Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

GlucoKoopman-UQ: Spectrally stable controlled Koopman dynamics for multi-horizon glucose forecasting, glycaemic-risk estimation, and empirical prediction intervals

Jean Marie Vianney Nyirimana

anne4505@gmail.com

University of Rwanda

Archana Mathur

Nitte Meenakshi Institute of Technology

Alexander Ngenzi

University of Rwanda

Research Article

Keywords: type 1 diabetes; continuous glucose monitoring; glucose forecasting; Koopman operator; uncertainty quantification; hypoglycaemia

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11846866/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

GlucoKoopman-UQ: Spectrally stable controlled Koopman dynamics for multi-horizon glucose forecasting, glycaemic-risk estimation, and empirical prediction intervals

Jean Marie Vianney Nyirimana[1,*], Archana Mathur[2], Alexander Ngenzi[3]

[1]African Centre of Excellence in Data Science (ACE-DS), College of Business and Economics, University of Rwanda, Kigali, Rwanda
[2]Nitte Meenakshi Institute of Technology, Bangalore, India
[3]College of Science and Technology, University of Rwanda, Kigali, Rwanda

[*]Corresponding author: Jean Marie Vianney Nyirimana, anne4505@gmail.com

Author contributions

Jean Marie Vianney Nyirimana: conceptualization, methodology, software, formal analysis, investigation, data curation, visualization, writing – original draft. Archana Mathur: supervision, methodology, validation, writing – review and editing. Alexander Ngenzi: supervision, validation, writing – review and editing. All authors read and approved the final manuscript.

Acknowledgments

We thank the people who took part in the OhioT1DM studies, and the Ohio University team that collected the data and makes it available to researchers.

GlucoKoopman-UQ: Spectrally stable controlled Koopman dynamics for multi-horizon glucose forecasting, glycaemic-risk estimation, and empirical prediction intervals

Abstract

Purpose: Koopman models move a learned latent state forward with a linear operator, which makes multi-step forecasting cheap to compute and easy to analyse. We asked whether this structure helps glucose forecasting in type 1 diabetes, and in particular whether forcing the operator to be stable pays off at longer horizons. Our model, GlucoKoopman-UQ, encodes the recent history of each participant, rolls the latent state forward under an operator with spectral radius below one plus a decaying treatment term, and outputs glucose, hypo- and hyperglycaemia probabilities, and prediction intervals. **Methods:** We used the 2018 and 2020 OhioT1DM cohorts (12 participants), 8-hour causal histories, and targets 15 to 60 minutes ahead. The model and five comparators were each trained with five seeds and scored on held-out test files. The primary endpoint was participant-level root mean squared error (RMSE) at 60 minutes. Exact Wilcoxon tests with Holm correction were used throughout, including for 13 ablations. **Results:** At 60 minutes the RMSE was 30.84 mg/dL (95% CI 28.74–33.04), lower than every comparator (adjusted $p = 0.0024$). Coverage of the 90% intervals ranged from 0.915 to 0.925. A ridge autoregressive model was slightly better at 15 minutes. Dropping the stability constraint helped at 15 minutes but hurt at 60 minutes in all 12 participants (adjusted $p = 0.0029$). Without treatment inputs, 60-minute RMSE rose by 1.47 mg/dL. **Conclusions:** The stability constraint buys long-horizon accuracy at a small short-horizon cost. Testing on unseen people and better hypoglycaemia recall are the main open problems.

Keywords: type 1 diabetes; continuous glucose monitoring; glucose forecasting; Koopman operator; uncertainty quantification; hypoglycaemia

1 Introduction

People with type 1 diabetes adjust insulin, food and activity many times a day, usually with only a rough idea of where their glucose is heading. Continuous glucose monitoring (CGM) gives a reading every five minutes, but the signal it produces is shaped by forces the sensor cannot see: insulin that is still acting, a meal being absorbed, exercise earlier in the day, stress, illness, and the time of day [1, 2]. A forecast that looks 30 or 60 minutes ahead has to account for these hidden drivers, and it has to say how confident it is.

The forecast horizon changes what the problem looks like. Over 15 minutes, glucose rarely moves far, so simply repeating the last reading is a surprisingly hard baseline, and a linear autoregression is harder still [3, 4]. Over an hour, recent insulin and carbohydrate start to dominate, and a model that only extrapolates the CGM trace loses ground. This is why a single average error across horizons can hide what a model is actually good at. Two further traps are common in this literature. Hypoglycaemia is rare, so a model can post a high area under the ROC curve while missing most events; and when thousands of overlapping windows come from a handful of people, window-level statistics overstate how much evidence there is.

Koopman operator theory offers a different way to build a forecaster. Instead of learning a nonlinear recurrent transition, one learns a lifting of the data into a latent space where the dynamics are close to

linear, and then propagates that latent state with a matrix [5, 6]. The appeal for glucose is practical. A linear rollout can be written in closed form for any number of steps, its long-run behaviour is governed by the eigenvalues of one matrix, and insulin and meals can enter as an explicit control input. What has not been shown is whether this structure helps on real CGM data, or under what conditions.

This paper reports GlucoKoopman-UQ, a forecaster built around a controlled Koopman rollout whose transition matrix is kept stable ($\rho(\mathbf{K}) < 1$). It makes four contributions. First, we derive a bound showing that, under this constraint, an error in the encoded state shrinks geometrically as the forecast is rolled forward (Proposition 1). The bound makes a testable prediction: stability should matter little at 15 minutes and more at 60. Second, one rollout feeds three outputs: point forecasts, three-class glycaemic-risk probabilities, and residual-quantile prediction intervals, which we evaluate together. Third, we test the prediction from the bound directly, with a participant-level interaction analysis on 13 ablations. Fourth, every model is compared in the same causal pipeline, with five seeds, the participant as the unit of analysis, and corrected exact tests, following the TRIPOD+AI reporting guidance [7] and recommended practice for comparing learners across datasets [8]. The prespecified primary hypothesis was that GlucoKoopman-UQ would have lower participant-level 60-minute RMSE than five classical and neural comparators.

2 Related work

2.1 Forecasting models and how they are compared

The OhioT1DM releases, which pair CGM with insulin, carbohydrate, activity and wearable signals, are the most widely used public benchmark for this task [1]. Published results on them are hard to line up against each other. Studies differ in which release they use, how long a history the model sees, how gaps are filled, which windows count as valid targets, and whether a person's own past data were in the training set. An early review of personalised prediction already pointed to this spread of set-ups [9]. Benchmarks that hold the pipeline fixed tell a sobering story: well-tuned classical and linear models often come close to deep networks, especially at short horizons [3, 4]. The same lesson appears in general forecasting, where a single linear layer can beat Transformer models on standard benchmarks [10].

Deep learning for glucose has nonetheless moved quickly. Convolutional–recurrent networks, dilated recurrent networks and CNN–LSTM hybrids learn short and long patterns in the CGM trace [11–13]. Other work has focused on the inputs: time-domain features, feature transformations, missing-data handling and engineered glucose dynamics [14–16], sometimes combined in ensembles [17]. Architectures from general forecasting are also being tried, including LSTMs [18], N-BEATS [19], patch-based Transformers [20] and the Temporal Fusion Transformer, which handles static, known-future and past-only covariates separately [21]. A recent comparison of Transformer variants found that the best one depended on horizon and history length, using OhioT1DM as an external test set [22], and foundation models pretrained on many time series are now being tested on CGM [23].

Glucose responses differ a great deal from one person to the next, so many groups personalise. Options include fine-tuning a shared network on each person [24], multitask learning across people [25, 26], meta-learning and evidential learning for people with little data [27–29], and multimodal attention or policy-conditioned models [30, 31]. Wearable-based systems [32], edge deployment [33] and federated training [34] add practical constraints. One distinction matters for reading all of this work: forecasting the future of a person whose earlier data trained the model is a different task from forecasting

for someone the model has never seen. Insulin and meals can also be modelled mechanistically, as in the UVA/Padova simulator used for in-silico trials [35], but in observational data their usefulness depends on how completely and accurately they were logged. Some studies find that preprocessing and feature availability move accuracy as much as the choice of network [4].

2.2 Uncertainty, calibration and event risk

A point forecast says nothing about how much to trust it. Recurrent networks that predict a variance, probabilistic and evidential models, Monte Carlo dropout and uncertainty-aware Temporal Fusion Transformers all attach some measure of spread [27, 36–38]. A predicted variance, however, does not guarantee that a 90% interval contains the truth 90% of the time; that has to be checked on held-out data. Conformal prediction turns residuals into intervals with a finite-sample coverage guarantee when the calibration and test points are exchangeable [39]. CGM windows from the same person overlap and drift, so exchangeability is doubtful. Time-series variants [40–42] and adaptive conformal inference, which adjusts its target online to restore long-run coverage under shift [43], are designed for this situation.

Event prediction raises its own issues. Hypoglycaemia occurs in only a few per cent of windows, so precision–recall curves say more than ROC curves [44], and neural class probabilities are often overconfident until recalibrated, for example by temperature scaling [45]. Recent multitask systems predict glucose and hypoglycaemia together [30]. Good discrimination on retrospective windows, though, is not evidence that a model is safe to use for treatment.

2.3 Koopman models and the gap

Koopman analysis began as a spectral tool for reducing complex flows [46]. Neural networks later made it practical to learn the lifting from data, through Koopman-invariant subspaces and autoencoders [47, 48], and recent sequence models such as consistent Koopman autoencoders [49] and Koopa [50] use learned operators for forecasting. Dynamic mode decomposition with control [51] and Koopman predictors for model predictive control [52] brought in external inputs, and stability-aware learned Koopman control has followed [53]. The only CGM application we found used dynamic mode decomposition on short sliding windows of glucose and flagged meals when the estimated dynamics lost stability [54].

Two points from this literature shaped our design. A Koopman model with no control input cannot represent the effect of insulin or a meal, which are the main reasons glucose changes over an hour. And an unconstrained operator applied many times can amplify its own errors. Azencot et al. argued the opposite trade-off: for data with strong short-term dependence, relaxing strict stability can improve short-range prediction [49]. If both views are right, the value of a stability constraint should depend on the horizon. As far as we could find, no published glucose model learns a controlled Koopman operator under a spectral constraint and uses it to produce forecasts, risk estimates and intervals at several horizons, and no study has checked whether the benefit of stability changes with horizon.

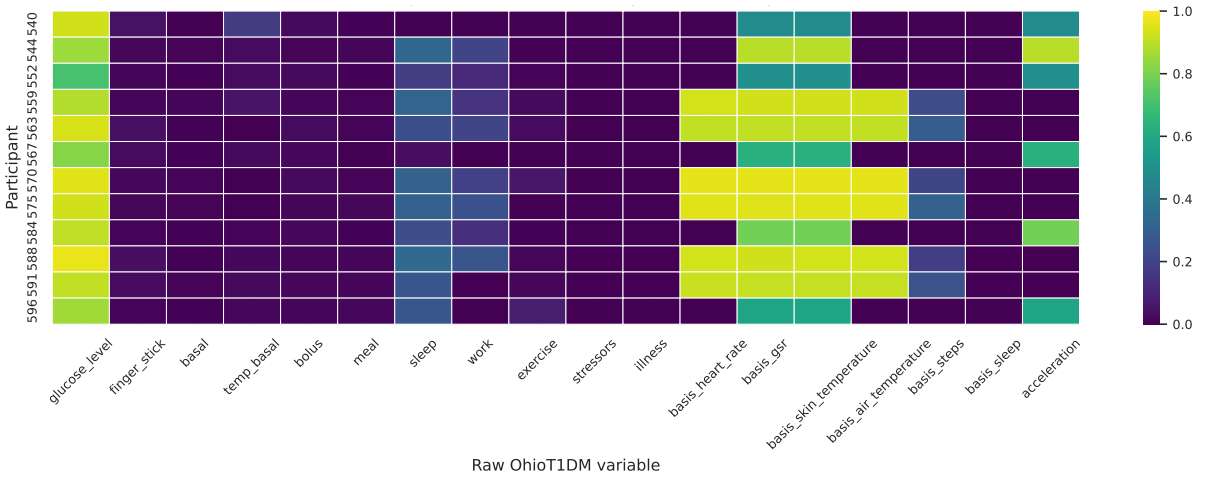

Fig. 1 Availability of raw OhioT1DM channels by participant. The main model uses only channels present in both cohorts; sparse and cohort-specific channels were tested in ablations

3 Materials and methods

3.1 Data and study design

We reanalysed the 2018 and 2020 OhioT1DM releases [1]. Each release contains six people with type 1 diabetes, giving 12 participants. Every participant has one training file and one later test file; after removing byte-identical duplicates in the archive, 24 XML files remained. The files hold CGM and finger-stick glucose, basal and bolus insulin, meals, exercise, sleep, work, stress, illness and hypoglycaemia logs, and wearable signals from BioHarness or Basis devices. In total there were 166,533 CGM readings. CGM availability per participant averaged 0.887 (range 0.708–0.967). Galvanic skin response and skin temperature were recorded for all 12 participants, while heart rate, air temperature, steps and acceleration exist for only six (Fig. 1; file-level counts in Online Resource 2, Tables S1–S2).

Because the same people appear in the training and test files, this design measures how well a model forecasts later periods for people it has already seen. It does not measure generalisation to new people. We therefore treat the participant, not the window, as the unit of inference: overlapping windows from one person are not independent replicates.

3.2 Data splits and leakage control

All signals were placed on a 5-minute grid using only past values. Each participant's test file was held out until final evaluation. Inside the training file, the last 15% of the timeline became a validation segment, separated from the fitting data by a 60-minute gap. We used this segment for model selection and early stopping, and afterwards to set interval widths and probability temperatures. The first hour of each test file could serve as history but never as a target. Scalers and imputation values were estimated on the fitting data only, and no test label was used for fitting, selection or calibration. Automated tests checked the time ordering, scaler isolation, target observability and window filters.

3.3 Preprocessing and window eligibility

Missing covariates were forward-filled for a limited time and paired with an indicator of whether the value was actually observed, so that a filled value is never confused with a measured zero. Let $m_{t,j} \in \{0,1\}$ mark whether channel j was observed at time t, let $\tau_j(t) = \max\{u < t : m_{u,j} = 1\}$ when such a time exists, and let b_j be the fill limit for that channel. Imputation and robust scaling were

$$\widetilde{x}_{t,j} = \begin{cases} x_{t,j}, & m_{t,j} = 1, \\ x_{\tau_j(t),j}, & \substack{\tau_j(t) \text{ exists}, \\ t-\tau_j(t) \le b_j}, \\ \mu_j^{\mathrm{tr}}, & \text{otherwise}, \end{cases} \tag{1}$$

$$x^{\star}_{t,j} = \frac{\widetilde{x}_{t,j} - \mu_j^{\mathrm{tr}}}{q^{\mathrm{tr}}_{0.75,j} - q^{\mathrm{tr}}_{0.25,j} + \varepsilon}, \tag{2}$$

with the median μ_j^{tr} and quartiles $q^{\mathrm{tr}}_{a,j}$ taken from the fitting data. The model receives both $x^{\star}_{t,j}$ and $m_{t,j}$. Channels missing entirely for a participant received the training fallback and a mask of zeros. Engineered inputs used only the present and past: backward glucose differences, velocity and acceleration, trailing glucose statistics, sine and cosine of the time of day, and causal insulin-on-board (IOB) and carbohydrate-on-board (COB) estimates.

A window was kept if at least 70% of its history had real CGM readings, the latest reading was at most 10 minutes old, no CGM gap exceeded 30 minutes, and every target was a measured value between 40 and 400 mg/dL. Filled values were never used as targets. This left 28,190 test windows at each horizon, or 2,349.2 per participant on average.

3.4 Forecasting task

At forecast origin t the input is eight hours of history,

$$\mathbf{X}_t = [\mathbf{x}_{t-95}, \ldots, \mathbf{x}_t] \in \mathbb{R}^{96\times 27}. \tag{3}$$

With step set $\mathscr{H} = \{3,6,9,12\}$ (15, 30, 45 and 60 minutes), the targets and glycaemic-state labels are

$$\mathbf{y}_t = [g_{t+h}]_{h\in\mathscr{H}} \in \mathbb{R}^4, \tag{4}$$

$$r_{t,h} = \begin{cases} 0, & g_{t+h} < 70, \\ 1, & 70 \le g_{t+h} \le 180, \\ 2, & g_{t+h} > 180, \end{cases} \tag{5}$$

following international consensus CGM ranges [2, 55]. The 27 input features cover current and past CGM, finger-stick glucose, insulin and IOB, carbohydrate and COB, galvanic skin response and skin temperature, the engineered glucose dynamics, time of day, and observation masks. Channels available for only half the participants, rare behavioural events and weekday were left out of the main model and tested as ablations.

3.5 GlucoKoopman-UQ

Encoder Glucose behaves differently at different time scales, so the encoder looks at the history at four resolutions, $\mathscr{S} = \{1,3,6,12\}$ steps (5, 15, 30 and 60 minutes). Each view $\mathscr{P}_s(\mathbf{X}_t)$ has its own encoder and a learned gate:

$$\mathbf{h}_t^{(s)} = f_{\theta_s}(\mathscr{P}_s(\mathbf{X}_t)), \qquad \mathbf{a}_t^{(s)} = \sigma\left(\mathbf{W}_s \mathbf{h}_t^{(s)} + \mathbf{b}_s\right), \tag{6}$$

$$\mathbf{h}_t = \sum_{s\in\mathscr{S}} \mathbf{a}_t^{(s)} \odot \mathbf{h}_t^{(s)}, \qquad \mathbf{z}_t = g_\theta(\mathbf{h}_t, \mathbf{e}_p), \tag{7}$$

where σ is the logistic function and $\mathbf{e}_p$ is a learned embedding of participant p. For a person not seen in training, the code uses the mean embedding; this path is unit-tested but was not evaluated here.

Controlled latent rollout The latent state is advanced one 5-minute step at a time:

$$\mathbf{z}_{t+k} = \mathbf{K}\mathbf{z}_{t+k-1} + \exp[-\boldsymbol{\lambda}(k-1)] \odot \mathbf{c}_t, \tag{8}$$

where $\mathbf{K}$ is the learned Koopman matrix, $\mathbf{c}_t$ encodes treatment and meal inputs, and the non-negative rates $\boldsymbol{\lambda}$ let their influence fade. The operator is constrained so that $\rho(\mathbf{K}) < 1$; across the five seeds the final spectral radius was between 0.934 and 0.939. Unrolling Eq. (8) gives

$$\mathbf{z}_{t+k} = \mathbf{K}^k \mathbf{z}_t + \sum_{j=1}^{k} \mathbf{K}^{k-j} \left[\exp\{-\boldsymbol{\lambda}(j-1)\} \odot \mathbf{c}_t\right]. \tag{9}$$

The first term carries the encoded state forward; the second adds up the fading effect of insulin and food. The next result shows what the constraint buys.

Proposition 1. *Let $\rho(\mathbf{K}) < \gamma < 1$. Then there is an induced matrix norm $\|\cdot\|_*$ with $\|\mathbf{K}\|_* \le \gamma$, and for every step $k \ge 1$*

$$\|\mathbf{z}_{t+k}\|_* \le \gamma^k \|\mathbf{z}_t\|_* + \frac{1-\gamma^k}{1-\gamma}\, \bar{c}_{t,k}, \tag{10}$$

$$\|\mathbf{z}_{t+k} - \widetilde{\mathbf{z}}_{t+k}\|_* \le \gamma^k \|\mathbf{z}_t - \widetilde{\mathbf{z}}_t\|_*, \tag{11}$$

where $\bar{c}_{t,k} = \max_{1\le j\le k} \|\exp\{-\boldsymbol{\lambda}(j-1)\} \odot \mathbf{c}_t\|_*$ and $\widetilde{\mathbf{z}}_{t+k}$ is the rollout from a perturbed encoding $\widetilde{\mathbf{z}}_t$ with the same control input.

Proof. A standard result [56, Lemma 5.6.10] states that every square matrix admits, for each $\varepsilon > 0$, an induced norm in which its norm is at most $\rho(\mathbf{K}) + \varepsilon$. Choosing $\varepsilon = \gamma - \rho(\mathbf{K})$ gives $\|\mathbf{K}\|_* \le \gamma$, hence $\|\mathbf{K}^m \mathbf{v}\|_* \le \gamma^m \|\mathbf{v}\|_*$ for any vector $\mathbf{v}$. Taking norms in Eq. (9) and using the triangle inequality yields $\|\mathbf{z}_{t+k}\|_* \le \gamma^k \|\mathbf{z}_t\|_* + \sum_{j=1}^{k} \gamma^{k-j} \bar{c}_{t,k}$, and summing the geometric series gives Eq. (10). Since $\boldsymbol{\lambda} \ge \mathbf{0}$, each entry of $\exp\{-\boldsymbol{\lambda}(j-1)\}$ lies in $(0,1]$, so $\bar{c}_{t,k}$ is bounded by a constant times $\|\mathbf{c}_t\|_2$ for all k, and the trajectory stays bounded. The control terms cancel when two rollouts are subtracted, leaving $\mathbf{z}_{t+k} - \widetilde{\mathbf{z}}_{t+k} = \mathbf{K}^k(\mathbf{z}_t - \widetilde{\mathbf{z}}_t)$, which gives Eq. (11). □

In words, Eq. (11) says that a mistake in the encoded state is damped by a factor γ at every step. An unconstrained operator offers no such protection: its errors can persist or grow, although its extra freedom

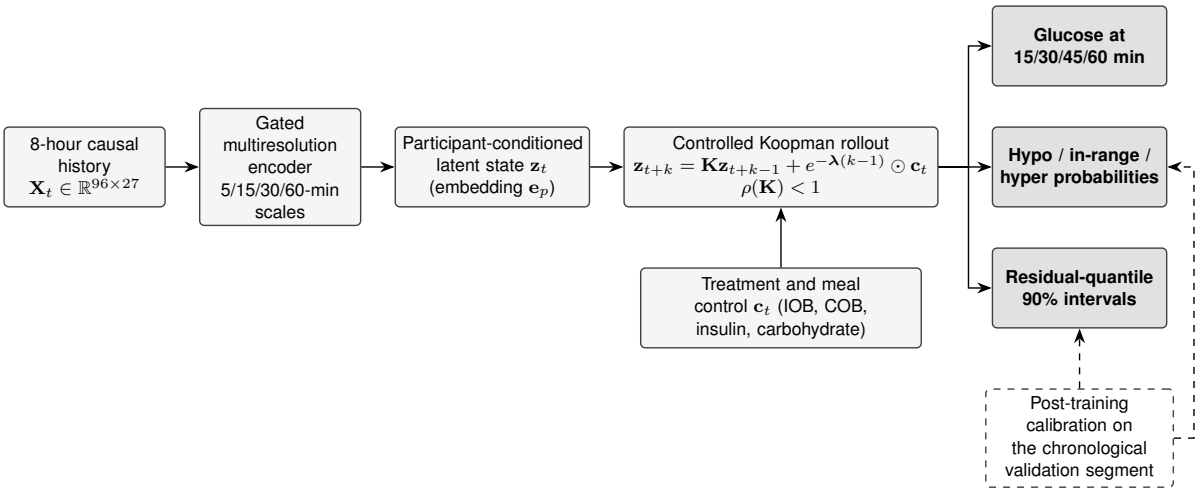

Fig. 2 GlucoKoopman-UQ architecture. The multiresolution encoder turns the multimodal history into a participant-conditioned latent state. That state is stepped forward by a stable Koopman matrix ($\rho(\mathbf{K}) < 1$) while insulin and meal inputs are added with fading weight; three decoders then read off glucose, class probabilities and intervals. Dashed arrows mark the calibration done after training

may fit the next few steps better. If the decoders are Lipschitz with constant L_h, the output error inherits the bound $L_h \gamma^k \|\mathbf{z}_t - \widetilde{\mathbf{z}}_t\|_*$. We therefore expected the constraint to cost little or even hurt at 15 minutes and to help at 60. Section 4.10 tests this.

Output heads and training objective For each horizon the rolled-forward state is decoded into a forecast, a positive scale and three risk logits:

$$\widehat{y}_{t,h} = d_{\mu,h}(\mathbf{z}_{t+h}), \qquad \widehat{s}_{t,h} = \mathrm{softplus}(d_{s,h}(\mathbf{z}_{t+h})) + \varepsilon, \tag{12}$$

$$\widehat{\mathbf{p}}_{t,h} = \mathrm{softmax}(\boldsymbol{\ell}_{t,h}), \qquad \boldsymbol{\ell}_{t,h} = d_{r,h}(\mathbf{z}_{t+h}). \tag{13}$$

The training loss combines a probabilistic forecast term with cross-entropy on the risk classes and parameter regularisation:

$$\mathscr{L}(\Theta) = \frac{1}{N|\mathscr{H}|} \sum_{i=1}^{N} \sum_{h\in\mathscr{H}} \left[\ell_{\mathrm{forecast}}(y_{i,h}, \widehat{y}_{i,h}, \widehat{s}_{i,h}) - \boldsymbol{\lambda}_r \sum_{c=0}^{2} r_{i,h,c} \log \widehat{p}_{i,h,c} \right] + \boldsymbol{\lambda}_\Theta \mathscr{R}(\Theta), \tag{14}$$

where $r_{i,h,c}$ is the one-hot class label. The model has 162,505 trainable parameters. Loss weights, layer sizes and optimiser settings were fixed in one configuration file shared by all seeds and ablations; the file is distributed with the code. The architecture is summarised in Fig. 2.

3.6 Post-training calibration

Interval widths were set from absolute residuals $e_{i,h} = |y_{i,h} - \widehat{y}_{i,h}|$ on the validation segment, $i = 1, \ldots, n_h$. For miscoverage $\alpha = 0.10$ we used the standard split-conformal quantile,

$$k_h = \lceil (n_h + 1)(1-\alpha) \rceil, \qquad q_h = e_{(\min\{k_h, n_h\}),h}, \tag{15}$$

$$\mathscr{C}_{t,h}^{1-\alpha} = [\widehat{y}_{t,h} - q_h,\ \widehat{y}_{t,h} + q_h], \tag{16}$$

where $e_{(k),h}$ is the kth smallest residual. Risk logits were rescaled by a temperature fitted on the same segment [45]:

$$\widehat{p}_{i,h,c}^{(T)} = \frac{\exp(\ell_{i,h,c}/T_h)}{\sum_{d=0}^{2}\exp(\ell_{i,h,d}/T_h)}, \qquad T_h^{\star} = \arg\min_{T_h>0} -\sum_{i=1}^{n_h}\sum_{c=0}^{2} r_{i,h,c}\log \widehat{p}_{i,h,c}^{(T)}. \tag{17}$$

Neither q_h nor $T_h^{\star}$ saw test labels. Because the validation segment had also been used for model selection, and because neighbouring windows are correlated, the formal split-conformal guarantee does not apply [39]. We report the resulting intervals as residual-quantile intervals and judge them by their measured coverage. Marginal coverage also says nothing about coverage within a particular person or glucose range [40–42].

3.7 Comparators

Five comparators were fixed before the analysis. *Persistence* repeats the latest glucose value. *Circadian persistence* predicts from glucose at the same time of day. *Ridge-ARX* is a regularised linear autoregression with exogenous inputs, and *KNN-ARX* a nearest-neighbour regression on the same kind of inputs. *DLinear* is a small direct linear network that has proved a strong baseline in general forecasting [10]. All comparators used the same causal preprocessing and were scored on exactly the same windows. During output checks we found one negative ARX extrapolation, so both ARX models were clipped to non-negative glucose before calibration, and all affected runs were recomputed before any test was carried out. Every comparator is reported, including where it beat our model.

3.8 Training

Neural models trained for at most 160 epochs with early stopping after 22 epochs without validation improvement, under seeds 17, 41, 73, 101 and 137. The seed controlled initialisation, data order and dropout. We kept the best validation checkpoint of each run. GlucoKoopman-UQ stopped after 26–28 epochs (best epoch 4–6), whereas DLinear ran for 144–160. Everything was trained on CPU (Online Resource 2, Table S14).

3.9 Ablations

Thirteen variants were each trained with the same five seeds (65 runs). Six changed the architecture: no participant embedding, no multiresolution encoder, no control input, an unconstrained operator, no risk objective, and a deterministic head. Seven changed the inputs: no treatment features, no shared wearable features, CGM-only inputs, and the addition of weekday, behavioural, rare-event or cohort-limited wearable channels. Each variant kept everything else identical. We had also planned sensitivity runs with 60-, 120- and 360-minute histories, a < 54 mg/dL hypoglycaemia threshold, per-participant training and leave-one-participant-out (LOPO) evaluation. These were not run (Online Resource 2, Table S13).

3.10 Outcomes

The primary endpoint was the mean over participants of each participant's RMSE at 60 minutes:

$$\mathrm{MAE}_{p,h} = \frac{1}{n_{p,h}} \sum_{i=1}^{n_{p,h}} |y_{p,i,h} - \widehat{y}_{p,i,h}|, \qquad \mathrm{RMSE}_{p,h} = \Big[\frac{1}{n_{p,h}} \sum_{i=1}^{n_{p,h}} (y_{p,i,h} - \widehat{y}_{p,i,h})^2 \Big]^{1/2}, \tag{18}$$

$$\overline{\mathrm{RMSE}}_h = \frac{1}{P} \sum_{p=1}^{P} \mathrm{RMSE}_{p,h}, \qquad P = 12. \tag{19}$$

Secondary outcomes were MAE, RMSE, mean absolute percentage error and R^2 at every horizon. For an interval $[L_{i,h}, U_{i,h}]$ we report coverage and mean width,

$$\mathrm{PICP}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} \mathbb{I}\{L_{i,h} \le y_{i,h} \le U_{i,h}\}, \qquad \mathrm{MPIW}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} (U_{i,h} - L_{i,h}), \tag{20}$$

together with normalised width and the Winkler score. Risk outputs were scored with class-wise AUROC, AUPRC, precision, recall and F1, expected calibration error (ECE) and the multiclass Brier score. Clinical agreement was summarised with Clarke error-grid zones [57] and Bland–Altman bias and limits of agreement [58], keeping in mind that error grids were built for meter accuracy rather than forecasting [59].

3.11 Statistical analysis

Each metric was first averaged over the five seeds within a participant,

$$\overline{M}_{p,m,h} = \frac{1}{S} \sum_{s=1}^{S} M_{p,m,h,s}, \qquad S = 5, \tag{21}$$

and all inference was done on these 12 participant values. We report means, standard deviations, medians and participant-level bootstrap 95% confidence intervals. Differences among models were screened with the Friedman statistic [8]

$$Q_F = \frac{12P}{M(M+1)} \sum_{m=1}^{M} \left(\overline{R}_m - \frac{M+1}{2} \right)^2, \tag{22}$$

where M is the number of models and $\overline{R}_m$ is the mean rank of model m. Pairwise contrasts between GlucoKoopman-UQ and each comparator used exact two-sided Wilcoxon signed-rank tests on the paired differences d_p, with Holm adjustment [60]

$$p^{\mathrm{Holm}}_{(j)} = \max_{1 \le k \le j} \min\{1, (J-k+1)\, p_{(k)}\} \tag{23}$$

over the five comparators within each horizon and metric. Ablations were analysed the same way, with separate Friedman tests for the architecture and feature suites and Holm correction over all 13 variants. Differences are always GlucoKoopman-UQ (or variant) minus the comparator, so negative error differences favour our model. Event prevalence and pooled Bland–Altman statistics are descriptive.

Post hoc interaction test After the prespecified analyses, Proposition 1 led us to ask whether each architecture component mattered more at 60 minutes than at 15. For component a and participant p, let

Table 1 Accuracy, risk and interval results of GlucoKoopman-UQ at each horizon

Horizon (min)	MAE (mg/dL)	RMSE (mg/dL)	R^2	Hypo recall	Hypo AUPRC	Hyper recall	Hyper AUPRC	PICP	MPIW (mg/dL)	Clarke A+B (%)
15	7.559	11.189	0.959	0.623	0.475	0.924	0.959	0.925	39.463	99.812
30	13.195	18.709	0.886	0.529	0.398	0.889	0.927	0.922	67.585	99.330
45	18.239	25.243	0.793	0.439	0.330	0.851	0.885	0.920	90.670	98.688
60	22.592	30.837	0.692	0.383	0.273	0.815	0.840	0.915	109.729	97.906

Means over the 12 participants, each first averaged over five seeds. PICP, observed coverage of the nominal 90% intervals; MPIW, mean interval width; AUPRC, area under the precision–recall curve.

$\Delta_{p,a,h}$ be the variant's seed-averaged RMSE minus the full model's. The contrast

$$I_{p,a} = \Delta_{p,a,60} - \Delta_{p,a,15} \tag{24}$$

is above zero if losing the component costs more accuracy at 60 minutes than at 15. Each set of 12 contrasts was tested against zero with an exact Wilcoxon test, with Holm correction over the six components. We also averaged $\Delta_{p,a,h}$ over the six components to get a composite effect. These tests were not planned in advance and are reported as exploratory.

3.12 Reproducibility and use of AI tools

An automated audit confirmed that all 1,440 required runs (six models, 12 participants, four horizons, five seeds) and all 65 ablation runs were present, and 35 unit tests cover the data handling, model stability, calibration and statistics. The aggregate results behind every table and figure, and the scripts that check them, are in Online Resource 1. An AI assistant (Claude, Anthropic) helped draft and edit parts of the text, check reference details, and write the scripts for the post hoc analysis and for redrawing three figures. All numbers come from the authors' analysis outputs and were checked against them by script. The authors reviewed every part of the manuscript and take full responsibility for it.

4 Results

4.1 Test windows and event rates

Each horizon had 28,190 eligible test windows. Hypoglycaemia was uncommon, with 754, 753, 759 and 767 positive windows at 15, 30, 45 and 60 minutes (2.67–2.72%). Hyperglycaemia was common, at 35.54–35.63% of windows (Online Resource 2, Table S3). With so few hypoglycaemic windows, AUPRC is the more honest summary for that class.

4.2 Point accuracy

Error grew steadily with horizon. MAE went from 7.559 mg/dL at 15 minutes to 22.592 mg/dL at 60 minutes, RMSE from 11.189 to 30.837 mg/dL, and R^2 fell from 0.959 to 0.692 (Table 1).

On the primary endpoint, GlucoKoopman-UQ had the lowest mean participant-level 60-minute RMSE, 30.837 mg/dL (95% CI 28.739–33.035; Table 2). The mean paired reductions were 45.550 mg/dL against circadian persistence, 7.184 against persistence, 3.664 against DLinear, 1.626 against Ridge-ARX and 1.367 against KNN-ARX. Every adjusted p-value was 0.0024, the smallest value an exact test can give

Table 2 Primary endpoint: mean participant-level RMSE (mg/dL) at 60 minutes for each model

Model	Mean	SD	Median	95% CI
Circadian persistence	76.386	14.575	70.264	[68.742, 84.702]
DLinear	34.501	4.569	34.023	[32.113, 37.019]
GlucoKoopman-UQ	30.837	3.960	30.115	[28.739, 33.035]
KNN-ARX	32.203	4.211	31.380	[29.965, 34.501]
Persistence	38.020	4.703	37.812	[35.557, 40.679]
Ridge-ARX	32.463	4.021	31.222	[30.394, 34.703]

Each participant's five seeds were averaged first, giving $n = 12$. Intervals come from resampling participants. Lower is better.

Table 3 Paired comparisons of 60-minute RMSE between GlucoKoopman-UQ and each comparator

Comparator	Mean paired Δ	Rank-biserial	Holm p	Significant
Circadian persistence	-45.550	-1.000	0.0024	Yes
DLinear	-3.664	-1.000	0.0024	Yes
KNN-ARX	-1.367	-0.974	0.0024	Yes
Persistence	-7.184	-1.000	0.0024	Yes
Ridge-ARX	-1.626	-1.000	0.0024	Yes

Δ is GlucoKoopman-UQ minus the comparator (mg/dL), so negative values favour GlucoKoopman-UQ. p-values are from exact two-sided Wilcoxon signed-rank tests on 12 participants, Holm-adjusted over the five comparators.

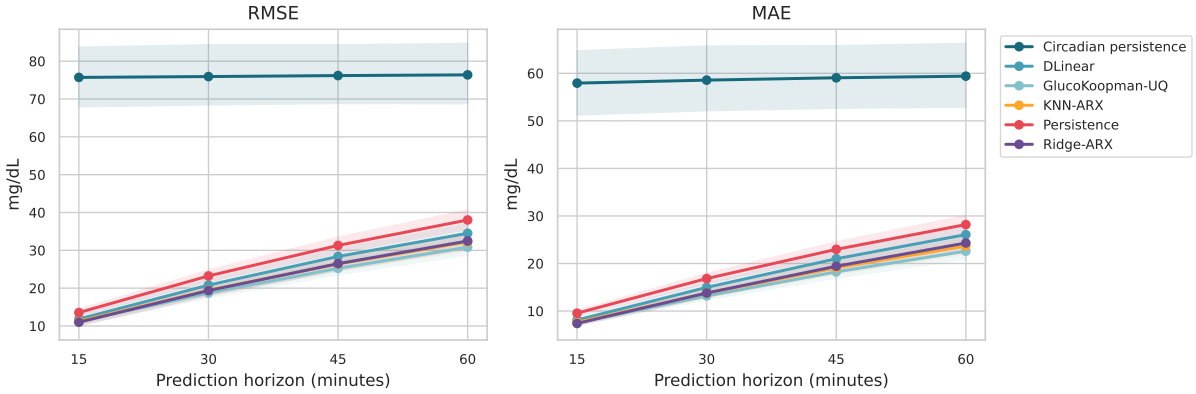

Fig. 3 MAE and RMSE by forecast horizon for all six models. Each point is a mean over participants after averaging the five seeds within each participant; bands are bootstrap 95% confidence intervals. Lower is better

with 12 participants after correction for five comparisons (Table 3). Friedman tests were significant for MAE and RMSE at every horizon ($p < 0.001$; Online Resource 2, Tables S4–S8). Fig. 3 shows the full horizon profile, and Fig. 4 the paired effects against each comparator.

The advantage did not hold at 15 minutes. There, Ridge-ARX had a slightly lower MAE (7.396 versus 7.559 mg/dL) and RMSE (10.990 versus 11.189 mg/dL). The 0.163 mg/dL MAE difference was significant after correction ($p = 0.0322$), but the 0.200 mg/dL RMSE difference was not ($p = 0.1846$), and GlucoKoopman-UQ did not differ significantly from KNN-ARX (Online Resource 2, Table S9). From 30 minutes onwards, GlucoKoopman-UQ had lower MAE and RMSE than every comparator (adjusted $p \leq 0.0029$).

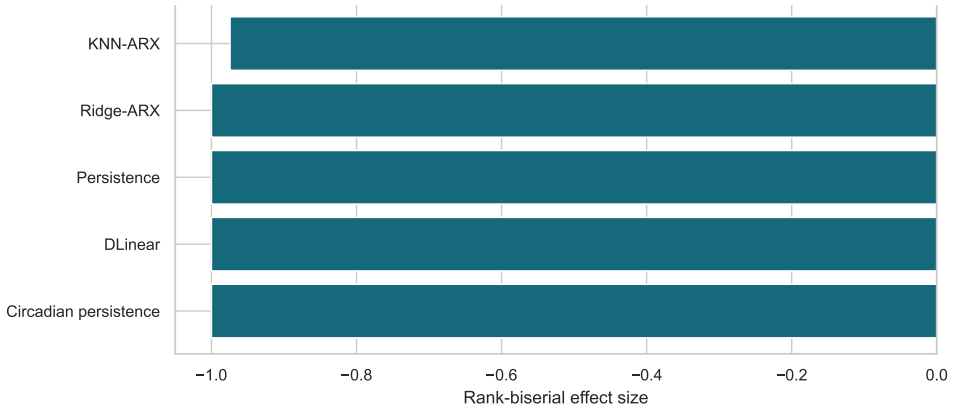

Fig. 4 Paired effect sizes (rank-biserial correlation) for GlucoKoopman-UQ against each comparator. Differences are GlucoKoopman-UQ minus comparator, so negative values favour GlucoKoopman-UQ

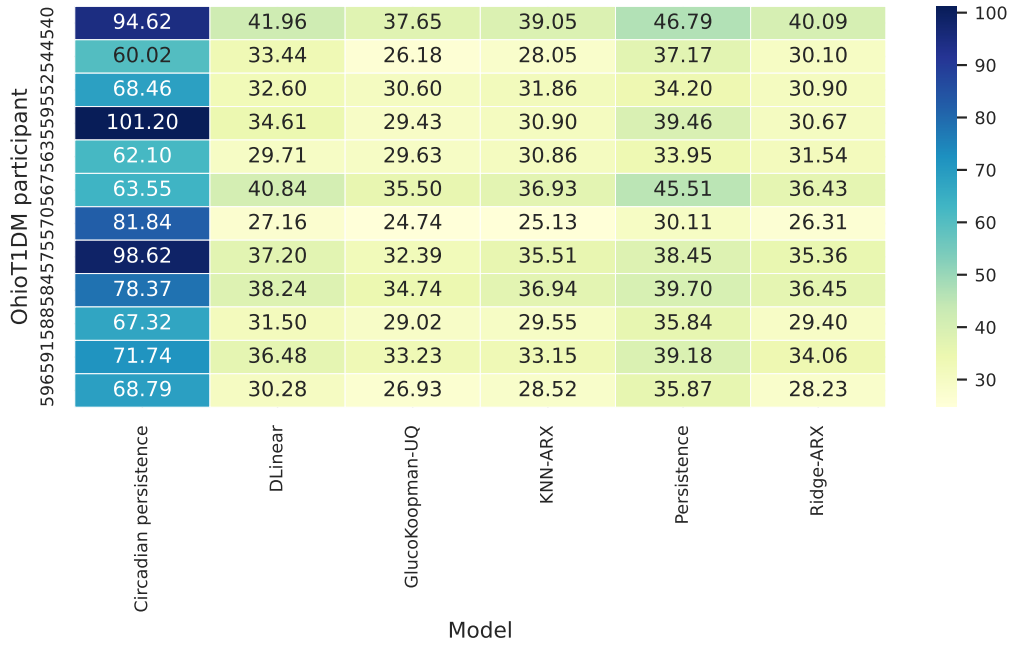

Fig. 5 RMSE (mg/dL) at 60 minutes for each participant (rows) and model (columns), after averaging the five seeds. Differences between participants are larger than most differences between models

4.3 Differences between participants and between seeds

At 60 minutes, RMSE ranged from 24.735 mg/dL for participant 570 to 37.646 mg/dL for participant 540 (Online Resource 2, Table S10). Varying the seed changed little: the within-participant standard deviation across seeds was generally below 1.2 mg/dL, far smaller than the spread between participants. Fig. 5 shows that both difficulty and model ranking varied from person to person. All five fits respected the stability constraint (Fig. 6).

4.4 Glycaemic risk

Hypoglycaemia became harder to anticipate as the horizon grew. Between 15 and 60 minutes, AUROC fell from 0.977 to 0.932, AUPRC from 0.475 to 0.273, and recall from 0.623 to 0.383. The high AUROC needs to be read against a prevalence of about 2.7%. Hyperglycaemia was much easier: at 60 minutes

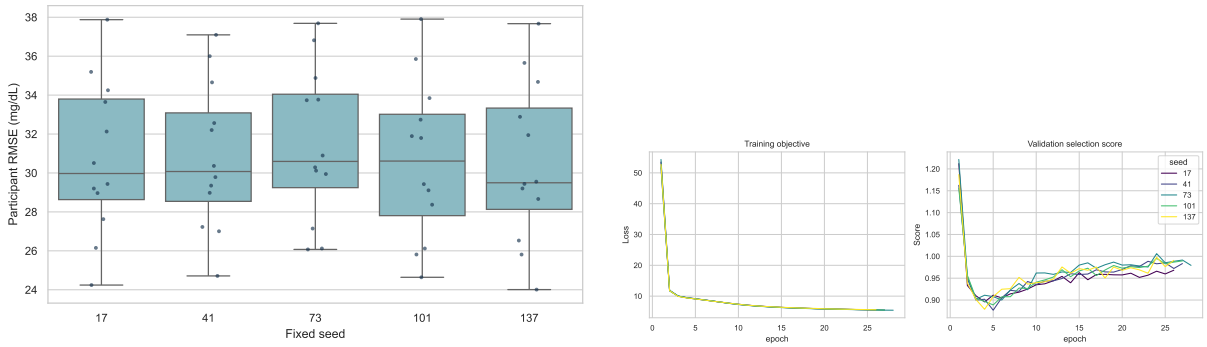

Fig. 6 Training behaviour. **a** Participant-level 60-minute RMSE for each seed. **b** Training loss and validation score by epoch. GlucoKoopman-UQ stopped early, whereas DLinear often ran close to the 160-epoch limit

Table 4 Hypo- and hyperglycaemia prediction by GlucoKoopman-UQ after temperature scaling

Horizon	Hypoglycaemia			Hyperglycaemia			Risk ECE	Multiclass Brier
(min)	AUROC	AUPRC	Recall	AUROC	AUPRC	Recall		
15	0.977	0.475	0.623	0.979	0.959	0.924	0.018	0.151
30	0.970	0.398	0.529	0.963	0.927	0.889	0.024	0.192
45	0.954	0.330	0.439	0.940	0.885	0.851	0.028	0.240
60	0.932	0.273	0.383	0.913	0.840	0.815	0.034	0.285

Classes: hypoglycaemia < 70, in range 70–180, hyperglycaemia > 180 mg/dL. ECE, expected calibration error. About 2.7% of windows were hypoglycaemic and 35.6% hyperglycaemic at each horizon.

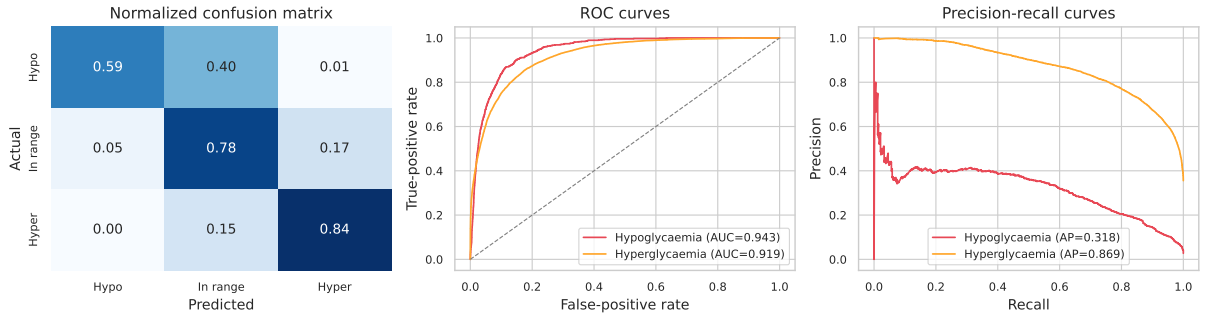

Fig. 7 Hypo- and hyperglycaemia discrimination at 60 minutes: confusion matrix, ROC curves and precision–recall curves. For the rare hypoglycaemia class the precision–recall curve is the more informative of the two

AUROC was 0.913, AUPRC 0.840 and recall 0.815. ECE rose from 0.018 to 0.034 and the Brier score from 0.151 to 0.285 with horizon (Table 4 and Figs. 7 and 8).

4.5 Prediction intervals

The nominal 90% intervals covered the truth slightly more often than required at every horizon: 0.925, 0.922, 0.920 and 0.915. The price was width. Mean interval width grew from 39.463 to 109.729 mg/dL and the Winkler score from 54.011 to 143.214 mg/dL (Table 5 and Fig. 9). Coverage was not uniform. At 60 minutes, averaged over participants, it was 0.946 for hypoglycaemic targets, 0.935 in range and 0.889 for hyperglycaemic targets, and it also varied between participants (Fig. 10).

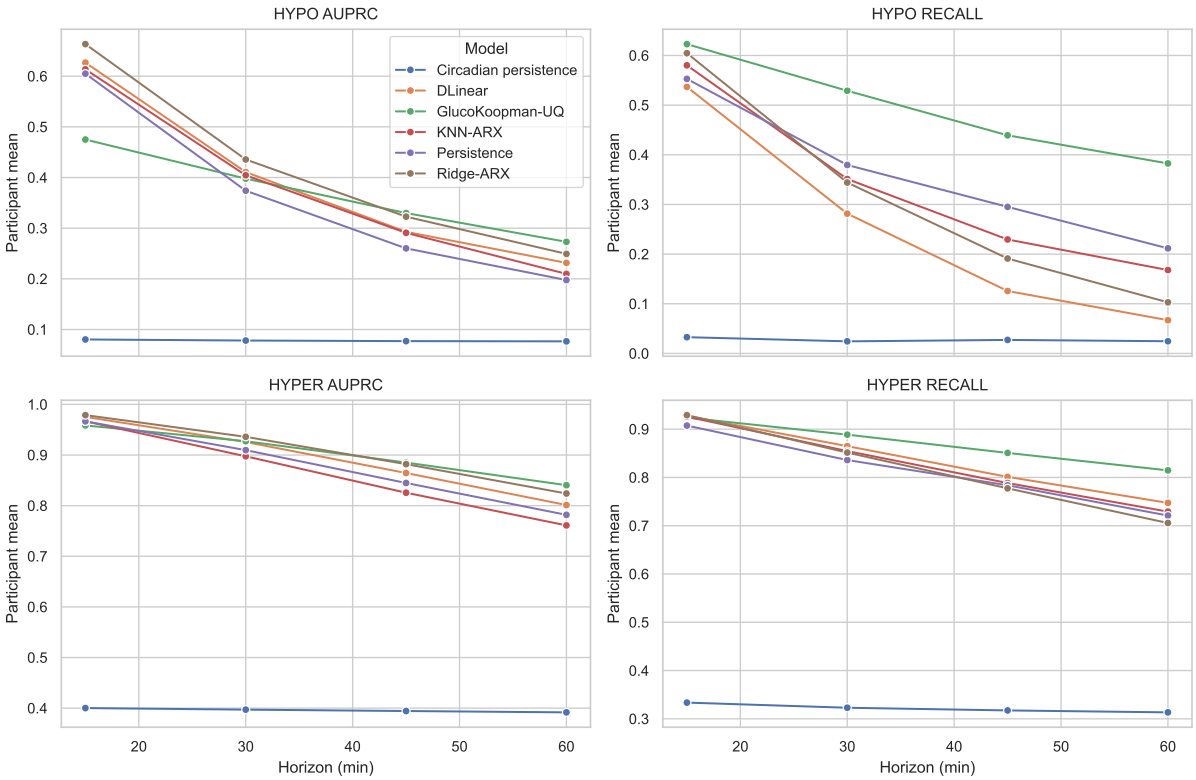

Fig. 8 AUPRC and recall for hypo- and hyperglycaemia by horizon and model

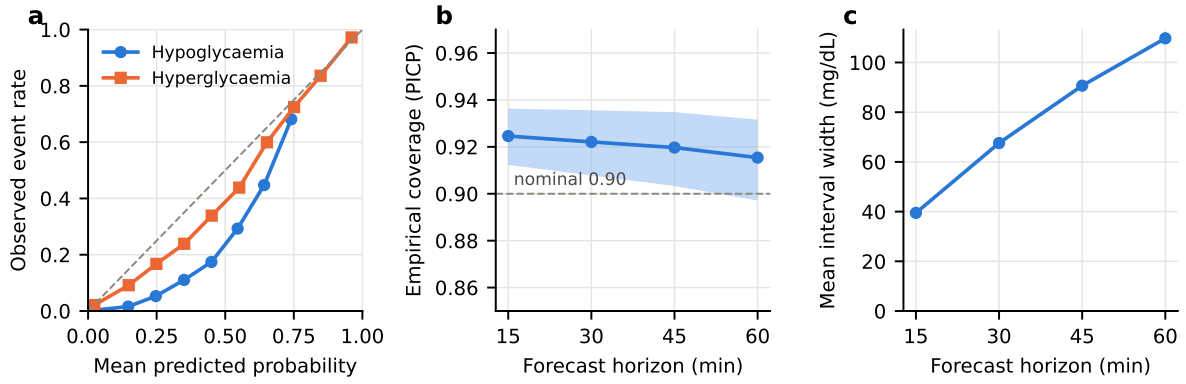

Fig. 9 Calibration of GlucoKoopman-UQ. **a** Observed event rate against predicted probability at 60 minutes, after temperature scaling; only bins holding 30 or more windows are drawn, and the dashed diagonal marks perfect agreement. **b** Share of test values inside the nominal 90% intervals, with a participant-bootstrap 95% band; the dashed line is 0.90. **c** Average interval width

Table 5 Coverage, width and Winkler score of the nominal 90% intervals

Horizon (min)	PICP	MPIW	Normalized MPIW	Winkler score
15	0.925	39.463	0.136	54.011
30	0.922	67.585	0.233	89.101
45	0.920	90.670	0.312	118.553
60	0.915	109.729	0.378	143.214

Widths and Winkler scores in mg/dL. For the same coverage, narrower is better; a PICP above 0.90 means the intervals were a little wider than needed.



Fig. 10 Interval coverage for each participant and horizon. Departures from the 0.90 target show why per-participant and per-range calibration is worth pursuing

4.6 Clinical agreement and large errors

Clarke zones A and B held 99.812%, 99.330%, 98.688% and 97.906% of forecasts from 15 to 60 minutes. Zone D rose from 0.184% to 1.917%, and zone E never exceeded 0.011%. Bland–Altman bias stayed small (-0.505 to 0.331 mg/dL), but the 95% limits of agreement widened from $[-21.706, 20.697]$ to $[-59.008, 59.671]$ mg/dL (Table 6). A small average bias therefore sits alongside individual errors that matter clinically. The residuals were heteroscedastic and heavy-tailed (Fig. 11). We exported the 80 largest errors at each horizon; the worst were 154.5, 162.4, 196.4 and 197.2 mg/dL from 15 to 60 minutes. The traces around them show sharp swings and reversals (Fig. 12), but the data cannot tell us what caused them.

4.7 Glucose range and cohort

Split by the glucose value being predicted, the 60-minute MAE was 21.241 mg/dL in range, 22.629 mg/dL below range and 25.679 mg/dL above range (Fig. 13). The six 2018 participants had a 60-minute RMSE of 29.739 mg/dL and the six 2020 participants 31.934 mg/dL, with coverage of 0.925 and 0.905 (Online

Table 6 Clarke error-grid zones and Bland–Altman agreement for GlucoKoopman-UQ

Horizon (min)	Zone A	Zones A+B	Zone C (%)	Zone D	Zone E	Bias	95% limits of agreement (mg/dL)
15	97.400	99.812	0.003	0.184	0.000	-0.505	[-21.706, 20.697]
30	89.643	99.330	0.009	0.659	0.002	-0.250	[-35.954, 35.455]
45	81.136	98.688	0.042	1.270	0.001	0.007	[-48.330, 48.345]
60	73.698	97.906	0.166	1.917	0.011	0.331	[-59.008, 59.671]

Clarke zones are averaged over participants. Bland–Altman statistics pool the 28,190 seed-averaged test windows per horizon and are descriptive only.

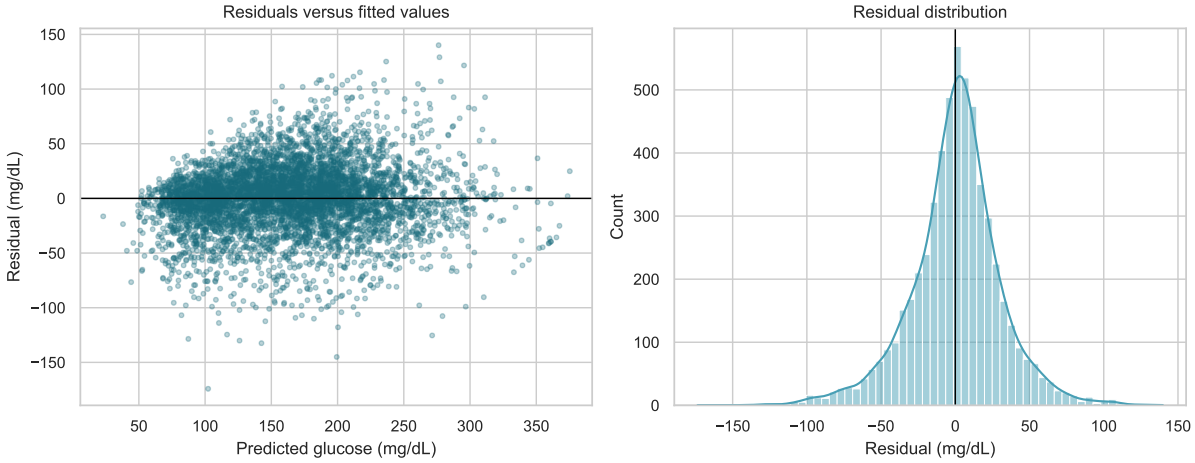

Fig. 11 Residuals against predicted glucose (left) and residual distributions by horizon (right). Spread grows with horizon and the tails are heavy, which mean bias alone does not reveal

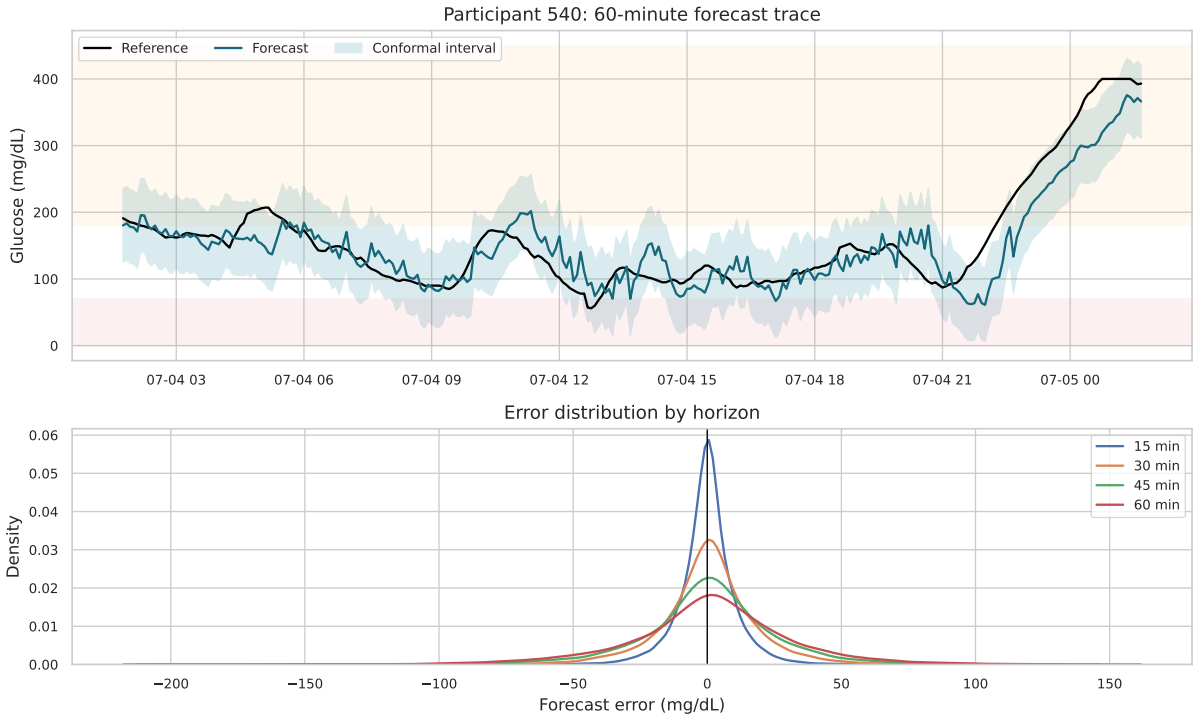

Fig. 12 Example 60-minute forecast trace with its interval for one participant (top), and error distributions by horizon (bottom). The trace is for illustration only

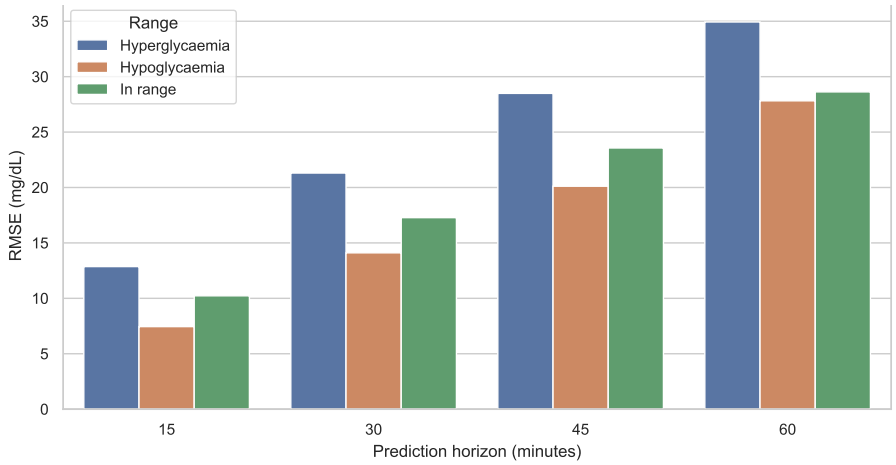

Fig. 13 GlucoKoopman-UQ error by target glucose range and horizon. Hyperglycaemic targets were hardest at 60 minutes

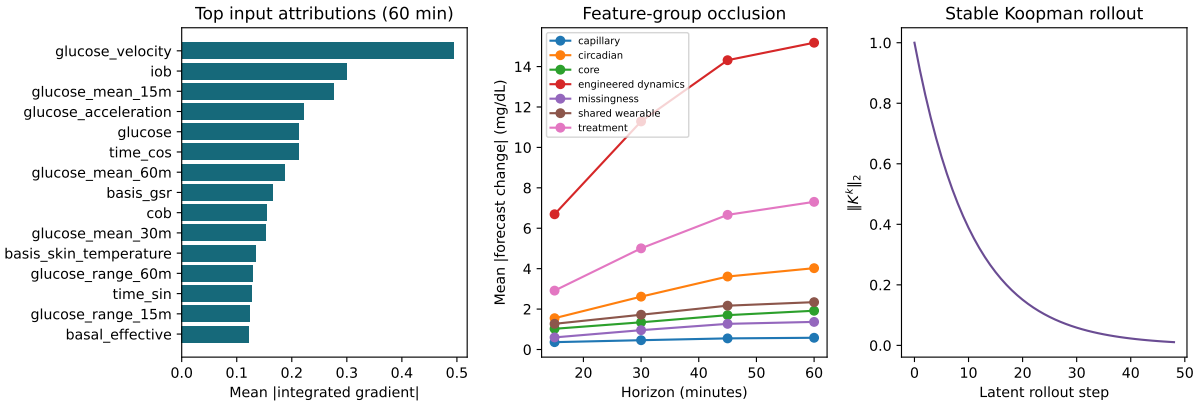

Fig. 14 Exploratory interpretability for the seed-17 model: input attributions, change in error when groups of inputs are occluded, and decay of the latent rollout. Results are for one seed and should not be read causally

Resource 2, Table S11). Groups of six are too small to treat as separate validation cohorts.

4.8 What the model relies on

For the seed-17 model, the inputs with the largest attributions were glucose velocity, IOB, the 15-minute trailing glucose mean, glucose acceleration and current glucose. Occluding groups of inputs changed the error most for engineered glucose dynamics and treatment features. The rollout plots show the control effect fading and the latent spectrum staying stable (Fig. 14). These are single-seed associations, not evidence of a physiological mechanism.

4.9 Ablations

All 65 ablation runs finished (Online Resource 2, Table S12). At 60 minutes the Friedman test was significant for the architecture suite ($Q_F = 18.36$, $p = 0.0054$) and more strongly for the feature suite ($Q_F = 39.00$, $p = 1.95 \times 10^{-6}$). Two changes clearly hurt. With CGM-only inputs, RMSE rose from 30.837 to 32.647 mg/dL (paired increase 1.810 mg/dL, adjusted $p = 0.0063$) and MAE from 22.592 to 23.913 mg/dL ($p = 0.0127$). Without treatment inputs, RMSE rose to 32.311 mg/dL (increase 1.474

Table 7 Effect of each ablation on 60-minute accuracy

Suite	Variant	RMSE	ΔRMSE (mg/dL)	Holm p	MAE (mg/dL)	Hypo AUPRC	Hyper AUPRC	PICP
Reference	**GlucoKoopman-UQ**	**30.837**	0.000	–	**22.592**	0.273	0.840	0.915
Architecture	Deterministic head	30.971	+0.135	0.5225	22.707	0.274	0.835	0.916
Architecture	No Koopman control	30.982	+0.145	1.0000	22.631	0.246	0.825	0.916
Architecture	No multiresolution	31.158	+0.322	0.9888	22.839	0.267	0.842	0.915
Architecture	No participant embedding	31.131	+0.294	0.9888	22.790	0.254	0.834	0.917
Architecture	No risk multitask objective	30.876	+0.040	1.0000	22.590	0.303	0.861	0.915
Architecture	Unconstrained dynamics	31.229	+0.393	0.1772	22.892	0.242	0.825	0.915
Feature	Core CGM only	32.647	**+1.810**	**0.0063**	23.913	0.272	0.830	0.914
Feature	No treatment inputs	32.311	**+1.474**	**0.0410**	23.713	0.258	0.828	0.913
Feature	No shared wearable inputs	30.691	−0.146	0.9888	22.525	0.282	0.846	0.915
Feature	Add weekday	30.931	+0.095	1.0000	22.644	0.245	0.837	0.915
Feature	Add behaviour	30.766	−0.070	1.0000	22.467	0.250	0.839	0.915
Feature	Add rare events	30.843	+0.007	1.0000	22.580	0.260	0.841	0.915
Feature	Add cohort-limited wearable inputs	30.725	−0.111	1.0000	22.424	0.247	0.842	0.915

Means over 12 participants after seed averaging. ΔRMSE is the variant minus the full model, so positive values mean lower accuracy. Holm p-values come from exact two-sided Wilcoxon tests corrected over the 13 variants. Confidence intervals are in Online Resource 1.

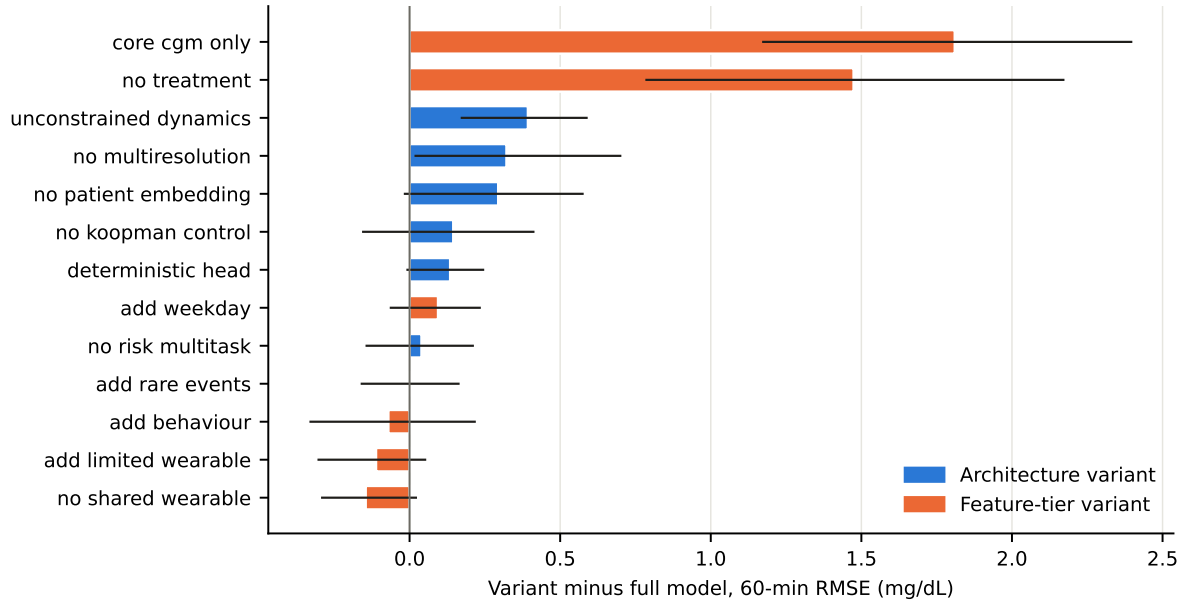

Fig. 15 Change in participant-level 60-minute RMSE for each ablation (variant minus full model; blue, architecture changes; orange, input changes). Positive values mean the variant was less accurate. Error bars are participant-bootstrap 95% confidence intervals; significance was tested separately with exact Wilcoxon tests and Holm correction over 13 variants

mg/dL, $p = 0.0410$) and MAE to 23.713 mg/dL ($p = 0.0293$) (Table 7 and Fig. 15).

No single architecture change significantly worsened 60-minute RMSE once corrected for 13 comparisons. The effects instead depended on horizon. The unconstrained operator lowered 15-minute RMSE by 0.500 mg/dL ($p = 0.0063$), but its 60-minute RMSE was 0.393 mg/dL higher (not significant after correction). Without the risk objective, 60-minute hyperglycaemia AUPRC was higher (0.861 versus 0.840; $p = 0.0063$), while hypoglycaemia recall was numerically lower (0.274 versus 0.383) but not significantly so ($p = 0.1206$). Adding weekday features lowered 60-minute hypoglycaemia AUPRC from 0.273 to 0.245 ($p = 0.0444$). The behavioural, rare-event and cohort-limited wearable channels did not improve 60-minute RMSE.

Table 8 Does each architecture component matter more at 60 minutes than at 15? (post hoc). For each variant, Δ_h is its mean participant-level RMSE minus that of the full model, in mg/dL; a positive value means the full model was better. The interaction column is $\Delta_{60} - \Delta_{15}$, tested over 12 participants with exact two-sided Wilcoxon tests and Holm-adjusted over the six variants. The last row averages Δ_h over all six

Removed or altered component	Δ_{15}	Δ_{60}	Interaction	Participants > 0	p	p_{Holm}
Unconstrained dynamics	-0.500	$+0.393$	$+0.893$	12/12	0.0005	0.0029
No participant embedding	-0.279	$+0.294$	$+0.574$	10/12	0.0034	0.0171
No multiresolution encoder	-0.054	$+0.322$	$+0.376$	10/12	0.0210	0.0840
Deterministic head	-0.057	$+0.135$	$+0.192$	9/12	0.0522	0.1567
No risk multitask objective	-0.126	$+0.040$	$+0.166$	10/12	0.0923	0.1846
No control injection	-0.087	$+0.145$	$+0.232$	8/12	0.1763	0.1846
Composite Δ_h at 15 / 30 / 45 / 60 min: -0.184 (1/12; $p = 0.0010$); -0.084 (6/12; $p = 0.3804$); $+0.047$ (8/12; $p = 0.1514$); $+0.221$ (10/12; $p = 0.0342$)						

4.10 Horizon dependence of the architecture (post hoc)

The interaction test asks whether a component's benefit grows with horizon, as the contraction bound predicts for the stability constraint. It did. Without the constraint, 15-minute RMSE was 0.500 mg/dL lower and 60-minute RMSE 0.393 mg/dL higher. The interaction of 0.893 mg/dL was positive in all 12 participants (exact $p = 0.00049$, adjusted $p = 0.0029$; Table 8 and Fig. 16). The unconstrained and constrained models crossed between 30 and 45 minutes. The participant embedding behaved the same way (interaction 0.574 mg/dL, 10 of 12 participants, adjusted $p = 0.0171$). The other four components all had positive interaction estimates, but none survived correction.

Averaged over the six architecture variants, removing any one component made 15-minute RMSE 0.184 mg/dL better (11 of 12 participants, $p = 0.0010$) and 60-minute RMSE 0.221 mg/dL worse (10 of 12, $p = 0.0342$; bootstrap 95% CI 0.056–0.361). The full model also had the best average 60-minute rank in the architecture suite (2.58 of 7), ahead of the variant without the risk objective (2.92); the unconstrained variant ranked last (5.67). Taken at face value, the architecture is mainly a bias towards long-horizon accuracy. Close to the forecast origin, more flexible variants that behave more like autoregression do a little better, which fits Ridge-ARX's small advantage at 15 minutes. Because these tests were defined after the planned ones, they need confirming in another cohort.

4.11 Computational cost

GlucoKoopman-UQ has 162,505 trainable parameters and needed 0.431 ms per window in our CPU timing test. DLinear, with 300 parameters, needed 0.0335 ms, and KNN-ARX 0.7465 ms (Online Resource 2, Table S15). These figures describe our software environment and are not a hardware benchmark.

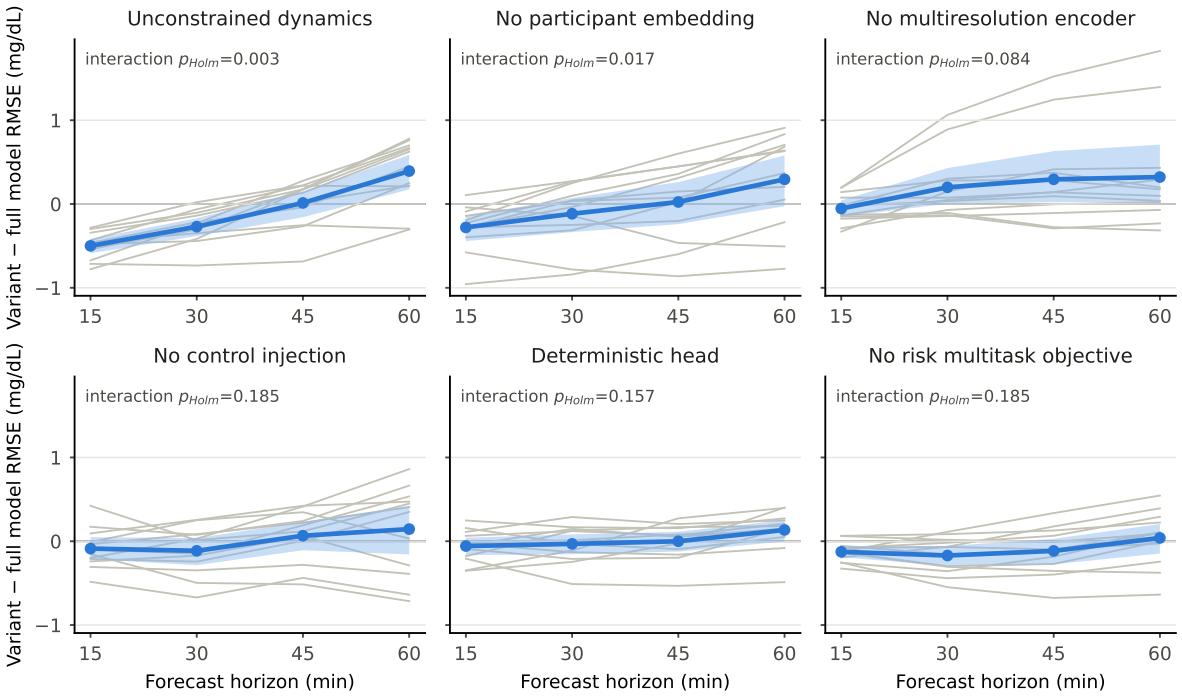

Fig. 16 RMSE of each architecture variant minus the full model, by horizon. Each grey line is one participant after seed averaging. The blue line is their mean, with a 95% band from resampling participants. A line rising above zero means the component matters more as the horizon lengthens. The p-value in each panel is for $\Delta_{60} - \Delta_{15}$ (exact Wilcoxon, Holm-adjusted over six panels)

5 Discussion

5.1 Main findings

Across 12 OhioT1DM participants and five seeds, GlucoKoopman-UQ had the lowest participant-level error from 30 to 60 minutes. Its 90% intervals were slightly conservative, and its hyperglycaemia probabilities separated the classes well. The 60-minute RMSE advantage was significant against all five comparators, but against Ridge-ARX and KNN-ARX it was modest, at 1.4–1.6 mg/dL. At 15 minutes, Ridge-ARX was as good or slightly better. We think this is the expected pattern rather than a weakness to explain away: over the next quarter of an hour glucose is highly autocorrelated, and a linear model already captures most of what can be predicted.

The more informative result is how the architecture's value shifts with horizon. Removing the stability constraint helped at 15 minutes and hurt at 60, in every participant. This is what Eq. (11) predicts. A contractive operator damps errors in the encoded state as it rolls forward, while an unconstrained one is free to fit the first few steps more closely but has no such damping. The participant embedding showed the same reversal, and the composite analysis points the same way. None of the six components on its own was indispensable at 60 minutes after correction over 13 variants. As a group, though, they push accuracy towards longer horizons, and their effect is additional to the larger gains from treatment and other non-CGM inputs.

5.2 Relation to other work

Earlier glucose studies report gains from deep sequence models, ensembles, personalisation, multitask learning and multimodal inputs [11, 12, 17, 25, 27, 31]. Comparing our numbers with theirs directly would be misleading, because participants, horizons, history lengths, target filters, pooled versus personal training, and leakage controls all differ. As one example, a Transformer comparison trained on a different cohort reported a 60-minute RMSE near 24.6 mg/dL with OhioT1DM as an external test set [22]. What we can offer is a single pipeline in which every model, including strong linear ones, is trained and scored the same way, with the participant as the unit of analysis.

The stability result also speaks to the Koopman literature. Azencot et al. argued that relaxing stability can help with data that have strong short-term dependence [49], and our 15-minute results agree. At 45–60 minutes the constrained operator was more accurate. The spectral radius stayed between 0.934 and 0.939, which rules out exploding trajectories but does not show that the latent dynamics are physiologically right.

5.3 Clinical reading of the uncertainty and risk outputs

The intervals change how a forecast should be read. Coverage stayed at or above 90%, but the average 60-minute interval was more than 100 mg/dL wide, nearly three times the 15-minute width. Coverage was also lower for hyperglycaemic targets and varied between people, so a single pooled coverage figure is not enough. The hypoglycaemia results carry a similar caution. An AUROC of 0.932 at 60 minutes looks strong, yet AUPRC was 0.273 and recall 0.383, because fewer than 3% of windows were hypoglycaemic. Systems meant for use should be judged on alarm burden, lead time, false alerts per day and decision-curve utility, not on window-level AUROC. Finally, Clarke zones A and B held over 97.9% of forecasts at every horizon, but error grids were designed for meters and can hide timing and calibration failures [59]. Errors approaching 200 mg/dL at the longest horizons make clear that this is a research forecaster and not a dosing tool.

5.4 Limitations

The study has several limitations. Only 12 participants from one public dataset were available, and the design tests later periods for people the model has already seen; participant embeddings may profit from that. LOPO and external validation are still needed. DLinear was the only neural comparator, so the size of the long-horizon advantage over recurrent, Transformer, N-BEATS or patch-based models [18–22] trained in the same pipeline remains to be measured. Wearable signals were patchy and some existed for only half the participants. Hypoglycaemia was rare, which limits what can be said about recall and rules out conclusions about severe events.

The validation segment served for model selection, early stopping and calibration, and neighbouring windows are correlated, so the formal split-conformal guarantee does not hold [39]. Splitting that segment into separate selection and calibration blocks, or using adaptive conformal inference [43], would address this. Targets outside 40–400 mg/dL were excluded. The ablations changed one thing at a time and cannot reveal interactions between components, and the horizon-interaction analysis was post hoc. The planned sensitivity runs were not carried out, and timings come from a CPU only. Retrospective accuracy does not show clinical benefit; that would need prospective, user-centred evaluation.

5.5 Future work

The next steps follow directly from these limitations: deep baselines in the same pipeline, LOPO and external validation, and a confirmatory test of the stability-by-horizon interaction in an independent cohort. Beyond that, we see value in event-level hypoglycaemia evaluation, calibration conditioned on participant and glucose range, and comparisons with personalised, federated, uncertainty-aware and foundation-model approaches under one causal pipeline [23, 28, 34, 38]. Calibration should be monitored over time, since intervals that were valid on past data can drift after changes in sensor, therapy or behaviour.

6 Conclusion

GlucoKoopman-UQ pairs a participant-conditioned multiresolution encoder with a controlled Koopman rollout whose operator is kept stable, and it adds temperature-scaled risk probabilities and residual-quantile intervals. On OhioT1DM it was the most accurate model from 30 to 60 minutes, with near-nominal interval coverage, but it did not beat Ridge-ARX at 15 minutes. Treatment and other non-CGM inputs were the largest single source of 60-minute accuracy. The stability constraint gave up a little accuracy at 15 minutes and gained it back at 60 in every participant, as the contraction bound predicts. Weak hypoglycaemia recall, wide long-horizon intervals, reuse of the validation data for calibration, and the lack of testing on new people all stand in the way of clinical claims. The model is ready for testing on new people and in prospective studies, but it should not drive insulin dosing.

Statements and Declarations

Funding No funding was received for this study.

Competing interests The authors have no relevant financial or non-financial interests to disclose.

Ethics approval Not applicable. This is a secondary analysis of the de-identified OhioT1DM dataset, obtained under its data use agreement. No new data were collected from human participants and no one was contacted, so ethical approval was not required. Ethical approval and consent for the original data collection were obtained by the original investigators [1].

Consent to participate Not applicable. The OhioT1DM participants gave their consent to the original study team [1].

Consent for publication Not applicable.

Data availability The OhioT1DM dataset is available to researchers from Ohio University on signing a data use agreement [1]. The authors may not redistribute the raw data. Aggregate participant-level results (CSV and JSON) behind every table and figure are provided as Online Resource 1, and additional tables as Online Resource 2.

Code availability The preprocessing, training, evaluation, statistical-testing and ablation code, with its fixed configuration and seeds, is available from the corresponding author on reasonable request. The scripts that validate the evidence and regenerate the post hoc analysis are included in Online Resource 1.

References

[1] Cindy Marling and Razvan Bunescu. The OhioT1DM dataset for blood glucose level prediction: Update 2020. *JMIR Diabetes*, 6(1):e16662, 2021. https://doi.org/10.2196/16662.

[2] Tadej Battelino, Charles M. Alexander, Stephanie A. Amiel, Guillermo Arreaza-Rubin, Roy W. Beck, Richard M. Bergenstal, Bruce A. Buckingham, James Carroll, Antonio Ceriello, Elaine Chow, Pratik Choudhary, Kelly Close, Thomas Danne, Sanjoy Dutta, Robert Gabbay, Satish Garg, Julie Heverly, Irl B. Hirsch, Tina Kader, Julia Kenney, Boris Kovatchev, Lori Laffel, David Maahs, Chantal Mathieu, Dídac Mauricio, Revital Nimri, Rimei Nishimura, Mauro Scharf, Stefano Del Prato, Eric Renard, Julio Rosenstock, Banshi Saboo, Kohjiro Ueki, Guillermo E. Umpierrez, Stuart A. Weinzimer, and Moshe Phillip. Continuous glucose monitoring and metrics for clinical trials: An international consensus statement. *The Lancet Diabetes & Endocrinology*, 11(1):42–57, 2023. https://doi.org/10.1016/S2213-8587(22)00319-9.

[3] Jinyu Xie and Qian Wang. Benchmarking machine learning algorithms on blood glucose prediction for type i diabetes in comparison with classical time-series models. *IEEE Transactions on Biomedical Engineering*, 67 (11):3101–3124, 2020. https://doi.org/10.1109/TBME.2020.2975959.

[4] Hoda Nemat, Heydar Khadem, Jackie Elliott, and Mohammed Benaissa. Data-driven blood glucose level prediction in type 1 diabetes: A comprehensive comparative analysis. *Scientific Reports*, 14(1):21863, 2024. https://doi.org/10.1038/s41598-024-70277-x.

[5] Petar Bevanda, Stefan Sosnowski, and Sandra Hirche. Koopman operator dynamical models: Learning, analysis and control. *Annual Reviews in Control*, 52:197–212, 2021. https://doi.org/10.1016/j.arcontrol.2021.09.002.

[6] Steven L. Brunton, Marko Budišić, Eurika Kaiser, and J. Nathan Kutz. Modern Koopman theory for dynamical systems. *SIAM Review*, 64(2):229–340, 2022. https://doi.org/10.1137/21M1401243.

[7] Gary S. Collins, Karel G. M. Moons, Paula Dhiman, Richard D. Riley, Andrew L. Beam, Ben Van Calster, Marzyeh Ghassemi, Xiaoxuan Liu, Johannes B. Reitsma, Maarten van Smeden, Anne-Laure Boulesteix, Jennifer C. Camaradou, Leo Anthony Celi, Spiros Denaxas, Alastair K. Denniston, Ben Glocker, Robert M. Golub, Hugh Harvey, Georg Heinze, Michael M. Hoffman, Andre Pascal Kengne, Elaine Lam, Noah Lee, Elizabeth W. Loder, Lena Maier-Hein, Bilal A. Mateen, Melissa D. McCradden, Lauren Oakden-Rayner, Johan Ordish, Richard Parnell, Sherri Rose, Karandeep Singh, Laure Wynants, and Patricia Logullo. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385:e078378, 2024. https://doi.org/10.1136/bmj-2023-078378.

[8] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[9] Silvia Oviedo, Josep Vehí, Remei Calm, and Joaquim Armengol. A review of personalized blood glucose prediction strategies for T1DM patients. *International Journal for Numerical Methods in Biomedical Engineering*, 33(6):e2833, 2017. https://doi.org/10.1002/cnm.2833.

[10] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11121–11128. AAAI Press, 2023. https://doi.org/10.1609/aaai.v37i9.26317.

[11] Kezhi Li, Chengyuan Liu, Taiyu Zhu, Pau Herrero, and Pantelis Georgiou. GluNet: A deep learning framework for accurate glucose forecasting. *IEEE Journal of Biomedical and Health Informatics*, 24(2):414–423, 2020. https://doi.org/10.1109/JBHI.2019.2931842.

[12] Taiyu Zhu, Kezhi Li, Jianwei Chen, Pau Herrero, and Pantelis Georgiou. Dilated recurrent neural networks for glucose forecasting in type 1 diabetes. *Journal of Healthcare Informatics Research*, 4(3):308–324, 2020. https://doi.org/10.1007/s41666-020-00068-2.

[13] Mehrdad Jaloli and Marzia Cescon. Long-term prediction of blood glucose levels in type 1 diabetes using a CNN–LSTM-based deep neural network. *Journal of Diabetes Science and Technology*, 17(6):1590–1601, 2023. https://doi.org/10.1177/19322968221092785.

[14] Ganjar Alfian, Muhammad Syafrudin, Muhammad Anshari, Filip Benes, Fransiskus Tatas Dwi Atmaji, Imam Fahrurrozi, Ahmad Fathan Hidayatullah, and Jongtae Rhee. Blood glucose prediction model for type 1 diabetes based on artificial neural network with time-domain features. *Biocybernetics and Biomedical Engineering*, 40 (4):1586–1599, 2020. https://doi.org/10.1016/j.bbe.2020.10.004.

[15] Hatim Butt, Ikramullah Khosa, and Muhammad Asif Iftikhar. Feature transformation for efficient blood glucose prediction in type 1 diabetes mellitus patients. *Diagnostics*, 13(3):340, 2023. https://doi.org/10.3390/diagnostics13030340.

[16] Mahmoud H. Osman, Mennatullah Mahmoud, Salma Zakzouk, Samah Mohamed, Ibrahim Gomaa, M. Saeed Darweesh, Sayed Taha, and Ahmed Soltan. Enhanced glucose forecasting using recurrent neural network and advanced feature engineering. *Scientific Reports*, 16:12036, 2026. https://doi.org/10.1038/s41598-026-41066-5.

[17] Hoda Nemat, Heydar Khadem, Mohammad R. Eissa, Jackie Elliott, and Mohammed Benaissa. Blood glucose level prediction: Advanced deep-ensemble learning approach. *IEEE Journal of Biomedical and Health Informatics*, 26(6):2758–2769, 2022. https://doi.org/10.1109/JBHI.2022.3144870.

[18] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. https://doi.org/10.1162/neco.1997.9.8.1735.

[19] Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*. OpenReview.net, 2020. URL https://openreview.net/forum?id=r1ecqn4YwB.

[20] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*. OpenReview.net, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

[21] Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021. https://doi.org/10.1016/j.ijforecast.2021.03.012.

[22] Meryem Altin Karagoz, Marc D. Breton, and Anas El Fathi. A comparative study of transformer-based models for multi-horizon blood glucose prediction. In *1st IFAC Workshop on Engineering Diabetes Technologies (EDT 2025)*. International Federation of Automatic Control, 2025. URL https://arxiv.org/abs/2505.08821. arXiv:2505.08821.

[23] Bowen Zhang, Hsiu-Wen Cheng, Hongyu Yang, Evie L. Shen, Joleen Vansomphone, Yuna Li, Kerry Zhou, Zitian Qu, Suning Zhao, Xiangning Deng, Hua Zhou, and Jin J. Zhou. Evaluating time-series foundation models and multimodal dietary context for CGM forecasting, 2026. URL https://arxiv.org/abs/2609.11872. arXiv:2609.11872.

[24] Wonju Seo, Sung-Woon Park, Namho Kim, Sang-Man Jin, and Sung-Min Park. A personalized blood glucose level prediction model with a fine-tuning strategy: A proof-of-concept study. *Computer Methods and Programs in Biomedicine*, 211:106424, 2021. https://doi.org/10.1016/j.cmpb.2021.106424.

[25] John Daniels, Pau Herrero, and Pantelis Georgiou. A multitask learning approach to personalized blood glucose prediction. *IEEE Journal of Biomedical and Health Informatics*, 26(1):436–445, 2022. https://doi.org/10.1109/JBHI.2021.3100558.

[26] Md Maruf Hossain Shuvo and Syed Kamrul Islam. Deep multitask learning by stacked long short-term memory for predicting personalized blood glucose concentration. *IEEE Journal of Biomedical and Health Informatics*, 27(3):1612–1623, 2023. https://doi.org/10.1109/JBHI.2022.3233486.

[27] Taiyu Zhu, Kezhi Li, Pau Herrero, and Pantelis Georgiou. Personalized blood glucose prediction for type 1 diabetes using evidential deep learning and meta-learning. *IEEE Transactions on Biomedical Engineering*, 70 (1):193–204, 2023. https://doi.org/10.1109/TBME.2022.3187703.

[28] Yiheng Shen and Samantha Kleinberg. Personalized blood glucose forecasting from limited CGM data using incrementally retrained LSTM. *IEEE Transactions on Biomedical Engineering*, 72(4):1266–1277, 2025. https://doi.org/10.1109/TBME.2024.3494732.

[29] Kihoon Moon, Jaehong Kim, Seohyun Yoo, and Jaehyuk Cho. Personalized blood glucose prediction in type 1 diabetes using meta-learning with bidirectional long short-term memory–transformer hybrid model. *Scientific Reports*, 15:30636, 2025. https://doi.org/10.1038/s41598-025-13491-5.

[30] Minjoo Hwang, Vega Pradana Rachim, Junyoung Yoo, Yein Lee, and Sung-Min Park. Generalized multi-task learning framework for glucose forecasting and hypoglycemia detection using simulation to reality. *npj Digital Medicine*, 8:612, 2025. https://doi.org/10.1038/s41746-025-01994-4.

[31] Sarmad Maqsood, Muhammad Abdullah Sarwar, Egle Belousovienė, and Rytis Maskeliūnas. GlucoNet-MM: A multimodal attention-based multi-task learning framework with decision transformer for personalised and explainable blood glucose forecasting. *Diabetes, Obesity and Metabolism*, 27(12):7416–7430, 2025. https://doi.org/10.1111/dom.70147.

[32] Taiyu Zhu, Chukwuma Uduku, Kezhi Li, Pau Herrero, Nick Oliver, and Pantelis Georgiou. Enhancing self-management in type 1 diabetes with wearables and deep learning. *npj Digital Medicine*, 5:78, 2022. https://doi.org/10.1038/s41746-022-00626-5.

[33] Taiyu Zhu, Lei Kuang, John Daniels, Pau Herrero, Kezhi Li, and Pantelis Georgiou. IoMT-enabled real-time blood glucose prediction with deep learning and edge computing. *IEEE Internet of Things Journal*, 10(5): 3706–3719, 2023. https://doi.org/10.1109/JIOT.2022.3143375.

[34] Dave Darpit, Kathan Vyas, Jagadish Kumaran Jayagopal, Alfredo Garcia, Madhav Erraguntla, and Mark Lawley. A personalized federated learning-based glucose prediction algorithm for high-risk glycemic excursion regions in type 1 diabetes. *Scientific Reports*, 15:38376, 2025. https://doi.org/10.1038/s41598-025-22316-4.

[35] Chiara Dalla Man, Francesco Micheletto, Dayu Lv, Marc Breton, Boris Kovatchev, and Claudio Cobelli. The UVA/PADOVA type 1 diabetes simulator: New features. *Journal of Diabetes Science and Technology*, 8(1): 26–34, 2014. https://doi.org/10.1177/1932296813514502.

[36] John Martinsson, Alexander Schliep, Björn Eliasson, and Olof Mogren. Blood glucose prediction with variance estimation using recurrent neural networks. *Journal of Healthcare Informatics Research*, 4(1):1–18, 2020. https://doi.org/10.1007/s41666-019-00059-y.

[37] Saúl Langarica, Maria Rodriguez-Fernandez, Francis J. Doyle, and Felipe Núñez. A probabilistic approach to blood glucose prediction in type 1 diabetes under meal uncertainties. *IEEE Journal of Biomedical and Health Informatics*, 27(10):5054–5065, 2023. https://doi.org/10.1109/JBHI.2023.3309302.

[38] Antonio J. Rodriguez-Almeida, Carmelo Betancort, Ana M. Wägner, Gustavo M. Callico, Himar Fabelo, Conceição Bartnæs, Eirik Årsand, and Miroslav Muzny. Incorporating uncertainty estimation and interpretability in personalized glucose prediction using the temporal fusion transformer. *Sensors*, 25(15):4647, 2025. https://doi.org/10.3390/s25154647.

[39] Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. https://doi.org/10.1561/2200000101.

[40] Kamilė Stankevičiūtė, Ahmed M. Alaa, and Mihaela van der Schaar. Conformal time-series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, pages 6216–6228. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/312f1ba2a72318edaaa995a67835fad5-Abstract.html.

[41] Chen Xu and Yao Xie. Sequential predictive conformal inference for time series. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 38707–38727. PMLR, 2023. URL https://proceedings.mlr.press/v202/xu23r.html.

[42] Roberto I. Oliveira, Paulo Orenstein, Thiago Ramos, and João Vitor Romano. Split conformal prediction and non-exchangeable data. *Journal of Machine Learning Research*, 25(225):1–38, 2024. URL https://jmlr.org/papers/v25/23-1553.html.

[43] Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/0d441de75945e5acbc865406fc9a2559-Abstract.html.

[44] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3):e0118432, 2015. https://doi.org/10.1371/journal.pone.0118432.

[45] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 2017.

[46] Igor Mezić. Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dynamics*, 41:309–325, 2005. https://doi.org/10.1007/s11071-005-2824-x.

[47] Naoya Takeishi, Yoshinobu Kawahara, and Takehisa Yairi. Learning Koopman invariant subspaces for dynamic mode decomposition. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://papers.nips.cc/paper/2017/hash/3a835d3215755c435ef4fe9965a3f2a0-Abstract.html.

[48] Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9:4950, 2018. https://doi.org/10.1038/s41467-018-07210-0.

[49] Omri Azencot, N. Benjamin Erichson, Vanessa Lin, and Michael W. Mahoney. Forecasting sequential data using consistent Koopman autoencoders. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*. PMLR, 2020. URL https://arxiv.org/abs/2003.02236.

[50] Yong Liu, Chenyu Li, Jianmin Wang, and Mingsheng Long. Koopa: Learning non-stationary time series dynamics with Koopman predictors. In *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., 2023. URL https://arxiv.org/abs/2305.18803.

[51] Joshua L. Proctor, Steven L. Brunton, and J. Nathan Kutz. Dynamic mode decomposition with control. *SIAM Journal on Applied Dynamical Systems*, 15(1):142–161, 2016. https://doi.org/10.1137/15M1013857.

[52] Milan Korda and Igor Mezić. Linear predictors for nonlinear dynamical systems: Koopman operator meets model predictive control. *Automatica*, 93:149–160, 2018. https://doi.org/10.1016/j.automatica.2018.03.046.

[53] Meixi Wang, Xuyang Lou, Wei Wu, and Baotong Cui. Koopman-based model predictive control with learned dynamics: Hierarchical neural network approach. *IEEE Transactions on Neural Networks and Learning Systems*, 35(3):3630–3639, 2024. https://doi.org/10.1109/TNNLS.2022.3194958.

[54] Ali Tavasoli and Heman Shakeri. Online meal detection based on CGM data dynamics. In *IFAC Diabetes Technology Conference*, Valencia, Spain, 2025. International Federation of Automatic Control. URL https://arxiv.org/abs/2507.00080. arXiv:2507.00080.

[55] Tadej Battelino, Thomas Danne, Richard M. Bergenstal, Stephanie A. Amiel, Roy Beck, Torben Biester, et al. Clinical targets for continuous glucose monitoring data interpretation: Recommendations from the international consensus on time in range. *Diabetes Care*, 42(8):1593–1603, 2019. https://doi.org/10.2337/dci19-0028.

[56] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 2nd edition, 2013. https://doi.org/10.1017/CBO9781139020411.

[57] William L. Clarke, Daniel Cox, Linda A. Gonder-Frederick, William Carter, and Stephen L. Pohl. Evaluating clinical accuracy of systems for self-monitoring of blood glucose. *Diabetes Care*, 10(5):622–628, 1987. https://doi.org/10.2337/diacare.10.5.622.

[58] J. Martin Bland and Douglas G. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476):307–310, 1986. https://doi.org/10.1016/S0140-6736(86)90837-8.

[59] Tiange Tian, Rebecca E. Aaron, Michael A. Kohn, and David C. Klonoff. The need for a modern error grid for clinical accuracy of blood glucose monitors and continuous glucose monitors. *Journal of Diabetes Science and Technology*, 18(1):3–9, 2024. https://doi.org/10.1177/19322968231214281.

[60] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- OnlineResource2SupplementaryTables.pdf