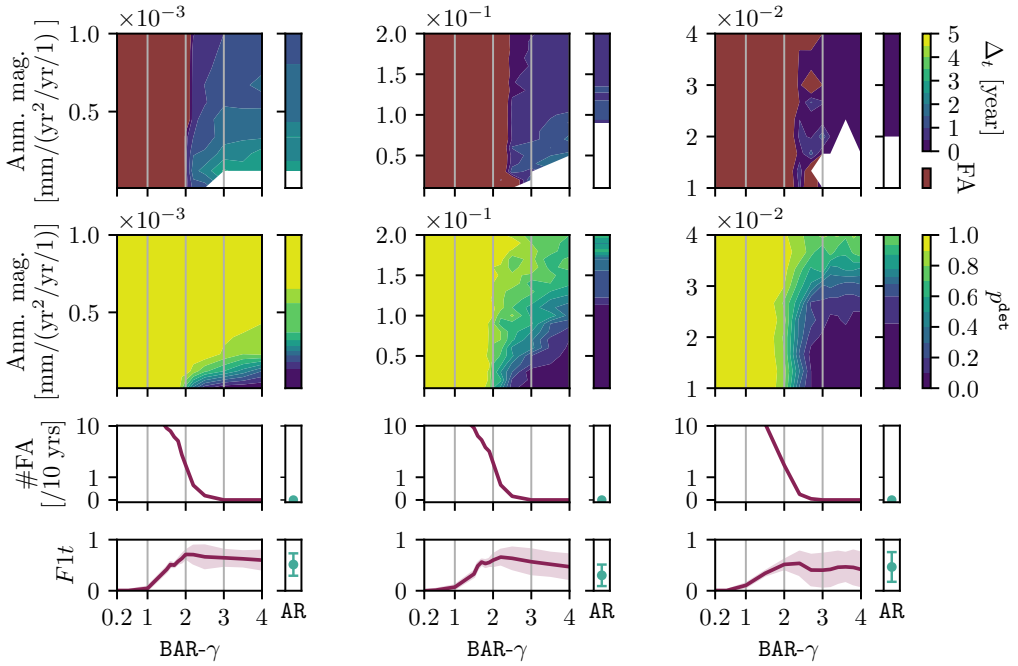
(e) M11D (f) M12D

(g) M13D (h) M14D

Figure B.1 Eight time series measuring the cracking opening (C.O.) and displacement (Disp.) on a dam. They present missing data, different data acquisition frequencies, trends and periodic patterns. All the time series are assumed to be anomaly-free given that no structural deterioration was observed during these periods.

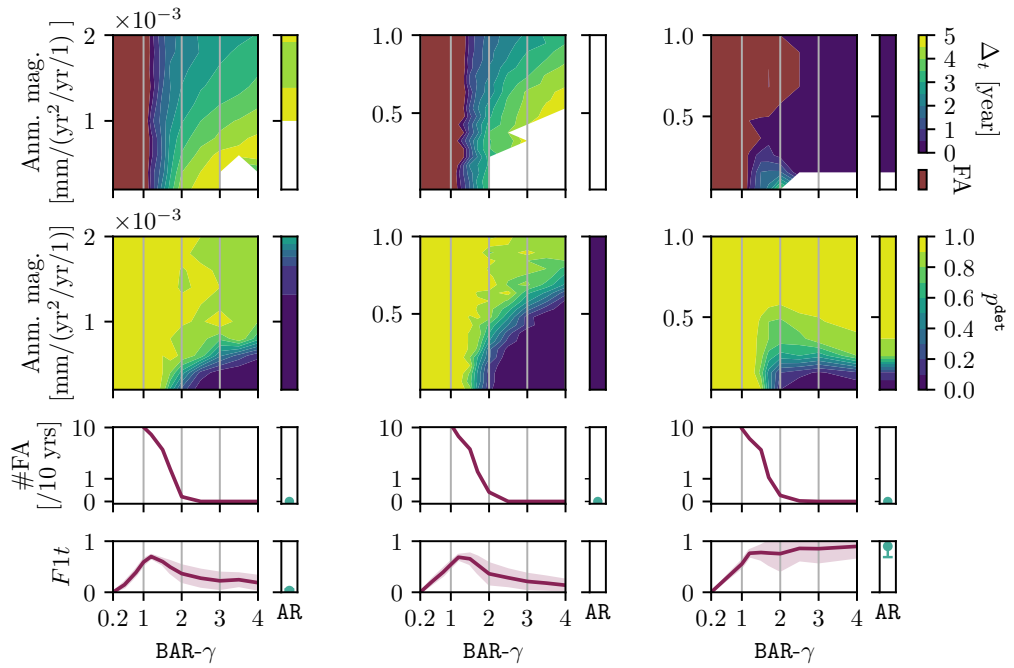
APPENDIX C BAR-γ PERFORMANCE ON SYNTHETIC TIME SERIES IN DAM DATASET

This appendix presents the BAR and the AR performance under different γ values evaluated on the synthetic datasets that are similar to seven time series measured on a dam. Three types of anomalies are introduced in the synthetic time series.



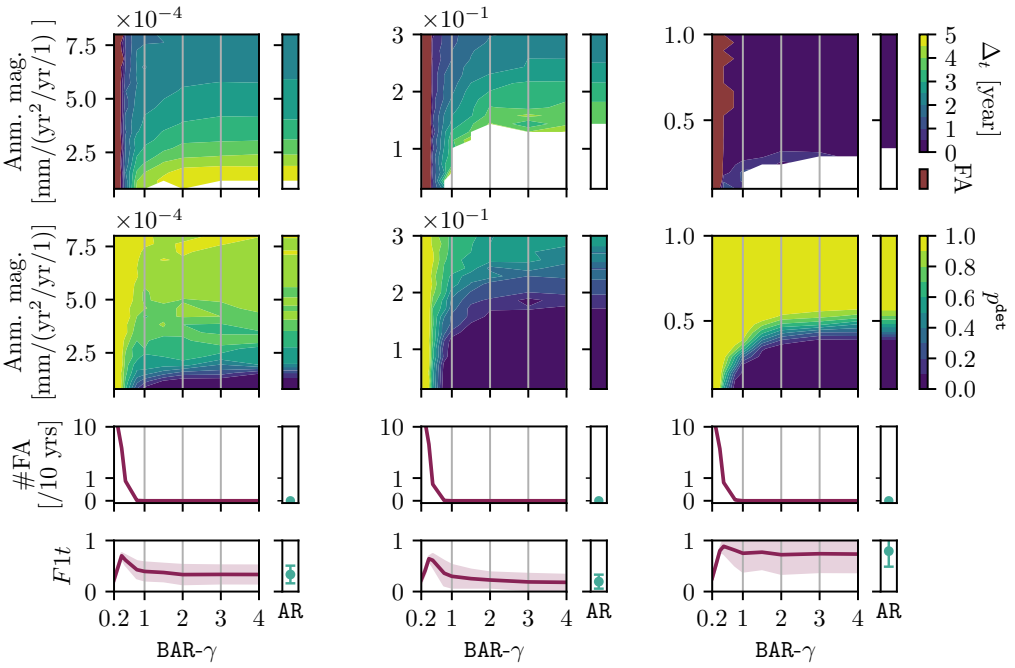
(a) Acceleration anomaly (b) Trend anomaly (c) Level anomaly

Figure C.1 BAR-γ and AR performance on synthetic time series A01C. Repetition N = 50.



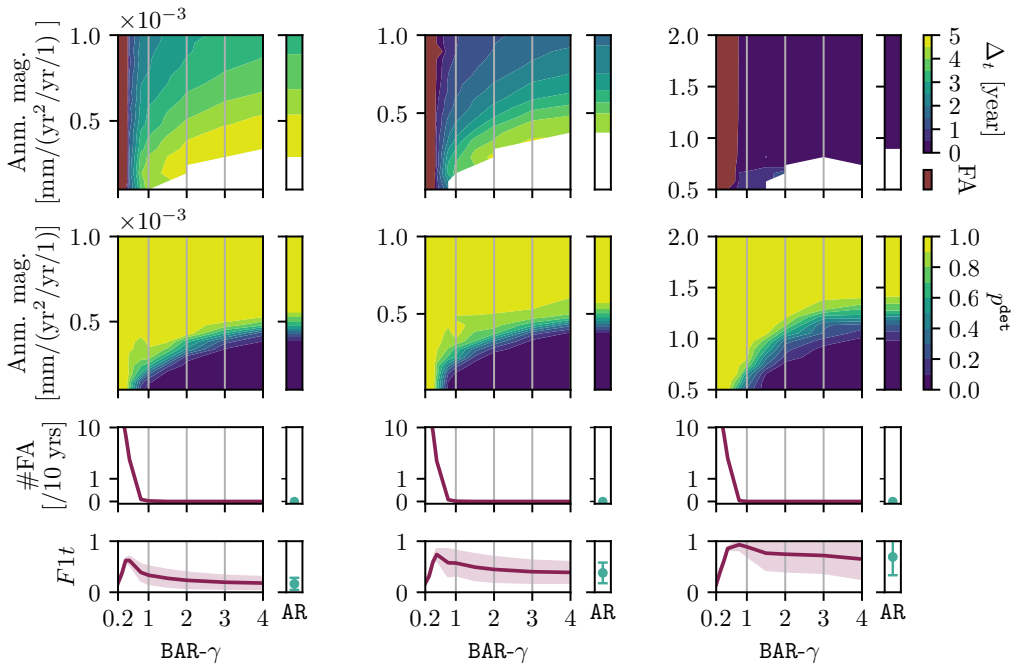
(a) Acceleration anomaly (b) Trend anomaly (c) Level anomaly

Figure C.2 BAR-γ and AR performance on synthetic time series A16D. Repetition N = 50.



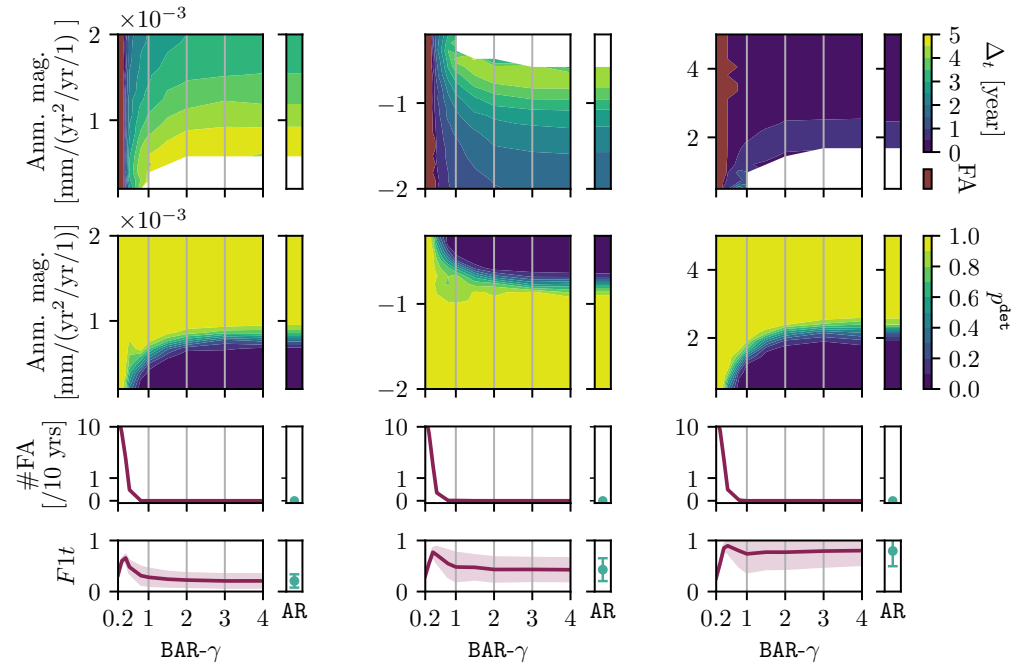
(a) Acceleration anomaly (b) Trend anomaly (c) Level anomaly

Figure C.3 BAR-γ and AR performance on synthetic time series M09C. Repetition N = 100.



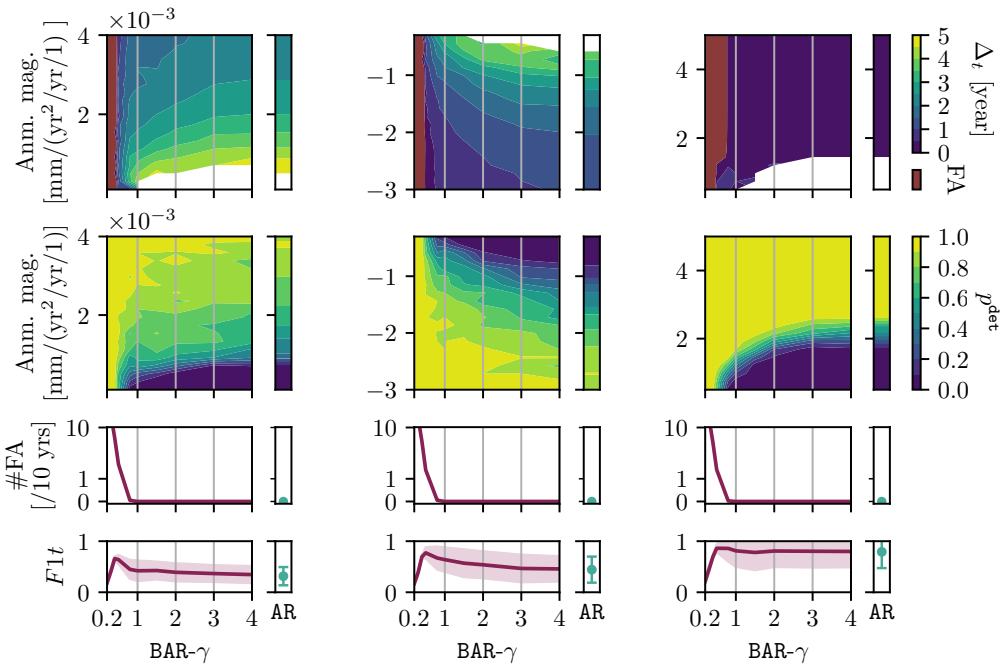
(a) Acceleration anomaly (b) Trend anomaly (c) Level anomaly

Figure C.4 BAR-γ and AR performance on synthetic time series M11D. Repetition N = 100.



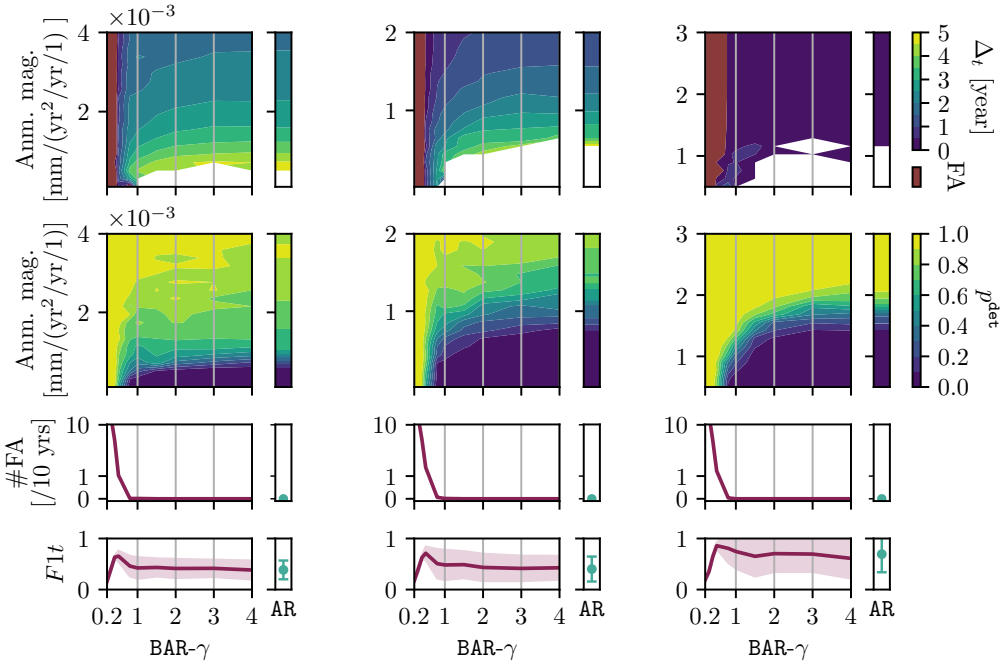
(a) Acceleration anomaly (b) Trend anomaly (c) Level anomaly

Figure C.5 BAR-γ and AR performance on synthetic time series M12D. Repetition N = 100.



(a) Acceleration anomaly (b) Trend anomaly (c) Level anomaly

Figure C.6 BAR-γ and AR performance on synthetic time series M13D. Repetition N = 100.



(a) Acceleration anomaly (b) Trend anomaly (c) Level anomaly

Figure C.7 BAR-γ and AR performance on synthetic time series M14D. Repetition N = 100.

APPENDIX D *F1T* SCORE SUMMARY ON 8 DAM TIME SERIES

This appendix presents the $F1t$ scores for `BAR` and `AR` verified on the synthetic time series involving trend and level anomalies.

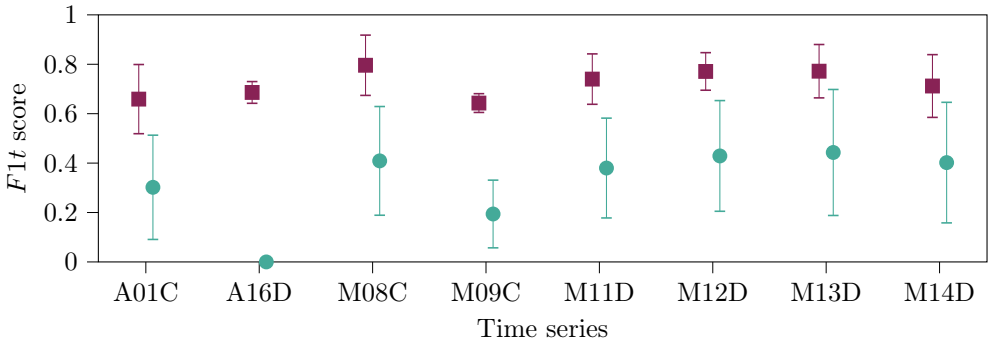


Figure D.1 $F1t$ scores for `BAR` and `AR` verified on the synthetic datasets similar to eight real time series. Trend anomalies are overlaid on the synthetic time series.

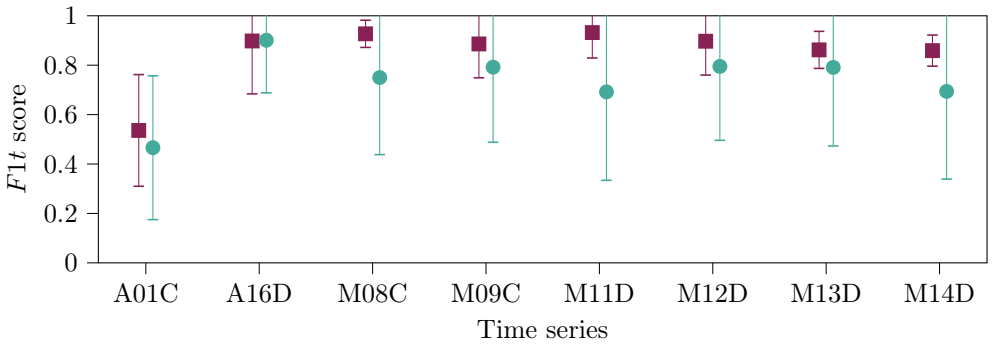
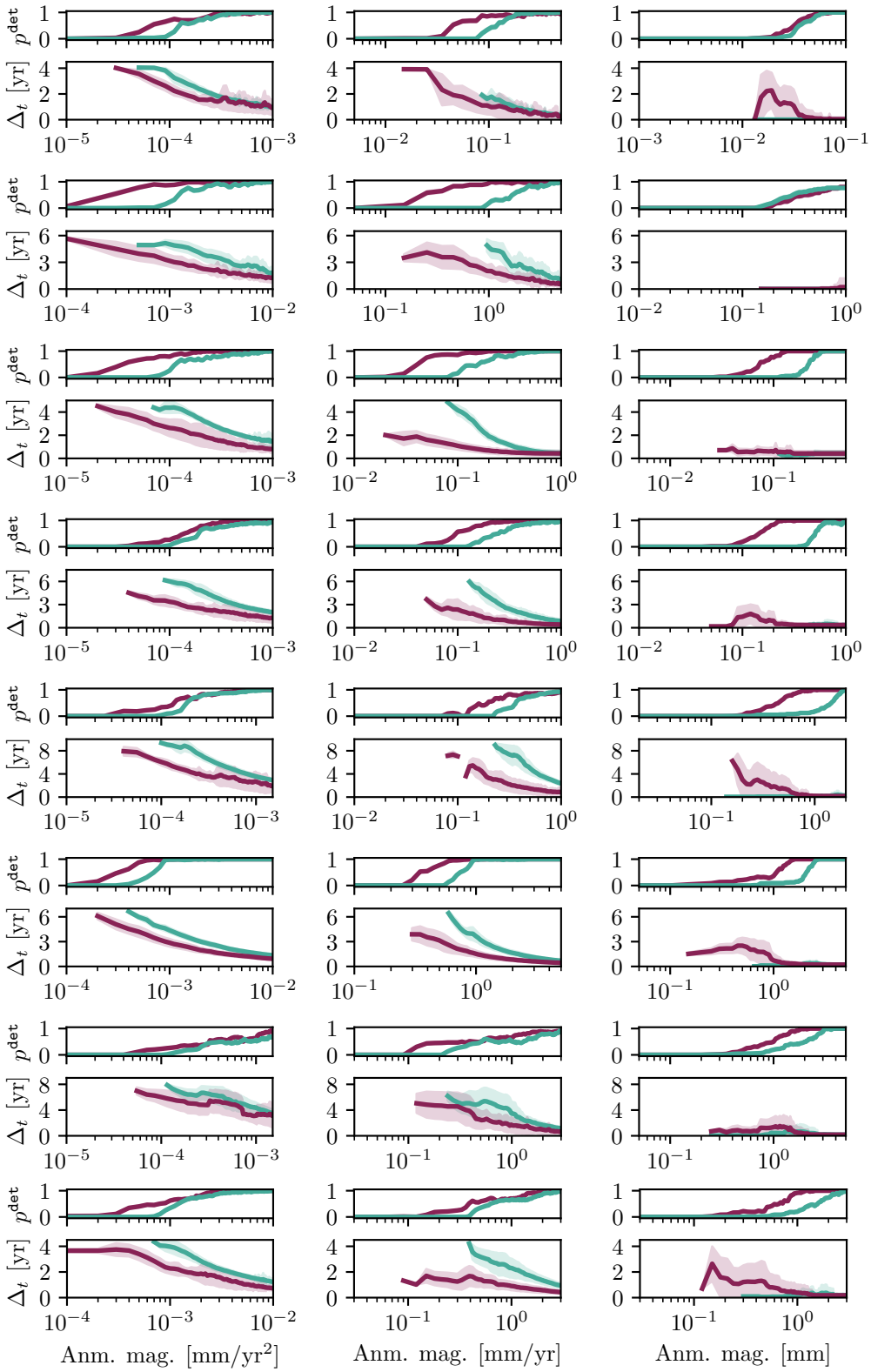


Figure D.2 $F1t$ scores for `BAR` and `AR` verified on the synthetic datasets similar to eight real time series. Level anomalies are overlaid on the synthetic time series.

APPENDIX E BAR-γ PERFORMANCE ON REAL DAM DATASET

This appendix presents the BAR performance with selected γ on the eight real time series measured on a dam involving three types of anomalies. For the plots of Pr, the higher the curve, the better, while for the plots of Δ_t, the lower and the more to the left, the better. Repetition is 30 for the automatically collected time series and 100 for the manually collected ones.



(a) Acceleration anomaly (b) Trend anomaly (c) Level anomaly

Figure E.1 From top to bottom, each row shows the performance of BAR-γ and AR on eight real time series of A01C ($\gamma = 2.8$), A16D ($\gamma = 2.0$), M08C ($\gamma = 0.4$), M09C ($\gamma = 0.4$), M11D ($\gamma = 0.6$), M12D ($\gamma = 0.7$), M13D ($\gamma = 0.9$) and M14D ($\gamma = 0.5$) involving three types of anomalies.

APPENDIX F SSM MATRICES AND PARAMETERS FOR RSI

This appendix provides all the model matrices and their parameters used in RSI.

F.1 Drift Model Matrices

The matrices defining the drift SSM ($\mathcal{M}^{\mathtt{d}}$) are given by

$$\begin{aligned}\mathbf{A}^{\mathtt{d}} &= \text{blockdiag}\left(\mathbf{A}^{\mathtt{LT}}, \mathbf{A}^{\mathtt{AR}}\right),\\ \mathbf{F}^{\mathtt{d}} &= \left[\mathbf{F}^{\mathtt{LT}}, \mathbf{F}^{\mathtt{AR}}\right],\\ \mathbf{Q}^{\mathtt{d}} &= \text{blockdiag}\left(\mathbf{Q}^{\mathtt{LT}}, \mathbf{Q}^{\mathtt{AR}}\right),\end{aligned}$$

where the matrices for `LT` and `AR` are provided in Section 2.3.2. The parameters are $\sigma^{\mathtt{LT}} = 10^{-5}$, $\phi^{\mathtt{AR}}$, while $\sigma^{\mathtt{AR}}$ and Δt are set to the same values as in the base model.

F.2 Model Matrices for Synthetic Time Series Generation

The synthetic time series used in Section 4.3.1 are generated by two SSMs comprising a local trend (`LT`) component, two periodic components with periods of 52 weeks (`PD(52)`) and 13 weeks (`PD(13)`), and an autoregressive (`AR`) component. The matrices defining the SSMs are

$$\begin{aligned}\mathbf{A}^{\mathtt{d}} &= \text{blockdiag}\left(\mathbf{A}^{\mathtt{LT}}, \mathbf{A}^{\mathtt{PD(52)}}, \mathbf{A}^{\mathtt{PD(13)}}, \mathbf{A}^{\mathtt{AR}}\right),\\ \mathbf{F}^{\mathtt{d}} &= \left[\mathbf{F}^{\mathtt{LT}}, \mathbf{F}^{\mathtt{PD(52)}}, \mathbf{F}^{\mathtt{PD(13)}}, \mathbf{F}^{\mathtt{AR}}\right],\\ \mathbf{Q}^{\mathtt{d}} &= \text{blockdiag}\left(\mathbf{Q}^{\mathtt{LT}}, \mathbf{Q}^{\mathtt{PD(52)}}, \mathbf{Q}^{\mathtt{PD(13)}}, \mathbf{Q}^{\mathtt{AR}}\right),\end{aligned}$$

where the matrices for `LT`, `AR`, and `PD(p)` are provided in Section 2.3.2.

The two SSMs share the same parameters except for $\phi^{\mathtt{AR}}$, which is 0.5 for the simple synthetic time series, and 0.9 for the complex one. The shared parameters are the process error standard deviation for each component $\sigma^{\mathtt{LT}} = 0$, $\sigma^{\mathtt{PD(52)}} = 0$, $\sigma^{\mathtt{PD(13)}} = 0$, $\sigma^{\mathtt{AR}} = 2$, and the time interval $\Delta t = 1$ week. The vectors of expected values for the initial hidden states of each component, $\boldsymbol{X}_0^{\mathtt{LT}}$, $\boldsymbol{X}_0^{\mathtt{PD(52)}}$, $\boldsymbol{X}_0^{\mathtt{PD(13)}}$, and $\boldsymbol{X}_0^{\mathtt{AR}}$, are

$$\boldsymbol{\mu}_0^{\mathtt{LT}} = \begin{bmatrix}0\\0\end{bmatrix},\ \boldsymbol{\mu}_0^{\mathtt{PD(52)}} = \begin{bmatrix}0\\25\end{bmatrix},\ \boldsymbol{\mu}_0^{\mathtt{PD(13)}} = \begin{bmatrix}0\\10\end{bmatrix},\ \boldsymbol{\mu}_0^{\mathtt{AR}} = [-0.0621].$$

Their initial covariance matrices are

$$\boldsymbol{\Sigma}_0^{\texttt{LT}} = \begin{bmatrix} 10^{-12} & 0 \\ 0 & 10^{-12} \end{bmatrix}, \ \boldsymbol{\Sigma}_0^{\texttt{PD(52)}} = \begin{bmatrix} 10^{-12} & 0 \\ 0 & 10^{-12} \end{bmatrix},$$
$$\boldsymbol{\Sigma}_0^{\texttt{PD(13)}} = \begin{bmatrix} 10^{-12} & 0 \\ 0 & 10^{-12} \end{bmatrix}, \ \boldsymbol{\Sigma}_0^{\texttt{AR}} = 6.36 \times 10^{-5}.$$

APPENDIX G EXPERIMENTAL SETUP FOR THE RSI AND RSIC EXPERIMENTS

This appendix presents the experimental setup of the detection methods compared in the RSI experiments of Section 4.3 (RSI, SKF, Matrix Profile, Prophet, LSTMED, and TranAD) and in the RSIC experiments of Section 5.3 (RSIC, RSI, SKF, DAMP, Prophet, LSTMED, and TranAD). With the exception of DAMP, which reports the top-k discords rather than alarms based on a threshold, each method quantifies an anomaly score $\eta_t^{\texttt{anm}}$ at each time step t as new observations arrive, and triggers an alarm when $\eta_t^{\texttt{anm}}$ exceeds a threshold $\hat{\eta}^{\texttt{anm}}$. All detection thresholds follow a unified calibration principle: for each method, the threshold is set to the most sensitive value that triggers no alarm on a stationary time series subset $\boldsymbol{y}_{1:\hat{\texttt{t}}}$. This common principle is instantiated through method-specific mechanisms described below, ensuring that all methods are calibrated to the same operating point.

G.1 RSI and RSIC

Both RSI and RSIC quantify the probability of anomaly, $\eta_t^{\texttt{anm}} = p_t^{\texttt{anm}} \in [0, 1]$, as described in Section 4.2.2, using a stationary SSM obtained from the training and validation sets. The hyperparameters of the base SSM-LSTM are tuned using the Optuna package available at `https://github.com/optuna/optuna`. The BNN intervention model uses two fully connected layers with 32 units in the first layer and 64 units in the second layer, ReLU activation functions, He initialization for the prior distributions of the network parameters, a batch size of 20, an early-stopping patience of 10, and $q = 6$, which controls the length of the logarithmically spaced historical input as described in Section 4.2.3, yielding an input length of 64. A sensitivity analysis of the main BNN hyperparameters is presented in Appendix H.

Following the unified calibration principle, the anomaly threshold is defined for all analyses as $\hat{\eta}^{\texttt{anm}} = \max(\max(\eta_{1:\hat{\texttt{t}}}^{\texttt{anm}}) \cdot \gamma,\ 0.1)$, with a margin $\gamma = 1.1$ and a minimum threshold of 0.1 to prevent detections arising from a near-zero anomaly score during the early phase of anomaly development, when the drift is too small to learn an effective intervention.

RSIC shares all of the above hyperparameters with RSI and introduces two additional ones: 1) the maximum classification time $\tilde{t}$; and 2) the quantile value $\tilde{p}$ used for the classification of the intervention type. We use $\tilde{t} = 5$ years and $\tilde{p} = 0.5\%$ in all RSIC experiments. Appendix K presents the sensitivity analysis of these two hyperparameters.

G.2 Switching Kalman Filter

SKF quantifies the probability of anomaly $\eta_t^{\texttt{anm}} = p_t^{\texttt{anm}} \in [0, 1]$ through its regime probabilities, as reviewed in Section 2.3.1. It employs the SSM-LSTM identical to the base model used in RSI and RSIC as its stationary model. This model consists of a local trend component, a LSTM component, and an online autoregressive component, and is trained using the observations from the training and validation sets. The non-stationary model is defined using the same LSTM and autoregressive components, while replacing the local trend component with a local acceleration component.

The SKF hyperparameters, such as the standard deviation of the transition error and the transition probability from the stationary to the non-stationary model, are tuned using the Ray package [133], available at `https://docs.ray.io/en/latest/tune/index.html`. The anomaly threshold is determined by grid search from 0 to 1 until no false alarms are triggered in the stationary training set.

G.3 Matrix Profile and DAMP

The RSI experiments employ the Matrix Profile, which measures the distance between the most recent subsequence of length n, $\boldsymbol{s}_t = \boldsymbol{y}_{t-n:t}$, and a stationary subsequence space $\mathcal{S}^{\texttt{MP}} = \{\boldsymbol{s}_{n+1}, \boldsymbol{s}_{n+2}, \ldots, \boldsymbol{s}_{\bar{\texttt{T}}}\}$, where $\bar{\texttt{T}} < \texttt{T}$, with $\texttt{T}$ the total length of the time series. This distance is used as the anomaly score, i.e., $\eta_t^{\texttt{anm}} = d(\boldsymbol{s}_t, \mathcal{S}^{\texttt{MP}})$. Following the unified calibration principle, its threshold is set to $\hat{\eta}^{\texttt{anm}} = \max(\eta_{1:\hat{\texttt{T}}}^{\texttt{anm}}) \cdot 1.1$, without a lower bound because the score lies in the positive real domain.

The RSIC experiments employ DAMP [47], an online variant computing left-discords, applied with a subsequence length of 52, corresponding to one year of weekly observations, and a zero look-ahead length, equivalent to a purely online operation. The location to start processing is set to the beginning of the test set. Rather than thresholding its score, DAMP identifies the top-k discords as the predicted anomalies for each test time series, by comparing, at each time step t, the query subsequence from $t - 51$ to t with all available past data up to $t - 51$. This approach leverages prior knowledge of the exact number of anomalies in each test time series, information that is not accessible to the other detection methods. By restricting the output to the top-k discords, DAMP inherently avoids triggering more than k false alarms per series, providing a structural advantage in terms of false alarm rate compared to the other methods.

G.4 Prophet

Prophet is typically used as an offline method that fits observations using a parametric regression model incorporating piecewise linear trends, seasonality, and holiday effects. Unlike the other methods, which rely only on part of the time series to construct a stationary model or representation, Prophet must be retrained on all past data; we adapt it to an online setting by retraining the regression model at each time step using all available past observations.

Anomalies are detected from the change in the fitted trend between consecutive segments n and $n+1$, i.e., $\Delta_{\texttt{trend}} = \texttt{trend}_{n+1} - \texttt{trend}_n$. The anomaly score is defined as the maximum trend deviation up to time t, $\eta_t^{\texttt{anm}} = \max(|\Delta_{\texttt{trend}}|_{1:t})$. The threshold is determined through a grid search starting from 0.1 and increasing in increments of 0.1 until no alarm is triggered on the stationary subset.

G.5 LSTMED

We employed the implementation provided by the Merlion [131] package, available at `https://github.com/salesforce/Merlion`, with the default configuration. To train the LSTMED model, we duplicated the stationary training set and introduced a total of 100 anomalies of both level and trend types with different magnitudes and starting times. After training the model on these time series, the anomaly score $\eta_t^{\texttt{anm}}$ is computed using the built-in function *DetectorBase.get_anomaly_score*, and the threshold follows the same calibration as RSI and RSIC: $\hat{\eta}^{\texttt{anm}} = \max(\max(\eta_{1:\hat{\texttt{T}}}^{\texttt{anm}}) \cdot \gamma,\ 0.1)$, with $\gamma = 1.1$.

G.6 TranAD

The implementation available at `https://github.com/imperial-qore/TranAD` is used, with the default configuration and training on the stationary training set, following its standard unsupervised protocol. The model reconstructs each time series segment and compares it with the corresponding observations, with the resulting deviation denoted as the anomaly score $\eta_t^{\texttt{anm}}$, following the procedure described in the original paper. Because this score is two-sided, the unified zero-false-alarm principle is instantiated as an envelope around the stationary subset's score range: an alarm is triggered when the score falls outside the upper bound $\max(\eta_{1:\hat{\texttt{T}}}^{\texttt{anm}}) \cdot 1.1$ or the lower bound $\min(\eta_{1:\hat{\texttt{T}}}^{\texttt{anm}}) \cdot 0.9$.

APPENDIX H HYPERPARAMETER SENSITIVITY IN RSI

The hyperparameters used in the RSI model include

1) hyperparameters associated with the base SSM-LSTM, which are optimized using the Optuna package available at https://github.com/optuna/optuna,

2) the coefficient γ, which controls the anomaly threshold in RSI according to $\hat{\eta}^{\texttt{anm}} = \max(\max(\eta^{\texttt{anm}}_{1:\hat{\texttt{T}}}) \cdot \gamma,\ 0.1)$, as presented in Appendix G, and

3) hyperparameters associated with the intervention model BNN, as presented in Appendix G.

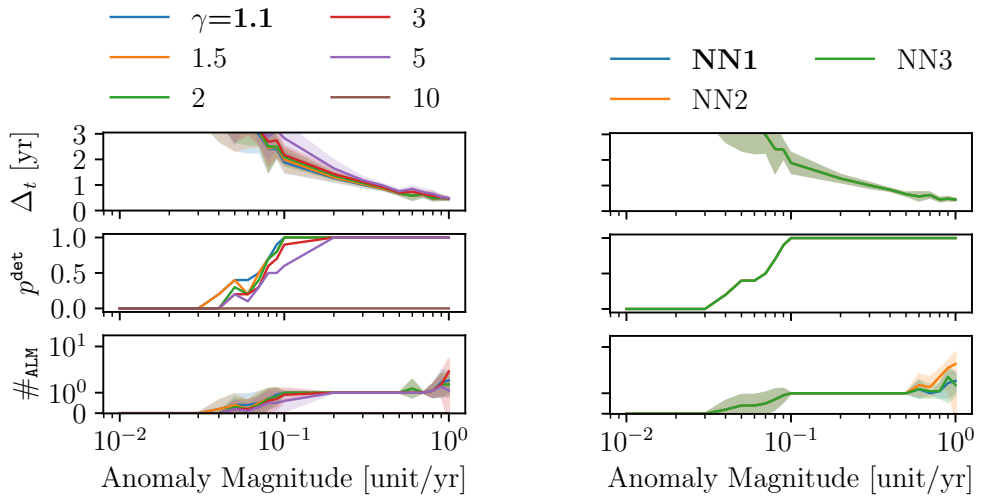
Table H.1 presents the MSEs of the BNN intervention model from the simple synthetic time series on the validation set under different network architectures, batch sizes, patience numbers, and input lengths. The input length refers to the total size of the historical window used to construct the logarithmically spaced input, e.g., an input length of 64 corresponds to $q = 6$ according to equation 4.4. The results show that the network architecture has the largest impact on BNN performance, yielding validation MSEs ranging from 0.12 to 0.23, whereas the other hyperparameters, when varied within reasonable ranges, result in validation MSEs of approximately 0.15.

Table H.1 Mean squared errors (MSEs) of the Bayesian neural network (BNN) used as the RSI intervention model on the validation set for different BNN hyperparameters, including network architecture, batch size, patience number, and input length.

BNN architecture							
Layer1 Unit#	32	64	256	32	*64*	64	256
Layer2 Unit#				32	*32*	64	256
MSE	0.23	0.20	0.17	0.16	*0.15*	0.14	0.12
Batch size	10	*20*	30	40	50		
MSE	0.15	*0.15*	0.15	0.16	0.16		
Patience	5	*10*	15	20	25		
MSE	0.16	*0.15*	0.15	0.15	0.15		
Input length	256	128	*64*	32	16		
MSE	0.16	0.15	*0.15*	0.20	0.25		

To further examine how variations in the validation MSE translate into anomaly detectability, we provide a probabilistic evaluation of the performance of RSI in Figure H.1 under different

network architectures, as well as different values of γ controlling the anomaly threshold. The tested network architectures range from the one yielding the lowest validation MSE (NN3: two layers with 256 units in each layer), through the architecture adopted in all experiments (NN1: two fully connected layers with 32 units in the first layer and 64 units in the second layer), to a more compact architecture achieving the highest validation MSE (NN2: a single layer with 32 units). The results show that γ has a larger impact on anomaly detectability than the network architecture. As γ increases, RSI becomes less sensitive to anomalies, leading to longer detection delays and lower detection probabilities. By contrast, RSI exhibits stable detection performance across different network architectures, despite validation MSEs ranging from 0.12 to 0.23.



(a) γ (b) BNN architecture

Figure H.1 Comparison of the RSI detection performance (detection time Δ_t, detection probability $p^{\mathtt{det}}$, and number of alarms $\#_{\mathtt{ALM}}$) across a range of anomaly magnitudes on validation time series for different (a) γ values and (b) BNN architectures. NN1 denotes the architecture used in all experiments (two fully connected layers with 32 and 64 units), NN2 denotes a single fully connected layer with 32 units, and NN3 denotes two fully connected layers with 256 units in each layer.

APPENDIX I RESIDUAL CHANGES UNDER RSI INTERVENTION

The intervention defined by equations 4.5–4.7 modifies the posterior distributions of the level and trend. By correcting these two baseline components in the new regime, the residual component remains stationary rather than drifting. This appendix presents an example illustrating how the intervention alters the dynamics of the residual hidden state by comparing its prior estimates and autocorrelation function before and after the intervention in Figure I.1.

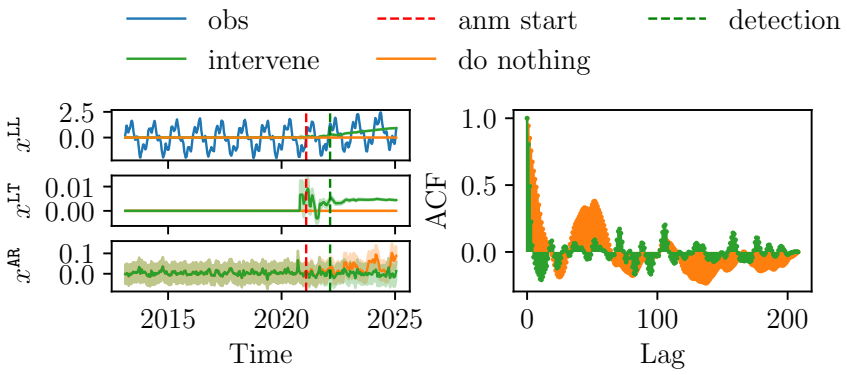


Figure I.1 Comparison of the prior distributions of the autoregressive (`AR`) hidden state and its autocorrelation function (ACF) before and after intervention on a non-stationary time series. The abbreviations "obs" and "anm" denote observations and anomalies, respectively.

When no intervention is applied, the level and trend baselines fail to capture the increasing trend induced by the anomaly. This trend is partially and wrongfully captured by the autoregressive (`AR`) component, causing the `AR` estimates to drift away from the stationary range and leading to a continuously increasing variance. By contrast, when an intervention is applied after the anomaly develops, the `AR` component remains stationary.

The autocorrelation function (ACF) of the prior `AR` estimates for all time steps after the anomaly is shown in the right subfigure. When no intervention is applied, the residual autocorrelation is consistently higher than in the intervened case across all lags, reflecting the model's attempt to accommodate the anomaly-induced drift through increased temporal dependence.

APPENDIX J TRAINING PROCESS OF BNN IN RSIC

The BNNs employed to predict interventions in RSIC are trained using the tractable approximate Gaussian inference (TAGI) method introduced in Section 2.5.2. The training set is constructed from the synthetic time series with anomalies, where we use the base SSM to generate a set of $\mathtt{N}$ synthetic stationary time series of length $\mathtt{L}$, and apply to each time series an anomaly involving either a shift in the level or in the trend with random magnitude and starting time, yielding $\{\boldsymbol{y}_{1:\mathtt{L}}^{\mathtt{anm1}}, \boldsymbol{y}_{1:\mathtt{L}}^{\mathtt{anm2}}, ..., \boldsymbol{y}_{1:\mathtt{L}}^{\mathtt{anm\ N}}\}$. These time series are then processed with the base and drift SSMs in order to generate the input-target training samples. For each time step $2^q + 1 < t < \mathtt{L}$ in all the synthetic time series, the input is collected as $\mu_{i|i}^{\mathtt{Td}},\ \forall i \in \{t, t-1, t-2, t-4, \ldots, t-2^q\}$, with $q \in \mathbb{Z}_{\geq 0}$. The target for the BNN predicting $\boldsymbol{I}^{\mathtt{L}}$ is defined as

$$\text{target}_t = \begin{cases} \{0\}, & \text{if } t < t_0^{\mathtt{anm}} \\ \{x_t^{\mathtt{L,anm}}\}, & \text{otherwise} \end{cases},$$

and for the BNN predicting $\boldsymbol{I}^{\mathtt{T}}$, it is

$$\text{target}_t = \begin{cases} \{0, 0\}, & \text{if } t < t_0^{\mathtt{anm}} \\ \{x_t^{\mathtt{L,anm}}, x_t^{\mathtt{T,anm}}\}, & \text{otherwise} \end{cases},$$

where $x_t^{\mathtt{L,anm}}$ and $x_t^{\mathtt{T,anm}}$ are the level and trend of the synthetic anomaly ($\mathtt{anm}\ n$) at time t, and $t_0^{\mathtt{anm}}$ denotes the anomaly starting time.

APPENDIX K HYPERPARAMETER SENSITIVITY IN RSIC

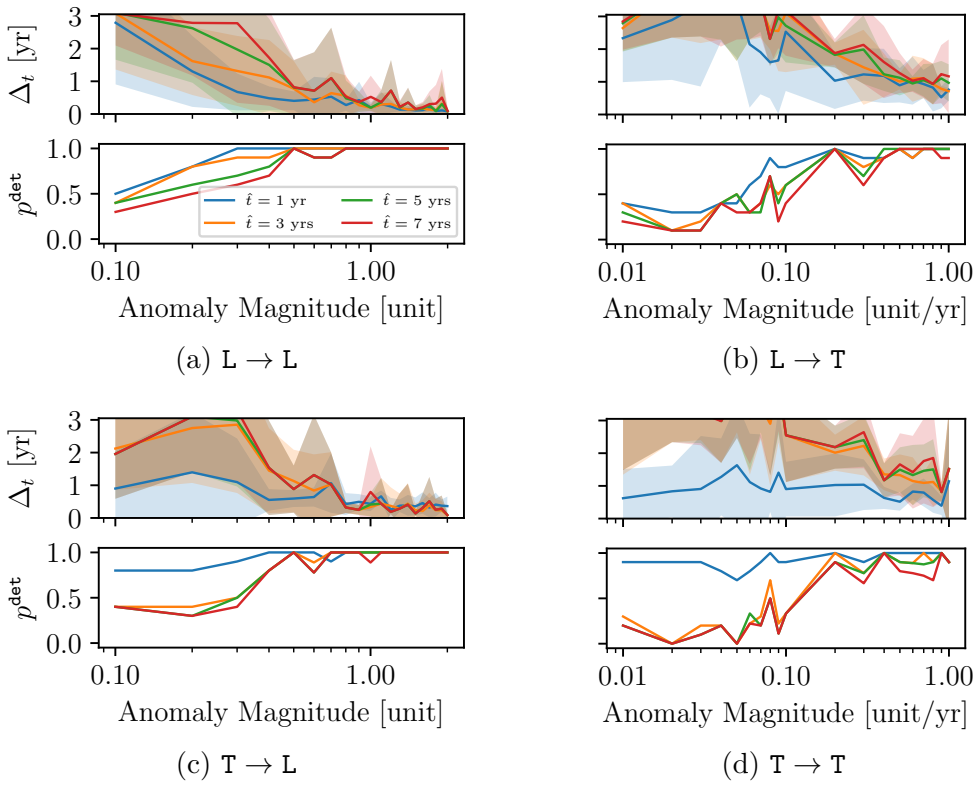
In addition to the hyperparameters inherited from RSI, the RSIC method introduces two hyperparameters that govern its classification procedure: the maximum waiting time $\tilde{t}$, which bounds the period after a detection during which RSIC may apply an intervention, and the quantile value $\tilde{p}$, which sets the confidence level required to classify the intervention type from the Beta posterior. This appendix examines the sensitivity of RSIC to these two hyperparameters and explains the choice of the values $\tilde{t} = 5$ years and $\tilde{p} = 0.5\%$ adopted throughout the experiments in Section 5.3. Section K.1 analyzes the influence of the maximum waiting time, while Section K.2 analyzes that of the quantile value. All experiments reported in this section are conducted on a separate validation set, distinct from the test set used in Section 5.3.

K.1 Maximum Waiting Time

The maximum waiting time $\tilde{t}$ bounds the period after a detection during which RSIC may apply an intervention. Starting from $t_{\texttt{det}}$, the method accumulates observations and applies the corresponding intervention as soon as one of the two intervention types is classified with sufficient confidence. If neither type reaches the required confidence within $\tilde{t}$, no intervention is applied. Figure K.1 reports the time to detection Δ_t, and the detection probability $p^{\texttt{det}}$ for $\tilde{t} \in \{1, 3, 5, 7\}$ years across the four anomaly configurations, and Table K.1 summarizes the corresponding false alarm rate, baseline estimation errors, and classification accuracy.

Across all anomaly configurations, increasing $\tilde{t}$ leads to a longer time to detection and a lower detection probability. Conversely, setting $\tilde{t} = 1$ year yields the highest detection probability and the shortest time to detection. This behaviour stems from the fact that a short waiting window leaves little time to reach the classification confidence, so an intervention is less likely to be classified and applied, frequently leaving the model uncorrected after the first anomaly. The resulting baseline deviation accumulates, and when the second anomaly occurs, the model departs further from the observations, becoming more prone to detect it. This oversensitivity to the second anomaly, however, comes at the expense of a higher false alarm rate. With $\tilde{t} = 1$ year, the false alarm rate rises sharply to between 4.75 and 7.62 per ten years, whereas it drops below one per ten years once $\tilde{t} \geq 5$ years in all configurations.

This tradeoff is mirrored in the baseline estimation and classification results. The local-level estimation error MSE_{L} is severely degraded at $\tilde{t} = 1$ year, reaching values one order



(a) L → L (b) L → T

(c) T → L (d) T → T

Figure K.1 Time to detection Δ_t and the detection probability $p^{\mathtt{det}}$ under different values of maximum waiting time $\tilde{t}$ across four anomaly configurations: level to level (L→L), level to trend (L→T), trend to level (T→L), and trend to trend (T→T).

Table K.1 Performance metrics and classification accuracy with regard to different $\tilde{t}$ on synthetic time series with different combinations of anomalies: level to level (L→L), level to trend (L→T), trend to level (T→L), and trend to trend (T→T).

Anomaly	$\tilde{t}$ [years]	# false alarm per 10 years	MSE_{L}	MSE_{T}	Classification accuracy	
					1st anm	2nd anm
L → L	1	5.07	0.341	1.05×10^{-5}	0.30	0.20
	3	1.41	0.110	1.44×10^{-5}	0.66	0.42
	5	**0.95**	**0.065**	$\mathbf{1.37 \times 10^{-5}}$	**0.82**	**0.47**
	7	0.82	0.064	1.42×10^{-5}	0.82	0.46
L → T	1	4.75	0.446	3.04×10^{-5}	0.34	0.16
	3	1.12	0.076	1.38×10^{-5}	0.65	0.74
	5	**0.90**	**0.040**	$\mathbf{1.49 \times 10^{-5}}$	**0.79**	**0.81**
	7	0.86	0.035	1.51×10^{-5}	0.81	0.81
T → L	1	7.48	0.633	2.21×10^{-5}	0.12	0.04
	3	1.12	0.055	1.23×10^{-5}	0.96	0.26
	5	**0.76**	**0.042**	$\mathbf{1.17 \times 10^{-5}}$	**1.00**	**0.29**
	7	0.67	0.037	1.17×10^{-5}	1.00	0.28
T → T	1	7.62	0.890	3.03×10^{-5}	0.15	0.00
	3	0.78	0.043	1.14×10^{-5}	0.92	0.63
	5	**0.66**	**0.022**	$\mathbf{1.16 \times 10^{-5}}$	**0.98**	**0.73**
	7	0.66	0.024	1.18×10^{-5}	0.99	0.72

of magnitude larger than those obtained with longer waiting times. This reflects the fact that a short waiting window frequently leaves the model uncorrected, allowing errors to accumulate in the baseline. The local-trend estimation error $\mathrm{MSE}_{\mathtt{T}}$, by contrast, remains comparatively insensitive to $\tilde{t}$. The classification accuracy follows the same pattern: a short waiting time provides too few observations to reliably distinguish the two intervention types, so the accuracy for the first anomaly is low at $\tilde{t} = 1$ year and improves markedly as $\tilde{t}$ increases, approaching its maximum around $\tilde{t} = 5$ years. The accuracy for the second anomaly remains lower in configurations involving a second level anomaly, consistent with the inherited trend error discussed in Section 5.3.1.

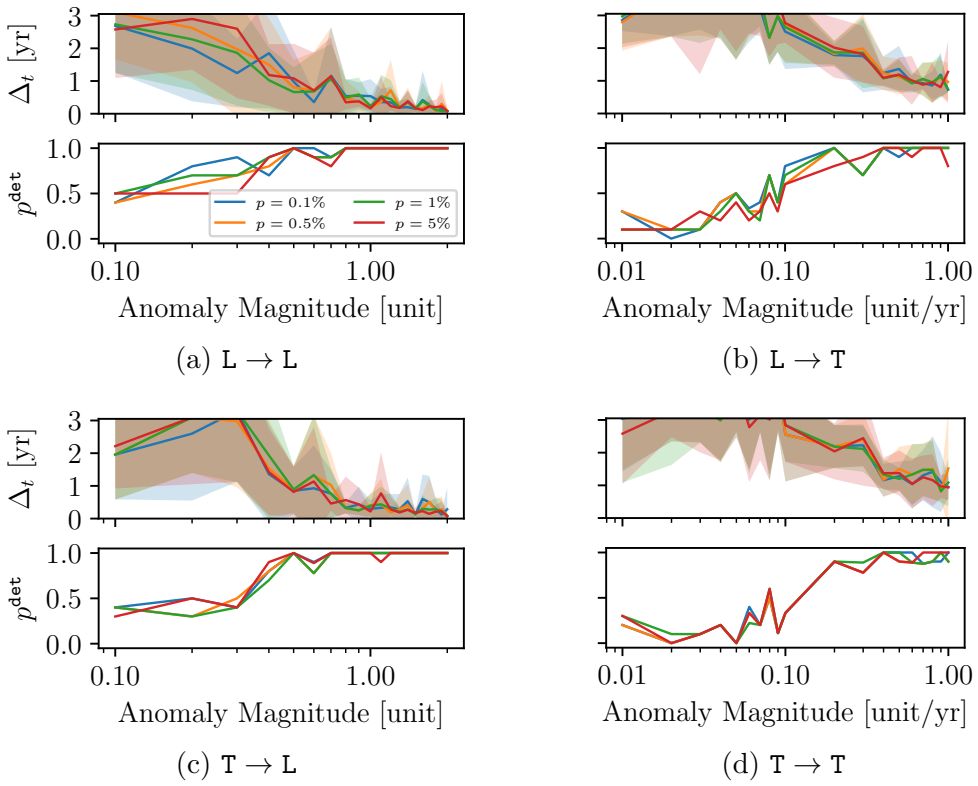
The selection of $\tilde{t}$ is therefore governed by the tradeoff between detectability and false alarm rate. Requiring the false alarm rate to remain below one per ten years across all configurations excludes $\tilde{t} = 1$ and $\tilde{t} = 3$ years, while the gains from $\tilde{t} = 5$ to $\tilde{t} = 7$ years are marginal across all metrics. We therefore set $\tilde{t} = 5$ years in all experiments, retaining a low false alarm rate and high classification accuracy without sacrificing detectability.

K.2 p Quantile

The quantile value $\tilde{p}$ sets the confidence required to classify the intervention type. Following the rule in Section 5.2.5, an intervention is applied only when the $\tilde{p}$-quantile of the Beta posterior for the corresponding type exceeds 0.5. A smaller $\tilde{p}$ evaluates the posterior further into its lower tail and therefore imposes a stricter requirement before an intervention can be applied, whereas a larger $\tilde{p}$ relaxes this condition and allows an earlier classification. Figure K.2 reports the time to detection Δ_t and the detection probability $p^{\mathtt{det}}$ for $\tilde{p} \in \{0.1\%, 0.5\%, 1\%, 5\%\}$ across the four anomaly configurations, and Table K.2 summarizes the corresponding false alarm rate, baseline estimation errors, classification accuracy, and time for diagnosis.

In contrast to the maximum waiting time, the detectability and the false alarm rate are comparatively insensitive to the choice of $\tilde{p}$. As shown in Figure K.2, the time to detection and detection probability curves nearly overlap across the four quantile values in all anomaly configurations. The false alarm rate similarly varies within a narrow band, remaining close to or below one per ten years across all values, with only a mild increase at $\tilde{p} = 5\%$ in the configurations involving a first level anomaly, where it reaches 1.34 and 1.06 per ten years. The baseline estimation errors $\mathrm{MSE}_{\mathtt{L}}$, $\mathrm{MSE}_{\mathtt{T}}$, and the classification accuracy also remain stable.

The selection of $\tilde{p}$ is therefore governed by the trade-off between the false alarm rate and the



(a) L → L (b) L → T

(c) T → L (d) T → T

Figure K.2 Time to detection Δ_t and the detection probability $p^{\mathtt{det}}$ under different quantile values $\tilde{p}$ across four anomaly configurations: level to level (L→L), level to trend (L→T), trend to level (T→L), and trend to trend (T→T).

Table K.2 Performance metrics and classification accuracy with regard to different quantile value $\tilde{p}$ on synthetic time series with different combinations of anomalies: level to level (L→L), level to trend (L→T), trend to level (T→L), and trend to trend (T→T).

Anomaly	$\tilde{p}$	# FA per 10 years	MSE_{L}	MSE_{T}	Class. acc.		Time for class.
					1st anm	2nd anm	[years]
L → L	0.1%	0.80	0.079	1.08×10^{-5}	0.83	0.49	1.96 ± 2.25
	0.5%	**0.95**	**0.065**	$\mathbf{1.37 \times 10^{-5}}$	**0.82**	**0.47**	$\mathbf{1.74 \pm 2.18}$
	1%	1.03	0.060	1.66×10^{-5}	0.76	0.47	1.49 ± 2.02
	5%	1.34	0.053	2.51×10^{-5}	0.60	0.48	0.95 ± 1.53
L → T	0.1%	0.75	0.043	1.23×10^{-5}	0.84	0.77	2.76 ± 1.81
	0.5%	**0.90**	**0.040**	$\mathbf{1.49 \times 10^{-5}}$	**0.79**	**0.81**	$\mathbf{2.68 \pm 1.75}$
	1%	0.89	0.037	1.60×10^{-5}	0.75	0.85	2.67 ± 1.77
	5%	1.06	0.035	1.96×10^{-5}	0.65	0.84	2.69 ± 1.78
T → L	0.1%	0.71	0.043	1.26×10^{-5}	0.99	0.22	1.79 ± 2.15
	0.5%	**0.76**	**0.042**	$\mathbf{1.17 \times 10^{-5}}$	**1.00**	**0.29**	$\mathbf{1.65 \pm 2.07}$
	1%	0.79	0.040	1.17×10^{-5}	1.00	0.3	1.43 ± 1.98
	5%	0.78	0.037	1.13×10^{-5}	0.99	0.38	1.09 ± 1.69
T → T	0.1%	0.62	0.024	1.18×10^{-5}	0.97	0.69	3.13 ± 1.9
	0.5%	**0.66**	**0.022**	$\mathbf{1.16 \times 10^{-5}}$	**0.98**	**0.73**	$\mathbf{3.12 \pm 1.86}$
	1%	0.59	0.022	1.15×10^{-5}	0.98	0.72	3.11 ± 1.85
	5%	0.63	0.022	1.18×10^{-5}	0.99	0.79	2.93 ± 1.9

time for diagnosis. Requiring the false alarm rate to remain below one per ten years across all configurations excludes $\tilde{p} = 1\%$ and $\tilde{p} = 5\%$, leaving $\tilde{p} = 0.1\%$ and $\tilde{p} = 0.5\%$. Among these two, $\tilde{p} = 0.5\%$ yields a shorter time for diagnosis while preserving a comparable false alarm rate and classification accuracy. We therefore set $\tilde{p} = 0.5\%$ in all experiments, providing a faster classification without compromising robustness.