A STATE-RESOLVED SYNTHETIC DATASET FOR BILATERAL FLAPPING-WING AERODYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Flapping-wing flight involves strongly coupled interactions among rapid bilateral wing motion, body dynamics, and unsteady aerodynamic responses. Physical data that reveal how these processes interact are essential for data-driven aerodynamic modeling, system identification, state estimation, reduced-order dynamics, control-oriented analysis, and scientific machine learning. We introduce FlapKAD-State, a state-resolved synthetic dataset comprising 2,000 curated rigid-wing flight episodes and 720,152 valid time steps generated using synchronized rigid-body integration and quasi-steady aerodynamic modeling. Our dataset provides comprehensive bilateral flap and wing-rotation kinematics, together with their corresponding angular velocities, local body linear velocity, body angular velocity, and independently resolved aerodynamic forces for the left and right wings. These variables provide complementary descriptions of bilateral wing kinematics, body motion, and aerodynamic response within a unified episode structure. To facilitate the use and evaluation of the dataset, we provide standardized episode-level partitions, variable-length masks, channel definitions, physical units, coordinate conventions, and data-quality documentation, together with reference benchmarks that illustrate representative learning scenarios.

1 INTRODUCTION

Flapping-wing aerodynamics arise from tightly unsteady interactions among wing motion, aerodynamic loading, and body dynamics. Accurate models of these interactions are essential for structure design, system identification, and control, yet collecting synchronized physical measurements remains challenges. Real-world experimental approaches are suffered due to sensor limitations and costly to scale, while analytical models often simplify the aerodynamic memory and coupling present in flapping motion. Data-driven modeling offers a complementary approach, but its progress depends on datasets that expose the relevant physical variables within consistent trajectories. Existing flapping-wing datasets typically support individual tasks, such as mapping wing kinematics to aerodynamic forces, reconstructing motion from force measurements, or estimating vehicle state from onboard sensors. However, these settings do not provide a unified interface for studying how wing motion, body state, and bilateral aerodynamic forces interact. As a result, determining whether physical states improve aerodynamic prediction, or whether explicit aerodynamic-force observations improve the prediction of future motion are still uncertain.

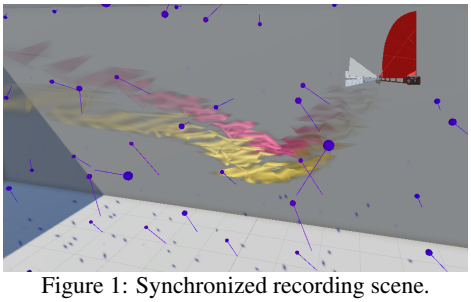

Figure 1: Synchronized recording scene.

We introduce FlapKAD-State, a state-oriented dataset for studying flapping-wing aerodynamics and coupled wing-force-body dynamics. Our dataset contains 2,000 curated variable-length episodes and 720,152 valid time steps recorded at 50 Hz. Each episode synchronizes bilateral wing angles and angular velocities, local body velocity, world-frame body angular velocity, and left-right vertical aerodynamic forces. Our unified physical interface supports controlled investigation of wing-force mapping, state-conditioned aerodynamic response, and future body-motion prediction within the same episode-level partition. FlapKAD-State supports three connected benchmark families. The first evaluates bidirectional mappings between bilateral wing kinematics and aerodynamic forces.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

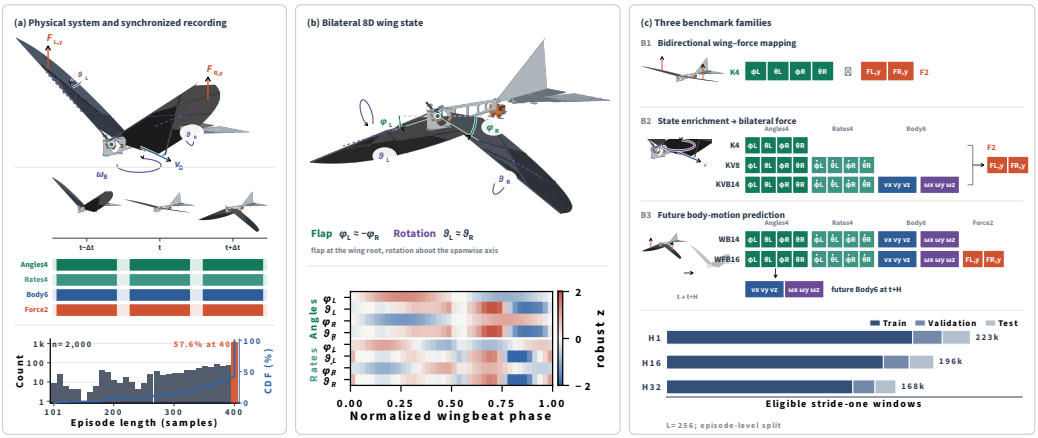

Figure 2: Overview of FlapKAD-State. **(a)** The synchronously records bilateral wing angles and angular rates, Body motion, and left- and right-wing vertical aerodynamic forces at 50 Hz. The episode-length distribution summarizes the variable-duration collection. **(b)** The bilateral wing representation contains four angles and their four angular rates. The phase-conditioned atlas summarizes the resulting eight-dimensional wing state over the wingbeat cycle. **(c)** The three benchmark families evaluate bidirectional wing-force mapping, state enrichment for bilateral-force prediction, and future Body prediction from matched state-only and force-conditioned inputs.

The second evaluates state-conditioned aerodynamic response using progressively wing and body states. The third evaluates coupled wing-force-body dynamics by testing whether bilateral aerodynamic forces provide incremental information for predicting body motion. Our contributions are:

- A curated collection of 2,000 variable-length flapping-wing episodes with synchronized bilateral wing motion, body dynamics, and aerodynamic force.
- A synchronized 14-dimensional wing-body state and bilateral aerodynamic-force interface containing 720,152 valid records at 50 Hz.
- Three connected benchmark families covering bidirectional wing-force mapping, state-conditioned aerodynamic response, and coupled wing-force-body dynamics.
- A unified multi-horizon evaluation protocol with reference models spanning linear, recurrent, temporal-mixing, sparse system-identification, and continuous-time approaches.

2 Related Work

Flapping-wing aerodynamic modeling. Flapping-wing loading arises from coupled translational, rotational, added-mass, wake-history, aeroelastic, and body-motion effects, producing phase-dependent responses that are difficult to represent using steady or memoryless formulations Dickinson et al. (1999); Sane & Dickinson (2002); Sane (2003); Shyy et al. (2010). System-identification, recurrent state-space, neural-surrogate, reduced-order, and data-driven quasi-steady methods have consequently been developed for aerodynamic prediction, model discovery, and kinematic optimization Bayiz & Cheng (2021b); Pereira et al. (2023); Lan et al. (2022); Corban et al. (2023); Kamimizu et al. (2025). These studies establish the importance of physical context, but typically organize their data around a particular platform, flow regime, or predefined modeling direction.

Open flapping-wing data resources. Open flapping-wing resources span several complementary physical scales. Force-resolved robotic-wing datasets provide synchronized kinematics, forces, and moments for forward state-space modeling and inverse wing-motion reconstruction Bayiz & Cheng (2021a;b); Sharvit et al. (2025). Flow-resolved experiments provide PIV fields for robotic bird-wing kinematic sweeps, petiolate-wing vortex formation, and deformable membrane-wing aerodynamics Johansson (2024); Phillips et al. (2017); Gehrke & Mulleners (2025). Other datasets characterize resonant spring-wing energetics, stability, responsiveness, and fluid-disturbance sensitivity

Lynch et al. (2021). Animal-flight resources contribute synchronized aerodynamic force, kinematic, acoustic, morphological, and body-oscillation measurements Hightower et al. (2021); Sikandar et al. (2025); Engels et al. (2020), while robotic-flight datasets provide high-rate pose, control, visual, inertial, and aggressive-maneuver records Kim et al. (2025); Hsiao et al. (2025). These resources are valuable but expose different subsets of flow, wing, force, body, or sensing variables. FlapKAD-State complements them by preserving bilateral wing motion, wing rates, body motion, and bilateral aerodynamic force in the same 2,000 episodes, enabling matched bidirectional, state-enrichment, and coupled-dynamics comparisons without merging incompatible platforms.

Temporal forecasting and dynamical-system learning. The evaluated references cover complementary temporal inductive biases: recurrent and temporal-convolutional modeling Cho et al. (2014); Bai et al. (2018), linear projection and modern forecasting architectures Zeng et al. (2023); Wu et al. (2021); Nie et al. (2023); Wu et al. (2023); Liu et al. (2024), temporal–channel mixing and exogenous forecasting Chen et al. (2023); Wang et al. (2024), and sparse or continuous-time dynamics Brunton et al. (2016); de Silva et al. (2020); Chen et al. (2018). FlapKAD-State uses these model families not to introduce another forecasting architecture, but to examine how physical information, prediction horizon, and dynamical assumptions affect force and body-motion prediction.

3 Methodology

3.1 Data Collection and State Representation

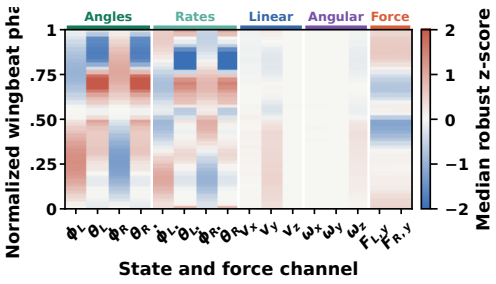

Figure 3: Phase-resolved state and force atlas.

FlapKAD-State was collected using a synchronized fixed-step and recording pipeline that integrates bilateral wing actuation, aerodynamic-force estimation, rigid-body dynamics, and episode-level logging. At each fixed step, the pipeline records the bilateral wing motion, the corresponding wing-motion rates, the rigid-body motion, and the independently evaluated aerodynamic load of each wing. This unified process preserves sample-level temporal alignment across wing, body, and force variables. The released state and aerodynamic-target vectors are:

$$\mathbf{s}(t) = \left[\mathbf{q}(t)^\mathsf{T}, \dot{\mathbf{q}}(t)^\mathsf{T}, \mathbf{v}_{\text{local}}(t)^\mathsf{T}, \boldsymbol{\omega}_{\text{world}}(t)^\mathsf{T}\right]^\mathsf{T} \in \mathbb{R}^{14}, \qquad \mathbf{F}(t) = \left[F_{L,y}(t), F_{R,y}(t)\right]^\mathsf{T} \in \mathbb{R}^2. \quad (1)$$

Where, $\mathbf{q}(t) \in \mathbb{R}^4$ contains the left and right flap and twist angles, and $\dot{\mathbf{q}}(t) \in \mathbb{R}^4$ contains their angular velocities. The body-motion variables comprise the local-frame linear velocity $\mathbf{v}_{\text{local}}(t) \in \mathbb{R}^3$ and the world-frame angular velocity $\boldsymbol{\omega}_{\text{world}}(t) \in \mathbb{R}^3$. The target $\mathbf{F}(t)$ contains the independently evaluated left- and right-wing aerodynamic-force. Detailed aerodynamic definitions, coordinate conventions, and derived force quantities are provided in Appendix B. Figure 3 illustrates the interface over a normalized wingbeat cycle. The phase-resolved representation places bilateral wing angles, wing-motion rates, body velocities, angular velocities, and aerodynamic-force channels on a common temporal coordinate, exposing their synchronized periodic structure without aggregating the bilateral force targets. Figure 5 verifies the physical consistency of the bilateral force interface.

3.2 Variable-Length Representation and Statistics

FlapKAD-State contains 720,152 valid records distributed across 2,000 variable-length episodes sampled at 50 Hz. Episode boundaries are preserved throughout data conversion, normalization, temporal-window construction, and evaluation. Sequences are padded only for batch processing, and a binary validity mask excludes all padded positions from normalization, learning, and metric computation. The predefined dataset split is applied at the episode level, preventing windows from the same episode from appearing in multiple splits. All state and force channels are finite and exhibit nonzero variance over the valid records. Episode lengths range from 100 to 400 samples, with a median of 400, corresponding to approximately 2–8 s of synchronized observation. The body angular-velocity channels are finite over all valid records and retain nonzero temporal variance in each episode. Figure 4 summarizes the empirical distributions of the 14 state channels and

two bilateral force channels in their original physical units. The grouped statistics expose their distinct numerical scales, while the duration panel characterizes the variable-length episode collection. Additional episode-level distributions and quality-control details are reported in Appendix C.

3.3 BENCHMARK FAMILIES

FlapKAD-State preserves synchronized episode-level trajectories rather than distributing task-specific windows. The same physical records can therefore support multiple learning tasks by selecting different input and target variables. We define three connected benchmark families.

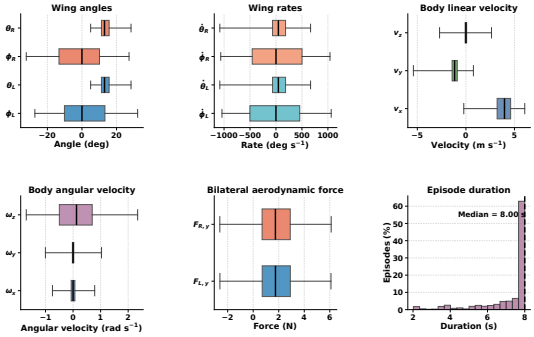

Figure 4: State, force, and episode-duration distributions. Boxes show interquartile ranges, medians, and 1st–99th percentile whiskers.

Benchmark 1: Bidirectional Wing-Force Mapping The first benchmark retains the bilateral correspondence between wing kinematics and aerodynamic force. Let $\mathbf{K}_4(t) = \mathbf{q}(t) \in \mathbb{R}^4$ denote the bilateral wing-angle state, and let $\mathbf{F}(t) \in \mathbb{R}^2$ denote the bilateral vertical-force target defined in Eq. equation 1. The forward and inverse tasks are

$$\widehat{\mathbf{F}}(t+H) = \mathcal{F}_{\boldsymbol{\theta}}(\mathbf{K}_{4,t-L+1:t}), \qquad \widehat{\mathbf{K}}_4(t+H) = \mathcal{G}_{\boldsymbol{\psi}}(\mathbf{F}_{t-L+1:t}). \tag{2}$$

where, $\mathcal{F}_{\boldsymbol{\theta}}$ and $\mathcal{G}_{\boldsymbol{\psi}}$ denote the learnable forward and inverse models, respectively. Both tasks retain the complete left–right wing interface and use the same physical episodes, temporal configuration, and episode-level data partition.

Benchmark 2: State-Enriched Force Prediction The second benchmark evaluates whether wing-motion rates and body motion provide predictive information beyond bilateral wing-angle history. We consider three nested input states and a common forecasting task:

$$\begin{gathered}\mathbf{z}_1(t) = \mathbf{q}(t) \in \mathbb{R}^4, \quad \mathbf{z}_2(t) = [\mathbf{q}^\top(t), \dot{\mathbf{q}}^\top(t)]^\top \in \mathbb{R}^8, \quad \mathbf{z}_3(t) = \mathbf{s}(t) \in \mathbb{R}^{14}; \\ \widehat{\mathbf{F}}(t+H) = \mathcal{A}_{\boldsymbol{\eta}}(\mathbf{z}_{j,t-L+1:t}), \qquad j \in \{1,2,3\}.\end{gathered} \tag{3}$$

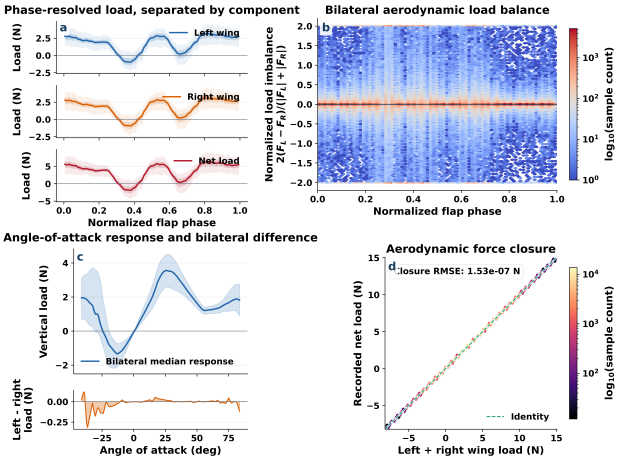

Figure 5: Aerodynamic-response, bilateral-load balance, and force-closure analysis.

Here, $\mathbf{q}(t)$ contains the realized bilateral flap and twist angles, $\dot{\mathbf{q}}(t)$ contains their angular velocities, and $\mathbf{s}(t)$ is the complete wing–body state defined in Eq. equation 1. Accordingly, $\mathbf{z}_1$ contains wing angles only, $\mathbf{z}_2$ additionally includes wing-motion rates, and $\mathbf{z}_3$ further includes body linear and angular motion. The history $\mathbf{z}_{j,t-L+1:t} = [\mathbf{z}_j(t-L+1), \ldots, \mathbf{z}_j(t)]$ contains the L most recent observations, while H is the forecast horizon. The target $\mathbf{F}(t+H) = [F_{L,y}(t+H), F_{R,y}(t+H)]^\top$ contains the future bilateral aerodynamic forces; $\widehat{\mathbf{F}}(t+H)$ denotes their prediction by the learnable temporal model $\mathcal{A}_{\boldsymbol{\eta}}$ with trainable parameters $\boldsymbol{\eta}$. The two nested comparisons isolate the incremental predictive value of wing-motion rates and body motion, respectively. Formal configuration names and channel indices are provided in Appendix D.3.

Benchmark 3: Coupled Wing–Force–Body Dynamics The third benchmark tests whether bilateral aerodynamic-force observations improve future body-motion prediction. The target and the two

matched forecasting tasks are

$$\mathbf{b}(t) = \left[\mathbf{v}_{\text{local}}^{\mathsf{T}}(t), \boldsymbol{\omega}_{\text{world}}^{\mathsf{T}}(t)\right]^{\mathsf{T}} \in \mathbb{R}^6, \quad \widehat{\mathbf{b}}_{\text{state}}(t+H) = \mathcal{D}_{\boldsymbol{\alpha}}(\mathbf{s}_{t-L+1:t}), \quad \widehat{\mathbf{b}}_{\text{force}}(t+H) = \mathcal{D}_{\boldsymbol{\beta}}([\mathbf{s}, \mathbf{F}]_{t-L+1:t}). \tag{4}$$

Where, $\mathbf{b}(t)$ combines local-frame body linear velocity and world-frame body angular velocity. The state-only model uses the 14-dimensional wing–body history $\mathbf{s}_{t-L+1:t}$, whereas the force-conditioned model additionally observes the bilateral aerodynamic-force history $\mathbf{F}_{t-L+1:t}$. The matched comparison therefore isolates the incremental predictive value of bilateral force observations. Detailed configurations and rollout conventions are provided in Appendix D.3.

3.4 Temporal Forecasting Formulation

All benchmark tasks follow a common temporal forecasting formulation. Given an input sequence $\mathbf{z}(t)$, the model observes the L-step history $\mathbf{z}_{t-L+1:t} = [\mathbf{z}(t-L+1), \ldots, \mathbf{z}(t)]$ and predicts the corresponding target at time $t+H$. Here, L is the observation-history length and H is the forecast horizon. The input and target variables are determined by the benchmark definitions above.

4 Experiments

Figure 6 summarizes the operating regimes represented by the released episodes. The broad joint coverage of wing-rate, body-motion, and force activity, together with the approximately balanced low–medium–high occupancy, shows that the collection is not dominated by a single physical regime. All experiments use the same predefined episode-level partition. Episodes are split before temporal-window construction, preventing windows from the same episode from appearing in multiple partitions.

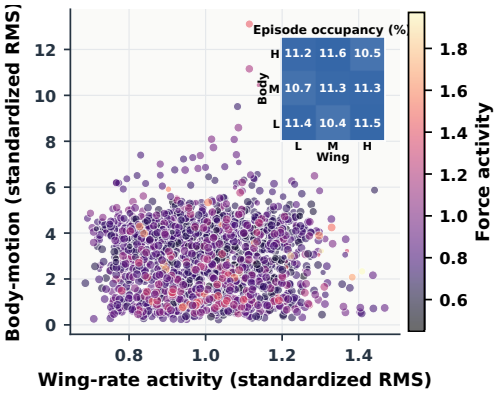

Figure 6: Episode-level coverage of wing, body, and force activity.

4.1 Unified Benchmark Protocol

Let $\mathbf{Z}_i(t)$ and $\mathbf{W}_i(t)$ denote the selected input and target representations for episode i. A benchmark sample is constructed as

$$\mathbf{X}_{i,t} = \mathbf{Z}_{i,t-L+1:t}, \qquad \mathbf{Y}_{i,t}^{(H)} = \mathbf{W}_i(t+H), \qquad L = 256,\ H \in \{1, 16, 32\}, \tag{5}$$

using unit stride. At 50 Hz, the input history spans 5.12 s, while the forecast horizons correspond to 20, 320, and 640 ms. The latter two intervals span approximately 1.3–1.5 and 2.7–3.1 wingbeat cycles over the simulated operating-frequency range. Let T_i denote the mask-defined valid length of episode i. A window is retained only when

$$L - 1 \le t \le T_i - H - 1, \tag{6}$$

ensuring that the complete input and target remain within one valid episode. No window crosses an episode boundary or includes padded storage positions. Consequently, the number of valid windows varies with episode length and forecast horizon, while the episode assignments remain fixed.

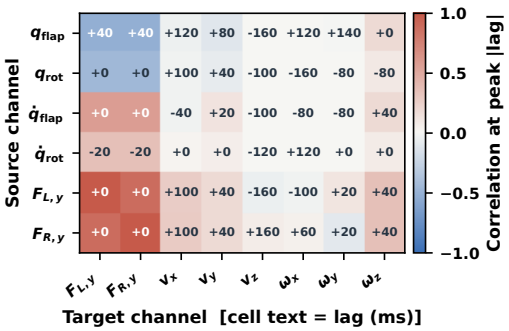

Figure 7: Peak-lag wing–force–body associations; cell values report lag in milliseconds.

Each selected channel is standardized using statistics estimated exclusively from mask-valid training observations:

$$\widetilde{u}_{t,c} = \frac{u_{t,c} - \mu_c^{\text{train}}}{\sigma_c^{\text{train}}}, \tag{7}$$

where u denotes an input or target channel. Separate statistics are estimated for each selected input and target channel and are applied unchanged to validation and test observations. Padded positions, validation observations, and test observations do not contribute to normalization.

4.2 Models, Training, and Selection

We evaluate Ridge, GRU, TSMixer, Grouped-State MLP, TimeXer, PySINDy, and Neural ODE, together with persistence for body-motion forecasting. Information-matched models use the task inputs defined in Sec. 3; TimeXer is reported separately because it additionally observes target history. For continuous-time references, body motion is the dynamical state and the remaining observations are exogenous controls propagated by a fixed zero-order hold when future values are unavailable. Stochastic models use three seeds and are reported as mean and sample standard deviation, while deterministic references are fitted once after validation-based hyperparameter selection. Matched configurations use identical optimization budgets and selection rules. Complete model definitions, parameter counts, hyperparameter grids, and implementation details are provided in Appendix D.

4.3 Evaluation Metrics

All errors are computed after predictions and references are transformed back to their original channel scales. Let $\mathcal{E}$ denote the evaluation episodes, N_e the number of valid windows in episode e, and D_y the target dimension. We report

$$\mathrm{MSE} = \frac{\sum_{e\in\mathcal{E}}\sum_{i=1}^{N_e}\|\widehat{\mathbf{y}}_{e,i}-\mathbf{y}_{e,i}\|_2^2}{D_y\sum_{e\in\mathcal{E}}N_e}, \quad \mathrm{MSE}_{\mathrm{EB}} = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}\frac{\sum_{i=1}^{N_e}\|\widehat{\mathbf{y}}_{e,i}-\mathbf{y}_{e,i}\|_2^2}{D_yN_e}. \tag{8}$$

Episode-balanced MSE is used for model selection. We additionally report element-wise MAE, the 95th percentile of absolute scalar error, and channel-wise MSE. For mixed-unit body-motion targets, aggregate errors are interpreted together with the corresponding channel-wise results.

4.4 Efficiency and Reproducibility

We report trainable parameter counts separately from predictive performance. All experiments use the released episode split, lineage-preserving artifact, and fixed preprocessing and selection procedures, while the test partition remains sealed during model development. Runtime, memory, software versions, and architecture-specific settings are reported in Appendix H.

5 Results

5.1 Benchmark 1: Bidirectional Wing-Force Mapping

Figure 8 summarizes relative performance using $L = 256$ and $H \in \{1, 16, 32\}$. Colors and hatching distinguish MSE, MAE, and E_{95}. Absolute errors are reported in Appendix Tables 1 and 2.

Forward prediction. TCN achieves the lowest MSE at all horizons, including 0.734 at $H = 1$ compared with 0.741 for GRU. GRU nevertheless yields the lowest E_{95} at $H = 1$, showing that the best average and tail errors need not coincide. The two models are nearly tied at $H = 16$, while TCN leads all metrics at $H = 32$, followed by PatchTST. Errors are not monotonic with horizon because each horizon induces a different eligible-window and target-position distribution.

Inverse prediction. The best architecture varies with horizon, shifting from GRU at $H = 1$ to TCN at $H = 16$ and PatchTST at $H = 32$. GRU's MSE rises from 0.052 to 0.234, while TimesNet remains comparatively stable, showing that strong one-step inverse performance does not necessarily persist at longer horizons.

5.2 Benchmark 2: State-Conditioned Aerodynamic Response

Figure 9 reports the state-enrichment comparison (top) and complete-state architecture comparison (bottom). The corresponding numerical results are provided in Appendix Tables 3 and 4.

Explicit wing rates. Explicit wing-rate channels yield negligible and inconsistent changes for Ridge and GRU, with all mean MSE differences below 0.5%. A 256-step wing-angle history therefore captures most of the predictive information provided by the rate channels.

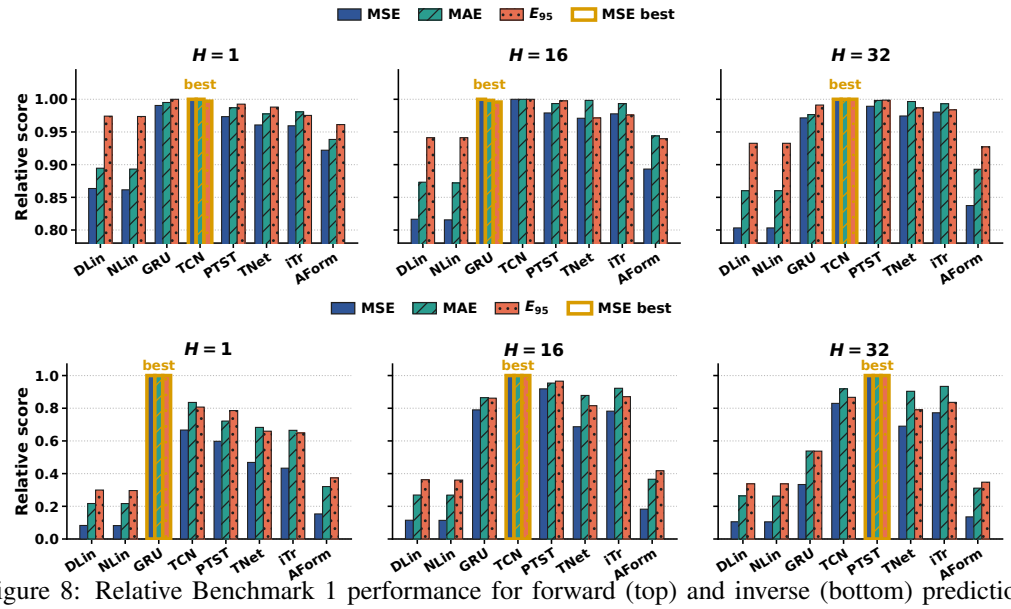

Figure 8: Relative Benchmark 1 performance for forward (top) and inverse (bottom) prediction. Errors are normalized metric-wise within each horizon; gold outlines mark the MSE-best model.

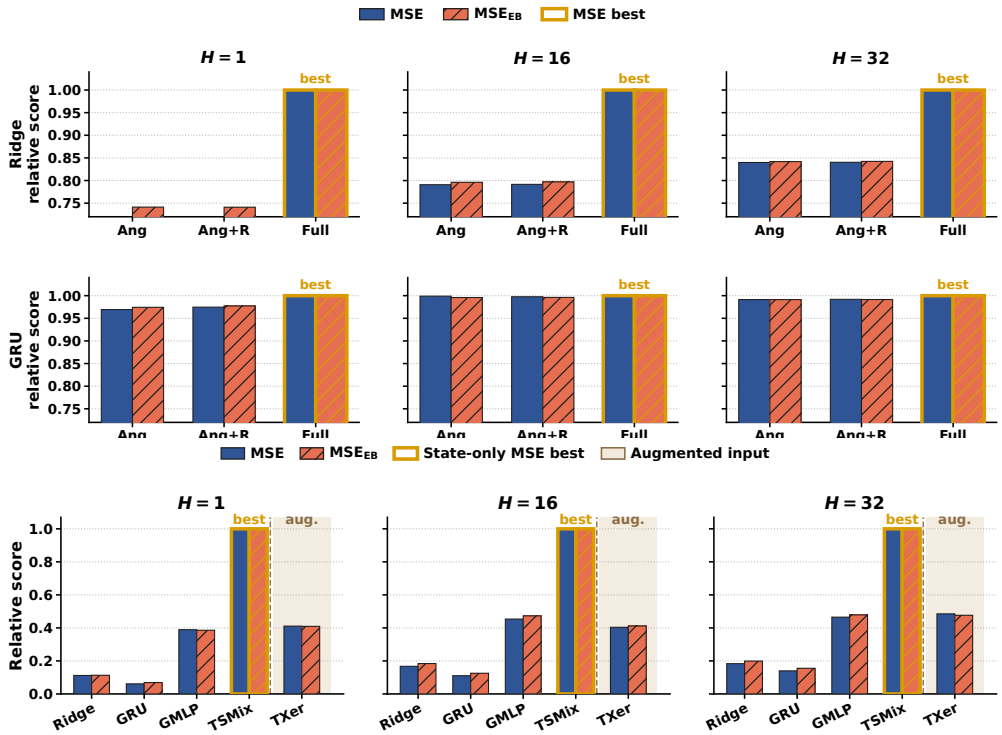

Figure 9: Benchmark 2 state enrichment (top) and architecture comparison (bottom). The upper panels compare wing angles, angles with rates, and the complete wing–body state within matched Ridge and GRU models. The lower panels compare architectures using the complete state; the shaded TimeXer setting additionally observes bilateral force history. Scores are normalized metric-wise within each comparison, gold outlines identifying the MSE-best matched configuration.

Body-state conditioning. Body motion provides substantially greater value. Ridge MSE decreases by 31.2%, 20.9%, and 15.9% across the three horizons, whereas GRU improves by 2.6% at $H = 1$ and by less than 1% at longer horizons. The additional information is therefore clear but strongly architecture-dependent.

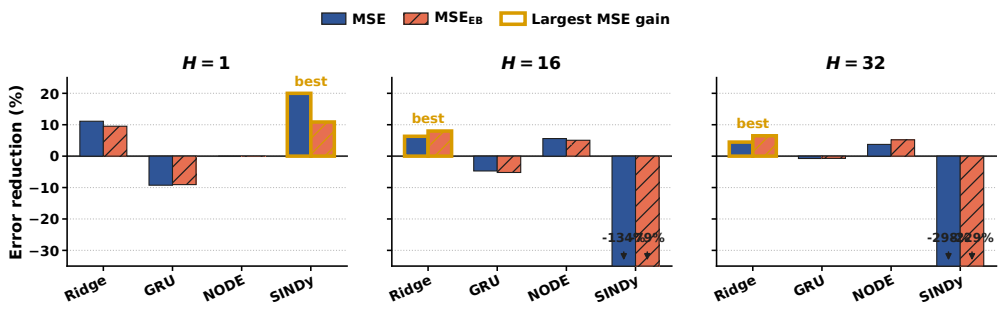

Figure 10: Effect of bilateral force conditioning on Benchmark 3 body-motion prediction. Positive values indicate error reduction relative to the matched state-only model, whereas negative values indicate degradation. Gold outlines mark the largest MSE gain at each horizon.

Architecture comparison. Through the complete wing–body state, TSMixer leads both MSE metrics at every horizon and substantially outperforms the other state-only references with approximately 0.140 million parameters. Grouped-State MLP remains less accurate, suggesting that joint temporal and cross-channel mixing is more effective than group-wise encoding. TimeXer receives additional force history but does not exceed TSMixer; because both its architecture and information set differ, this result does not isolate the value of force history.

5.3 Benchmark 3: Coupled Wing–Force–Body Dynamics

Figure 7 reveals channel-dependent temporal associations among bilateral wing motion, aerodynamic loading, and body motion. Several relationships peak at nonzero lags, motivating the multi-horizon comparison between state-only and force-conditioned forecasting. These correlations are interpreted as temporal associations rather than causal effects.

Figures 10 and 11 report the model-level effect of bilateral force history and its channel-wise Ridge decomposition, respectively. Complete numerical results are provided in Appendix Tables 5 and 6.

Persistence and linear prediction. Figure 10 shows that Persistence is competitive at one step but degrades rapidly with horizon. State-only Ridge substantially improves over Persistence, while adding bilateral force history yields a further consistent Ridge improvement across all horizons. The corresponding episode-balanced gains indicate that the benefit extends across episodes.

Recurrent prediction. Bilateral force history increases GRU error at all three horizons. The inconsistent seed-level effects indicate no stable incremental benefit for the evaluated GRU, without implying that force history is intrinsically uninformative.

Continuous-time neural prediction. Force conditioning provides no robust Neural ODE improvement at $H = 1$, where the two MSE metrics change in opposite directions. At $H = 16$ and $H = 32$, however, bilateral force history consistently reduces both metrics.

Sparse controlled dynamics. Force conditioning improves one-step PySINDy but substantially degrades the longer recursive forecasts. This reversal reflects sensitivity to the zero-order-hold assumption for unavailable future controls rather than evidence that force is intrinsically harmful.

Channel-wise force effects. Figure 11 shows that bilateral force history reduces Ridge error for all six body-motion channels at every horizon. The largest one-step gains occur for v_y and ω_z, whereas v_z and ω_z show the strongest longer-horizon improvements. The aggregate benefit therefore spans both translational and rotational motion.

5.4 Cross-Benchmark Analysis

Across the three benchmarks, predictive performance depends jointly on the task, forecast horizon, and model inductive bias. Forward wing-to-force rankings are relatively stable, whereas inverse prediction is substantially more horizon-dependent. The state-enrichment experiments distinguish redundant from complementary observations. Explicit wing rates add little beyond the long wing-angle history, whereas body motion provides clear additional information whose exploitation re-

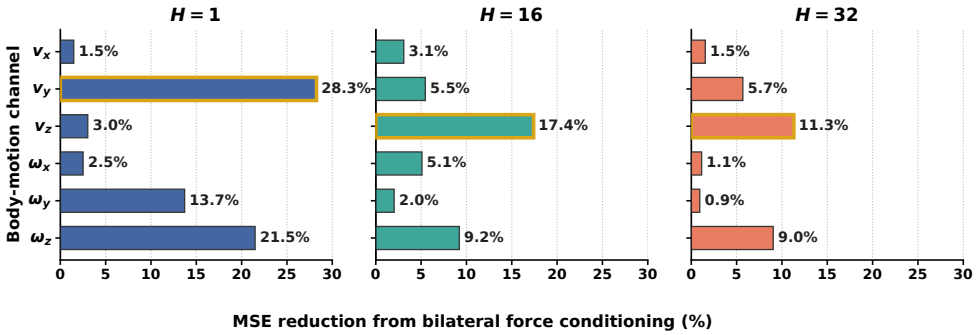

Figure 11: Channel-wise Ridge MSE reduction from bilateral force conditioning at the three forecast horizons. Gold outlines mark the largest reduction within each horizon.

mains architecture-dependent. TSMixer performs best with the complete wing–body state, supporting joint temporal and cross-channel mixing over recurrent or group-separated processing. Bilateral force history also provides no architecture-independent benefit for body-motion prediction. Its value ranges from consistent Ridge improvements to horizon-dependent Neural ODE gains, unstable recursive PySINDy behavior, and no stable GRU benefit. These results show that the utility of additional physical variables depends not only on their information content, but also on how each model represents temporal dynamics and incorporates that information.

6 Discussion

The aerodynamic contribution of each wing is evaluated independently within the fixed-step simulation loop. At each aerodynamic sampling location, the relative-flow velocity is resolved in the local chordwise, surface-normal, and spanwise directions. The local flow determines the instantaneous angle of attack and dynamic pressure used in the aerodynamic calculation. The resulting force is applied to the rigid body at the prescribed location, after which the updated body state and bilateral wing loads are recorded at the same simulation step. The formal dataset interface retains the simulator-y force component of each wing, $F_{L,y}(t)$ and $F_{R,y}(t)$. The corresponding aggregate load is a derived quantity, $F_{\mathrm{net},y}(t) = F_{L,y}(t) + F_{R,y}(t)$, rather than an additional learning target. Figure 5 examines the phase-resolved response, bilateral loading balance, angle-of-attack dependence, and numerical closure of the released force channels. Complete aerodynamic equations, coordinate conventions, and force application details are provided in Appendix B.

7 Conclusion

This work introduced FlapKAD-State, a state-rich dataset and benchmark for learning flapping-wing aerodynamics and coupled wing–force–body dynamics. The release contains 2,000 curated variable-length episodes and 720,152 valid records at 50 Hz, synchronizing bilateral wing angles and angular velocities, local-frame body linear velocity, world-frame body angular velocity, and left–right aerodynamic forces. Its 14-dimensional wing–body state and two-dimensional bilateral force interface support bidirectional wing–force prediction, state-enriched aerodynamic forecasting, and future body-motion prediction under a common multi-horizon protocol. The experiments show that richer physical observations do not yield architecture-independent gains: explicit wing rates add little beyond a long wing-angle history, body motion provides substantial additional value for linear force prediction, and TSMixer achieves the strongest complete-state force-prediction performance among the evaluated references. Bilateral force history consistently improves Ridge body-motion forecasting and benefits Neural ODE at longer horizons, but produces no stable GRU improvement and makes long PySINDy rollouts sensitive to the assumed future-control trajectory. These results distinguish the availability of additional physical information from a model's ability to exploit that information under a given temporal formulation. FlapKAD-State provides a reproducible foundation for aerodynamic prediction, inverse modeling, system identification, and coupled-dynamics forecasting.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript organization, language editing, LaTeX formatting, and the presentation and interpretation of experimental results. All AI-assisted text, and technical claims were reviewed and verified by the authors against the underlying data, source code, experiment logs, and cited literature. Generative AI tools were not treated as authors, did not independently conduct experiments, and did not determine the final scientific claims. The authors take full responsibility for the content of the paper and all associated artifacts.

REFERENCES

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, 2018. URL https://arxiv.org/abs/1803.01271.

Yagiz Bayiz and Bo Cheng. Flapping wing aerodynamics with PRSSM. Dryad, 2021a. URL https://doi.org/10.5061/dryad.zgmsbccbs. Dataset.

Yagiz E. Bayiz and Bo Cheng. State-space aerodynamic model reveals high force control authority and predictability in flapping flight. *Journal of The Royal Society Interface*, 18(181):20210222, 08 2021b. ISSN 1742-5689. doi: 10.1098/rsif.2021.0222. URL https://doi.org/10.1098/rsif.2021.0222.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Sparse identification of nonlinear dynamics with control (sindyc), 2016. URL https://arxiv.org/abs/1605.06682.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf.

Si-An Chen, Chun-Liang Li, Nate Yoder, Sercan O. Arik, and Tomas Pfister. Tsmixer: An all-mlp architecture for time series forecasting, 2023. URL https://arxiv.org/abs/2303.06053.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014. URL https://arxiv.org/abs/1406.1078.

Baptiste Corban, Michael Bauerheim, and Thierry Jardin. Discovering optimal flapping wing kinematics using active deep learning. *Journal of Fluid Mechanics*, 974:A54, 2023. doi: 10.1017/jfm.2023.832.

Brian M. de Silva, Kathleen Champion, Markus Quade, Jean-Christophe Loiseau, J. Nathan Kutz, and Steven L. Brunton. Pysindy: A python package for the sparse identification of nonlinear dynamical systems from data. *Journal of Open Source Software*, 5(49):2104, 2020. doi: 10.21105/joss.02104. URL https://doi.org/10.21105/joss.02104.

Michael H. Dickinson, Fritz-Olaf Lehmann, and Sanjay P. Sane. Wing rotation and the aerodynamic basis of insect flight. *Science*, 284(5422):1954–1960, 1999. doi: 10.1126/science.284.5422.1954. URL https://www.science.org/doi/abs/10.1126/science.284.5422.1954.

Thomas Engels, Henja-Niniane Wehmann, and Fritz-Olaf Lehmann. Three-dimensional wing structure attenuates aerodynamic efficiency in flapping fly wings. *Journal of The Royal Society Interface*, 17(164):20190804, 03 2020. ISSN 1742-5689. doi: 10.1098/rsif.2019.0804. URL https://doi.org/10.1098/rsif.2019.0804.

Alexander Gehrke and Karen Mulleners. Highly deformable flapping membrane wings suppress the leading edge vortex in hover to perform better. *Proceedings of the National Academy of Sciences*, 122(6):e2410833121, 2025. doi: 10.1073/pnas.2410833121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2410833121.

Ben J Hightower, Patrick WA Wijnings, Rick Scholte, Rivers Ingersoll, Diana D Chin, Jade Nguyen, Daniel Shorr, and David Lentink. How oscillating aerodynamic forces explain the timbre of the hummingbird's hum and other animals in flapping flight. *eLife*, 10:e63107, mar 2021. ISSN 2050-084X. doi: 10.7554/eLife.63107. URL https://doi.org/10.7554/eLife.63107.

Yi-Hsuan Hsiao, Andrea Tagliabue, Owen Matteson, Suhan Kim, Tong Zhao, Jonathan P. How, and YuFeng Chen. Aerobatic maneuvers in insect-scale flapping-wing aerial robots via deep-learned robust tube model predictive control. *Science Advances*, 11(49):eaea8716, 2025. doi: 10.1126/sciadv.aea8716. URL https://www.science.org/doi/abs/10.1126/sciadv.aea8716.

L. Christoffer Johansson. Aerodynamic efficiency explains flapping strategies used by birds. *Proceedings of the National Academy of Sciences*, 121(46):e2410048121, 2024. doi: 10.1073/pnas.2410048121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2410048121.

Yu Kamimizu, Hao Liu, and Toshiyuki Nakata. Data-driven discovery and formulation refines the quasi-steady model of flapping-wing aerodynamics, 2025. URL https://arxiv.org/abs/2508.18703.

Suhan Kim, Yi-Hsuan Hsiao, Zhijian Ren, Jiashu Huang, and Yufeng Chen. Acrobatics at the insect scale: A durable, precise, and agile micro–aerial robot. *Science Robotics*, 10(98):eadp4256, 2025. doi: 10.1126/scirobotics.adp4256. URL https://www.science.org/doi/abs/10.1126/scirobotics.adp4256.

Bluest Lan, You-Jun Lin, Yu-Hsiang Lai, Chia-Hung Tang, and Jing-Tang Yang. A neural network approach to estimate transient aerodynamic properties of a flapping wing system. *Drones*, 6(8), 2022. ISSN 2504-446X. doi: 10.3390/drones6080210. URL https://www.mdpi.com/2504-446X/6/8/210.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting, 2024. URL https://arxiv.org/abs/2310.06625.

James Lynch, Jeff Gau, Simon Sponberg, and Nick Gravish. Dimensional analysis of spring-wing systems reveals performance metrics for resonant flapping-wing flight. *Journal of The Royal Society Interface*, 18(175):20200888, 02 2021. ISSN 1742-5689. doi: 10.1098/rsif.2020.0888. URL https://doi.org/10.1098/rsif.2020.0888.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers, 2023. URL https://arxiv.org/abs/2211.14730.

João A. F. Pereira, Emanuel A. R. Camacho, Flávio D. Marques, and André R. R. Silva. Fast flapping aerodynamics prediction using a recurrent neural network. *Engineering Proceedings*, 56 (1), 2023. ISSN 2673-4591. doi: 10.3390/ASEC2023-16272. URL https://www.mdpi.com/2673-4591/56/1/219.

Nathan Phillips, Kevin Knowles, and Richard J. Bomphrey. Petiolate wings: effects on the leading-edge vortex in flapping flight. *Interface Focus*, 7(1):20160084, 02 2017. ISSN 2042-8898. doi: 10.1098/rsfs.2016.0084. URL https://doi.org/10.1098/rsfs.2016.0084.

Sanjay P. Sane. The aerodynamics of insect flight. *Journal of Experimental Biology*, 206(23):4191–4208, 12 2003. ISSN 0022-0949. doi: 10.1242/jeb.00663. URL https://doi.org/10.1242/jeb.00663.

Sanjay P. Sane and Michael H. Dickinson. The aerodynamic effects of wing rotation and a revised quasi-steady model of flapping flight. *Journal of Experimental Biology*, 205(8):1087–1096, 04 2002. ISSN 0022-0949. doi: 10.1242/jeb.205.8.1087. URL https://doi.org/10.1242/jeb.205.8.1087.

Hadar Sharvit, Raz Karl, and Tsevi Beatus. A deep inverse-mapping model for a flapping robotic wing. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=254NJe9JEw.

W. Shyy, H. Aono, S.K. Chimakurthi, P. Trizila, C.-K. Kang, C.E.S. Cesnik, and H. Liu. Recent progress in flapping wing aerodynamics and aeroelasticity. *Progress in Aerospace Sciences*, 46(7):284–327, 2010. ISSN 0376-0421. doi: https://doi.org/10.1016/j.paerosci.2010.01.001. URL https://www.sciencedirect.com/science/article/pii/S0376042110000023.

Usama Bin Sikandar, Brett R. Aiello, and Simon Sponberg. Body oscillations couple with wing flapping to reduce aerodynamic power in wild silk moth flight. *Journal of The Royal Society Interface*, 22(229):20250061, 08 2025. ISSN 1742-5689. doi: 10.1098/rsif.2025.0061. URL https://doi.org/10.1098/rsif.2025.0061.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. Timexer: Empowering transformers for time series forecasting with exogenous variables. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 469–498. Curran Associates, Inc., 2024. doi: 10.52202/079017-0015. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/0113ef4642264adc2e6924a3cbbdf532-Paper-Conference.pdf.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: decomposition transformers with auto-correlation for long-term series forecasting. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis, 2023. URL https://arxiv.org/abs/2210.02186.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i9.26317. URL https://doi.org/10.1609/aaai.v37i9.26317.

A Appendix

	$H = 1$			$H = 16$			$H = 32$		
Model	MSE	MAE	E_{95}	MSE	MAE	E_{95}	MSE	MAE	E_{95}
DLinear	0.850	0.683	1.786	0.861	0.694	1.770	0.808	0.679	1.694
NLinear	0.852	0.684	1.787	0.862	0.695	1.770	0.808	0.679	1.694
GRU	<u>0.741</u>	<u>0.614</u>	**1.740**	<u>0.703</u>	<u>0.607</u>	1.673	0.668	0.598	1.594
TCN	**0.734**	**0.611**	<u>1.745</u>	**0.703**	**0.606**	**1.666**	**0.649**	**0.584**	**1.580**
PatchTST	0.754	0.619	1.753	0.718	0.610	<u>1.670</u>	<u>0.656</u>	<u>0.585</u>	<u>1.582</u>
TimesNet	0.764	0.625	1.761	0.724	0.607	1.715	0.666	0.586	1.601
iTransformer	0.765	0.623	1.784	0.719	0.610	1.707	0.662	0.588	1.606
Autoformer	0.796	0.651	1.810	0.787	0.642	1.773	0.775	0.654	1.704

Table 1: Normalized test errors for Benchmark 1 forward prediction. Lower values are better. Bold and underlined entries denote the best and second-best values within each horizon. Rankings are determined using the unrounded values.

	$H = 1$			$H = 16$			$H = 32$		
Model	MSE	MAE	E_{95}	MSE	MAE	E_{95}	MSE	MAE	E_{95}
DLinear	0.628	0.585	1.579	0.688	0.617	1.633	0.739	0.641	1.676
NLinear	0.633	0.586	1.596	0.693	0.618	1.642	0.742	0.643	1.677
GRU	**0.052**	**0.127**	**0.473**	0.100	0.192	0.687	0.234	0.314	1.053
TCN	<u>0.078</u>	<u>0.152</u>	<u>0.586</u>	**0.079**	**0.166**	**0.592**	<u>0.094</u>	0.184	<u>0.653</u>
PatchTST	0.087	0.176	0.602	<u>0.086</u>	<u>0.174</u>	<u>0.613</u>	**0.078**	**0.169**	**0.566**
TimesNet	0.111	0.186	0.717	0.115	0.189	0.726	0.113	0.187	0.716
iTransformer	0.120	0.191	0.729	0.101	0.180	0.680	0.101	<u>0.181</u>	0.677
Autoformer	0.339	0.396	1.265	0.433	0.454	1.418	0.576	0.543	1.628

Table 2: Normalized test errors for Benchmark 1 inverse prediction. Lower values are better. Bold and underlined entries denote the best and second-best values within each horizon. Rankings are determined using the unrounded values.

B Simulation and Aerodynamic Validation

B.1 Simulation and Data Acquisition

FlapKAD-State was generated using a rigid-wing flapping-wing aerial-vehicle simulation implemented in Unity. The simulator integrates rigid-body dynamics, bilateral wing actuation, aerodynamic-force estimation, flight-control components, episode management, and synchronized data logging within a common fixed-step loop. The simulation and recording interval is $\Delta t = 0.02\,\mathrm{s}$, corresponding to 50 Hz.

Aerodynamic forces are evaluated independently for the left and right wings. For each wing, the simulator estimates the velocity of an aerodynamic sampling point relative to the ambient flow and resolves the velocity along the local chordwise, surface-normal, and spanwise directions. The local flow components determine the instantaneous angle of attack, while the relative-flow magnitude determines the dynamic pressure used to evaluate lift and drag. The resulting force is applied to the rigid body at the prescribed aerodynamic-force location.

Each episode begins from a reset state and includes a settling interval before data acquisition. During this interval, rigid-body motion and aerodynamic-force application are temporarily suspended so that reset transients are not recorded as physical observations. Logging begins after the settling procedure is complete and the controller, rigid-body, and aerodynamic components have been restored.

Episode conditions vary through the configured initialization, gait, operating-condition, and controller settings. Wingbeat frequency is randomized within a narrow operating range rather than fixed to one value. An episode terminates when its maximum duration is reached or when a predefined safety or validity condition is violated. Episode boundaries, source identifiers, valid lengths, and termination information are retained explicitly.

At every recorded step, the simulator synchronously records bilateral wing motion, aerodynamic quantities, controls, and vehicle states. FlapKAD-State exposes the variables required to construct the 14-dimensional wing-body state, bilateral aerodynamic-force interface, and three benchmark families defined in the main paper. Additional simulator variables remain available in the source records for quality control and physical-consistency analysis.

B.2 Aerodynamic Variable Definitions

Let $U_L(t)$ and $U_R(t)$ denote the nonnegative relative-flow speed magnitudes at the left- and right-wing aerodynamic sampling points. The descriptive reference speed and dynamic pressure are

$$U_{\mathrm{ref}}(t) = \frac{U_L(t) + U_R(t)}{2}, \qquad q_{\mathrm{ref}}(t) = \frac{1}{2}\rho U_{\mathrm{ref}}^2(t), \tag{9}$$

where ρ denotes fluid density.

The formal force interface retains the vertical aerodynamic component of each wing along the Unity world-frame y axis:

$$\mathbf{F}(t) = [F_{L,y}(t) \quad F_{R,y}(t)]^{\mathsf{T}}. \tag{10}$$

The aggregate vertical load is

$$F_{\mathrm{net},y}(t) = F_{L,y}(t) + F_{R,y}(t). \tag{11}$$

For descriptive aerodynamic analysis, $D_F(t) = q_{\mathrm{ref}}(t)S$, where S is the simulator reference area. The corresponding coefficients are

$$\begin{aligned} C_{F,L,y}(t) &= \frac{F_{L,y}(t)}{D_F(t)}, \qquad C_{F,R,y}(t) = \frac{F_{R,y}(t)}{D_F(t)}, \\ C_{F,\mathrm{net},y}(t) &= \frac{F_{\mathrm{net},y}(t)}{D_F(t)}. \end{aligned} \tag{12}$$

These normalized quantities support aerodynamic characterization and the normalized Benchmark 1 analysis. Benchmarks 2 and 3 use the bilateral forces in Eq. equation 10 directly.

The simulator additionally records

$$Re(t) = \frac{U_{\mathrm{ref}}(t)c}{\nu}, \qquad AR_{\mathrm{sim}} = \frac{R}{c}, \tag{13}$$

where c, R, and ν denote the reference chord, simulator wing span, and kinematic viscosity.

B.3 Aerodynamic Consistency and Bilateral Force Validation

Bilateral aerodynamic balance is characterized by

$$I_F(t) = \frac{2[F_{L,y}(t) - F_{R,y}(t)]}{|F_{L,y}(t)| + |F_{R,y}(t)| + \epsilon_F}, \tag{14}$$

where $\epsilon_F > 0$ prevents division by zero. Values near zero indicate balanced loading, while the sign identifies the wing with the larger instantaneous vertical load.

The phase-resolved bilateral loads exhibit similar patterns over the wingbeat cycle. Their differences retain variation associated with asymmetric operating conditions. The angle-of-attack analysis exhibits a nonlinear relationship between local angle of attack and vertical loading, with dispersion arising from flow speed, dynamic pressure, phase, and recent motion history.

Internal force consistency is evaluated using

$$r_F(t) = F_{\mathrm{main},y}(t) - [F_{L,y}(t) + F_{R,y}(t)]. \tag{15}$$

		$H = 1$		$H = 16$		$H = 32$	
Model	Input	MSE	$\mathrm{MSE}_{\mathrm{EB}}$	MSE	$\mathrm{MSE}_{\mathrm{EB}}$	MSE	$\mathrm{MSE}_{\mathrm{EB}}$
Ridge	$\mathbf{K}_4$	2.688	2.718	3.136	3.145	3.680	3.681
Ridge	$\mathbf{KV}_8$	2.689	2.719	3.133	3.142	3.678	3.679
Ridge	$\mathbf{KVB}_{14}$	**1.850**	**2.015**	**2.480**	**2.504**	**3.091**	**3.099**
GRU	$\mathbf{K}_4$	3.494 ± 0.095	3.389 ± 0.070	3.761 ± 0.019	3.693 ± 0.015	4.093 ± 0.003	4.006 ± 0.004
GRU	$\mathbf{KV}_8$	3.475 ± 0.094	3.379 ± 0.069	3.766 ± 0.023	3.692 ± 0.022	4.091 ± 0.010	4.005 ± 0.012
GRU	$\mathbf{KVB}_{14}$	$\mathbf{3.387 \pm 0.130}$	$\mathbf{3.302 \pm 0.127}$	$\mathbf{3.757 \pm 0.038}$	$\mathbf{3.678 \pm 0.032}$	$\mathbf{4.058 \pm 0.023}$	$\mathbf{3.971 \pm 0.025}$

Table 3: State-enrichment results for Benchmark 2. Ridge is deterministic after validation-based regularization selection; GRU values are mean ± sample standard deviation over three seeds. Bold entries denote the best input within each model family and horizon.

		$H = 1$		$H = 16$		$H = 32$	
Model	Information	MSE	$\mathrm{MSE}_{\mathrm{EB}}$	MSE	$\mathrm{MSE}_{\mathrm{EB}}$	MSE	$\mathrm{MSE}_{\mathrm{EB}}$
Ridge	State	1.850	2.015	2.480	2.504	3.091	3.099
GRU	State	3.387 ± 0.130	3.302 ± 0.127	3.757 ± 0.038	3.678 ± 0.032	4.058 ± 0.023	3.971 ± 0.025
Grouped MLP	State	0.532 ± 0.005	0.591 ± 0.007	0.915 ± 0.013	0.975 ± 0.005	1.219 ± 0.001	1.289 ± 0.009
TSMixer	State	$\mathbf{0.207 \pm 0.024}$	$\mathbf{0.228 \pm 0.025}$	$\mathbf{0.415 \pm 0.028}$	$\mathbf{0.461 \pm 0.032}$	$\mathbf{0.567 \pm 0.015}$	$\mathbf{0.617 \pm 0.010}$
TimeXer	State + force	0.504 ± 0.007	0.557 ± 0.012	1.028 ± 0.055	1.120 ± 0.058	1.169 ± 0.084	1.295 ± 0.083

Table 4: Architecture comparison for Benchmark 2 using the complete 14-dimensional wing–body state. TimeXer additionally observes bilateral force history and is therefore reported separately from the state-only models. Values for stochastic models are mean ± sample standard deviation over three seeds.

The measured root-mean-square closure error is

$$\mathrm{RMSE}_F = \sqrt{\frac{1}{N}\sum_{i=1}^{N} r_F^2(t_i)} = 1.53 \times 10^{-7}\,\mathrm{N}. \tag{16}$$

This confirms numerical consistency between the aggregate load and the bilateral force channels.

C Dataset Construction and Quality Audits

C.1 Episode Selection and Quality Control

The synchronized simulator records are processed episode by episode. An initial transient interval is removed before quality screening. Episodes are excluded if formal physical variables contain non-finite values, retained duration is insufficient, bilateral wing speed exceeds the prescribed threshold, or aggregate aerodynamic load falls outside the admissible range.

The resulting dataset contains 2,000 curated variable-length episodes and 720,152 valid observations. Released metadata records the source row, episode identifier, episode ordering, mask-defined valid length, and fixed train-validation-test assignment for every episode. Each quality record contains original and retained length, retained duration, validity status, filtering reason, maximum bilateral wing speed, maximum absolute aggregate load, mean body velocity, and mean absolute angle of attack.

C.2 Released State-Rich Representation

The public artifact stores

$$\begin{aligned} \mathbf{X} &\in \mathbb{R}^{2000\times 400\times 14}, \\ \mathbf{F} &\in \mathbb{R}^{2000\times 400\times 2}, \\ \mathbf{M} &\in \{0,1\}^{2000\times 400}. \end{aligned} \tag{17}$$

		$H = 1$		$H = 16$		$H = 32$	
Model	Input	MSE	$\mathrm{MSE}_{\mathrm{EB}}$	MSE	$\mathrm{MSE}_{\mathrm{EB}}$	MSE	$\mathrm{MSE}_{\mathrm{EB}}$
Persistence	Last state	0.030	0.050	0.535	0.809	1.170	1.320
Ridge	State	0.018	0.042	0.269	0.488	0.514	0.725
Ridge	State + force	**0.016**	**0.038**	**0.252**	**0.449**	**0.491**	**0.678**
GRU	State	$\mathbf{0.605 \pm 0.037}$	$\mathbf{0.761 \pm 0.037}$	$\mathbf{0.826 \pm 0.023}$	$\mathbf{0.980 \pm 0.031}$	$\mathbf{1.059 \pm 0.031}$	$\mathbf{1.259 \pm 0.054}$
GRU	State + force	0.661 ± 0.113	0.830 ± 0.133	0.865 ± 0.068	1.031 ± 0.085	1.067 ± 0.083	1.268 ± 0.110
Neural ODE	State	0.011 ± 0.000	$\mathbf{0.030 \pm 0.000}$	0.125 ± 0.003	0.257 ± 0.012	0.214 ± 0.004	0.308 ± 0.011
Neural ODE	State + force	$\mathbf{0.011 \pm 0.000}$	0.030 ± 0.000	$\mathbf{0.118 \pm 0.004}$	$\mathbf{0.244 \pm 0.022}$	$\mathbf{0.206 \pm 0.008}$	$\mathbf{0.292 \pm 0.013}$
PySINDy	State	0.025	0.046	**0.570**	**0.935**	**1.300**	**1.597**
PySINDy	State + force	**0.020**	**0.041**	1.336	1.674	5.169	5.262

Table 5: Complete Benchmark 3 body-motion prediction results. GRU and Neural ODE values are mean $\pm$ sample standard deviation over three seeds; Ridge and PySINDy are deterministic after validation-based selection. Bold entries indicate the better matched input within each model family and horizon.

	$H = 1$		$H = 16$		$H = 32$	
Channel	WB	WFB	WB	WFB	WB	WFB
v_x	0.01281	**0.01262**	0.46636	**0.45195**	1.14027	**1.12268**
v_y	0.01058	**0.00759**	0.42169	**0.39866**	0.77745	**0.73323**
v_z	0.00759	**0.00736**	0.12983	**0.10723**	0.24508	**0.21736**
ω_x	0.02702	**0.02634**	0.18302	**0.17370**	0.26156	**0.25858**
ω_y	0.01174	**0.01013**	0.09441	**0.09250**	0.14553	**0.14417**
ω_z	0.03946	**0.03098**	0.31578	**0.28670**	0.51681	**0.47010**

Table 6: Channel-wise Ridge MSE for Benchmark 3. WB and WFB denote the state-only and force-conditioned inputs, respectively. Bold values identify the lower channel error at each horizon. Linear velocities are measured in local body coordinates and angular velocities in world coordinates.

The state order is

$$\mathbf{x}(t) = \big[\phi_L, \theta_L, \phi_R, \theta_R, \dot{\phi}_L, \dot{\theta}_L, \dot{\phi}_R, \dot{\theta}_R, \\ v_x, v_y, v_z, \omega_x, \omega_y, \omega_z\big]^{\mathsf{T}}. \tag{18}$$

Wing angles are expressed in degrees, wing rates in degrees per second, body linear velocity in metres per second, and body angular velocity in radians per second. Only positions satisfying $M_{i,t} = 1$ contribute to normalization, temporal sampling, and evaluation.

C.3 Artifact Consistency Audits

All state and force values are finite over the 720,152 valid observations. Every stored episode length equals the corresponding row-wise validity-mask sum, and every formal channel has nonzero global variance. Temporal alignment is checked across bilateral wing angles, wing rates, body velocity, body angular velocity, and bilateral force. Every angular-velocity component has nonzero temporal variance in all 2,000 episodes.

C.4 Bilateral Kinematic Consistency

Under the simulator coordinate convention,

$$\phi_L(t) \approx -\phi_R(t), \qquad \theta_L(t) \approx \theta_R(t). \tag{19}$$

The instantaneous residuals are

$$r_\phi(t) = |\phi_L(t) + \phi_R(t)|, \qquad r_\theta(t) = |\theta_L(t) - \theta_R(t)|. \tag{20}$$

After sign alignment, bilateral flap trajectories exhibit close agreement over the wingbeat cycle, while twist trajectories show similarly small discrepancies under the shared sign convention.

Table 7: Global body angular-velocity statistics in radians per second.

Channel	Mean	Std.	Minimum	Maximum
ω_x	0.0004	0.3068	-7.1479	7.6803
ω_y	0.0022	0.2919	-5.2008	4.9401
ω_z	0.1241	0.8979	-7.9730	6.1906

Table 8: Channel-wise normalized kinematic-prediction MSE for the best aggregate model at each horizon.

Horizon and model	Flap	Twist	Aggregate MSE
$H = 1$, GRU	**0.011**	0.092	0.052
$H = 16$, TCN	**0.020**	0.137	0.079
$H = 32$, PatchTST	**0.024**	0.132	0.078

D Detailed Benchmark Results

All percentage changes are computed from unrounded values. Metrics are rounded to three decimals unless additional precision is required.

D.1 Benchmark 1: Channel-Wise Kinematic Prediction

Twist-angle error remains higher than flap-angle error for all horizon-specific winners. This pattern occurs across GRU, TCN, and PatchTST and is therefore not specific to one architecture. The benchmark does not isolate the underlying cause, so the difference is treated as an empirical property rather than a causal conclusion.

D.2 Benchmark 2: Complete Ridge Results

Wing-rate increment. The changes from $\mathbf{K}_4$ to $\mathbf{KV}_8$ are negligible. MSE changes are -0.030%, $+0.082\%$, and $+0.068\%$ at $H = 1, 16, 32$, while episode-balanced changes remain within 0.11%. A 256-step angle history therefore contains nearly all linear predictive information supplied by explicit rates.

Body-state increment. Relative to $\mathbf{KV}_8$, $\mathbf{KVB}_{14}$ reduces MSE by 31.207%, 20.851%, and 15.943%; episode-balanced reductions are 25.885%, 20.289%, and 15.761%. The improvement therefore arises from body motion rather than explicit wing-rate channels.

D.3 Benchmark Input Configurations

The state-enrichment benchmark uses three nested input configurations:

$$\begin{aligned} \textsc{WingAngle-4}: &\quad \mathbf{q}(t) \in \mathbb{R}^4, \\ \textsc{WingState-8}: &\quad [\mathbf{q}(t)^\mathsf{T}, \dot{\mathbf{q}}(t)^\mathsf{T}]^\mathsf{T} \in \mathbb{R}^8, \\ \textsc{WingBodyState-14}: &\quad [\mathbf{q}(t)^\mathsf{T}, \dot{\mathbf{q}}(t)^\mathsf{T}, \mathbf{v}_{\mathrm{local}}(t)^\mathsf{T}, \boldsymbol{\omega}_{\mathrm{world}}(t)^\mathsf{T}]^\mathsf{T} \in \mathbb{R}^{14}. \end{aligned} \tag{21}$$

D.4 Benchmark 2: GRU Seed-Level State Increments

At $H = 1$, every seed improves from $\mathbf{KV}_8$ to $\mathbf{KVB}_{14}$, supporting a measurable short-horizon body-state increment. At $H = 32$, all seeds again improve, but mean gain remains below 1%. No horizon establishes a meaningful and stable increment for explicit wing-rate channels.

Table 9: Complete Ridge results for Benchmark 2.

Horizon	Input	α	MSE	EBMSE	MAE	E_{95}
$H = 1$	$\mathbf{K}_4$	1,000	2.688	2.718	1.208	3.246
$H = 1$	$\mathbf{KV}_8$	1,000	2.689	2.719	1.209	3.247
$H = 1$	$\mathbf{KVB}_{14}$	10,000	**1.850**	**2.015**	**0.923**	**2.691**
$H = 16$	$\mathbf{K}_4$	100	3.136	3.145	1.334	3.471
$H = 16$	$\mathbf{KV}_8$	1,000	3.133	3.142	1.334	3.474
$H = 16$	$\mathbf{KVB}_{14}$	10,000	**2.480**	**2.504**	**1.123**	**3.123**
$H = 32$	$\mathbf{K}_4$	1,000	3.680	3.681	1.465	3.726
$H = 32$	$\mathbf{KV}_8$	1,000	3.678	3.679	1.464	3.722
$H = 32$	$\mathbf{KVB}_{14}$	10,000	**3.091**	**3.099**	**1.278**	**3.503**

Table 10: Paired Test-MSE improvements for evaluated GRU configurations in Benchmark 2. Values are percentages.

Horizon	Comparison	Seed 42	Seed 1024	Seed 3407	Mean	Improved seeds
$H = 1$	$\mathbf{K}_4 \rightarrow \mathbf{KV}_8$	0.482	4.455	-3.512	0.475	2/3
$H = 1$	$\mathbf{KV}_8 \rightarrow \mathbf{KVB}_{14}$	3.071	3.902	0.719	2.564	3/3
$H = 1$	$\mathbf{K}_4 \rightarrow \mathbf{KVB}_{14}$	3.538	8.183	-2.768	2.984	2/3
$H = 16$	$\mathbf{K}_4 \rightarrow \mathbf{KV}_8$	-0.771	-0.731	1.121	-0.127	1/3
$H = 16$	$\mathbf{KV}_8 \rightarrow \mathbf{KVB}_{14}$	1.865	-0.028	-1.144	0.231	1/3
$H = 16$	$\mathbf{K}_4 \rightarrow \mathbf{KVB}_{14}$	1.109	-0.759	-0.010	0.113	1/3
$H = 32$	$\mathbf{K}_4 \rightarrow \mathbf{KV}_8$	-0.208	0.299	0.038	0.043	2/3
$H = 32$	$\mathbf{KV}_8 \rightarrow \mathbf{KVB}_{14}$	0.671	1.195	0.581	0.816	3/3
$H = 32$	$\mathbf{K}_4 \rightarrow \mathbf{KVB}_{14}$	0.464	1.491	0.619	0.858	3/3

E Detailed Coupled-Dynamics Results

E.1 Persistence and Linear Prediction

Relative to Persistence, state-only Ridge reduces MSE by 38.544%, 49.777%, and 56.047%. Force-conditioned Ridge reduces MSE by 46.525%, 52.906%, and 58.049%. The increasing advantage over Persistence confirms that learned history dependence becomes more important at longer offsets.

E.2 Cross-Model Effect of Force Conditioning

Ridge consistently benefits at every horizon. Neural ODE obtains multi-seed evidence of force value at $H = 16$ and $H = 32$. PySINDy obtains the largest one-step improvement but becomes less robust under longer recursive rollouts. GRU does not obtain a positive average effect under the evaluated configuration.

E.3 GRU Seed-Level Force Effects

At $H = 32$, two seeds improve, but degradation for seed 1024 is sufficiently large to produce worse mean performance. Counting improved seeds alone is therefore insufficient.

E.4 Neural ODE Seed-Level Force Effects

The one-step Neural ODE effect is not robust under episode-balanced evaluation. At $H = 16$ and $H = 32$, the mean gains exceed 1% and at least two seeds improve, providing the strongest multi-seed evidence of incremental force value for this model.

Table 11: Complete Persistence and Ridge results for Benchmark 3.

Horizon	Model and input	α	MSE	EBMSE	MAE	E_{95}
$H = 1$	Persistence	–	0.030	0.050	0.090	0.297
$H = 1$	Ridge, $\mathbf{WB}_{14}$	100	0.018	0.042	0.059	0.185
$H = 1$	Ridge, $\mathbf{WFB}_{16}$	100	**0.016**	**0.038**	**0.049**	**0.154**
$H = 16$	Persistence	–	0.535	0.809	0.485	1.444
$H = 16$	Ridge, $\mathbf{WB}_{14}$	10,000	0.269	0.488	0.288	0.921
$H = 16$	Ridge, $\mathbf{WFB}_{16}$	10,000	**0.252**	**0.449**	**0.277**	**0.890**
$H = 32$	Persistence	–	1.170	1.320	0.747	2.271
$H = 32$	Ridge, $\mathbf{WB}_{14}$	10,000	0.514	0.725	0.407	1.309
$H = 32$	Ridge, $\mathbf{WFB}_{16}$	10,000	**0.491**	**0.678**	**0.392**	**1.261**

Table 12: Relative effect of force conditioning in Benchmark 3. Positive values indicate error reduction.

	$H = 1$		$H = 16$		$H = 32$	
Model	ΔMSE	ΔEB-MSE	ΔMSE	ΔEB-MSE	ΔMSE	ΔEB-MSE
Ridge	+12.988%	+7.537%	+6.229%	+8.058%	+4.554%	+6.432%
GRU	−8.773%	−8.656%	−4.605%	−5.097%	−0.998%	−1.014%
Neural ODE	+0.658%	−0.244%	+5.677%	+4.925%	+3.377%	+4.972%
PySINDy	+19.864%	+10.121%	−134.234%	−79.037%	−297.452%	−229.609%

E.5 PySINDy Configurations and Rollout Behavior

Validation selects second-order systems for one-step estimation and sparser first-order systems for longer rollouts. The one-step WFB model contains more force-dependent structure and improves every aggregate metric. At longer horizons, WFB systems remain finite but accumulate much larger recursive-rollout error. This behavior is consistent with model and zero-order-held control mismatch, not numerical overflow.

F Channel-Wise Body-Motion Analysis

F.1 Ridge Channel Errors

At $H = 1$, the largest Ridge gains occur for v_y and ω_z, with reductions of 28.240% and 21.493%. At $H = 16$, the largest gains occur for v_z and ω_z, with reductions of 17.410% and 9.210%. At $H = 32$, these channels improve by 11.311% and 9.038%. Force therefore benefits both translational and rotational components, with axis-dependent magnitude.

F.2 PySINDy Channel Errors

At $H = 1$, every channel improves, with the largest reductions for v_y, ω_z, ω_y, and ω_x. At $H = 16$ and $H = 32$, WFB retains a small advantage only for ω_x. Long-horizon degradation is dominated by ω_z and v_y, demonstrating strongly axis-dependent recursive instability.

G External Experimental Forward Comparison

G.1 Shared-Channel Protocol

This comparison evaluates whether forward architecture-level learning patterns observed in FlapKAD-State also appear in an independent experimental flapping-wing dataset. The experimental data do not expose the complete 14-dimensional state, so the comparison is restricted to shared wing-motion and aerodynamic-response channels. Simulation flap and experimental stroke describe

Table 13: Paired GRU force-conditioning effects. Positive values indicate improvement.

Horizon	Metric	Seed 42	Seed 1024	Seed 3407	Mean	Improved seeds
$H = 1$	MSE	4.013	-20.334	-9.998	-8.773	1/3
$H = 1$	EB-MSE	4.023	-20.476	-9.515	-8.656	1/3
$H = 16$	MSE	0.492	-11.472	-2.836	-4.605	1/3
$H = 16$	EB-MSE	0.383	-12.248	-3.427	-5.097	1/3
$H = 32$	MSE	7.468	-13.231	2.767	-0.998	2/3
$H = 32$	EB-MSE	8.696	-16.160	4.421	-1.014	2/3

Table 14: Paired Neural ODE force-conditioning effects. Positive values indicate improvement.

Horizon	Metric	Seed 42	Seed 1024	Seed 3407	Mean	Improved seeds
$H = 1$	MSE	0.713	-0.377	1.637	0.658	2/3
$H = 1$	EB-MSE	1.091	-0.726	-1.095	-0.244	1/3
$H = 16$	MSE	1.143	9.093	6.795	5.677	3/3
$H = 16$	EB-MSE	-6.941	12.416	9.301	4.925	2/3
$H = 32$	MSE	7.410	4.398	-1.677	3.377	2/3
$H = 32$	EB-MSE	11.224	0.987	2.704	4.972	3/3

the same upward-and-downward wing motion, while simulation wing rotation and experimental pitch describe the corresponding rotational coordinate.

Both datasets use episode-level partitions, a 256-step input history, direct offsets $H \in \{1, 16, 32\}$, training-derived normalization, the same model implementations, and normalized-space errors. Each dataset is split and standardized independently, so normalized magnitudes describe relative task difficulty rather than equality of physical error. The seven common models are DLinear, NLinear, GRU, TCN, PatchTST, TimesNet, and iTransformer, all evaluated with seed 42.

G.2 Forward Prediction on Simulation Data

TCN obtains the lowest simulation MSE at all three horizons. GRU remains in the leading group at $H = 1$ and $H = 16$, while PatchTST is second at $H = 32$. DLinear and NLinear consistently occupy the lowest-performing group.

G.3 Forward Prediction on Experimental Data

TCN and GRU form the leading experimental group. TCN obtains the lowest error at $H = 1$, while GRU leads at $H = 16$ and $H = 32$. The linear baselines again occupy the lowest-performing group.

G.4 Forward Architecture-Ranking Agreement

Forward prediction shows strong and statistically significant ranking agreement at $H = 1$ and $H = 16$. Both settings have complete Top-2 overlap, and GRU and TCN form the leading group in both datasets. At $H = 32$, agreement is moderate: TCN remains in the experimental Top-2, but GRU rises from the middle simulation group to first experimentally. The persistence of recurrent and convolutional models near the top and linear baselines near the bottom provides external evidence that the simulation reproduces nontrivial architecture-level forward-learning patterns.

This is not a same-platform fidelity test. The datasets do not contain matched trials, identical geometry, identical operating conditions, or measurements from the same system. The comparison does not establish zero-shot transfer or equality of physical error; it quantifies relative architecture agreement across independent simulation and experimental data.

Table 15: Validation-selected PySINDy configurations and test results.

Horizon	Input	Degree	Threshold	α	Terms	MSE	EB-MSE	E_{95}
$H = 1$	**WB**	2	0.01	10^{-5}	197	0.025	0.046	0.259
$H = 1$	**WFB**	2	0.01	10^{-3}	288	**0.020**	**0.041**	**0.205**
$H = 16$	**WB**	1	0.05	10^{-5}	29	**0.570**	**0.935**	**1.491**
$H = 16$	**WFB**	1	0.10	10^{-5}	35	1.336	1.674	2.374
$H = 32$	**WB**	1	0.05	10^{-5}	29	**1.300**	**1.597**	**2.366**
$H = 32$	**WFB**	1	0.10	10^{-5}	35	5.169	5.262	4.812

Table 16: Channel-wise Ridge MSE for state-only and force-conditioned inputs.

	$H = 1$		$H = 16$		$H = 32$	
Channel	WB	WFB	WB	WFB	WB	WFB
v_x	0.01281	**0.01262**	0.46636	**0.45195**	1.14027	**1.12268**
v_y	0.01058	**0.00759**	0.42169	**0.39866**	0.77745	**0.73323**
v_z	0.00759	**0.00736**	0.12983	**0.10723**	0.24508	**0.21736**
ω_x	0.02702	**0.02634**	0.18302	**0.17370**	0.26156	**0.25858**
ω_y	0.01174	**0.01013**	0.09441	**0.09250**	0.14553	**0.14417**
ω_z	0.03946	**0.03098**	0.31578	**0.28670**	0.51681	**0.47010**

H Model Configurations and Reproducibility

All benchmark families use $L = 256$, $H \in \{1, 16, 32\}$, and stride one. The fixed partition contains 1,600 training, 200 validation, and 200 test episodes. Normalization statistics are estimated exclusively from valid training observations.

The state-rich GRU uses two recurrent layers, hidden size 128, dropout 0.1, AdamW with learning rate 10^{-3} and weight decay 10^{-4}, a 50-epoch budget, gradient clipping at 1.0, and early-stopping patience of eight epochs. Ridge uses $\alpha \in \{1, 10, 100, 1000, 10000\}$. PySINDy uses polynomial degree $\{1, 2\}$, STLSQ threshold $\{0.01, 0.05, 0.1\}$, and $\alpha \in \{10^{-5}, 10^{-3}\}$. PySINDy uses zero-order-held controls and explicit Euler rollout with $\Delta t = 0.02$ s. Neural ODE uses a neural derivative function and fixed-step fourth-order Runge–Kutta integration.

I Interpretation Boundaries

The force interface contains left- and right-wing vertical aerodynamic-force components rather than complete three-dimensional forces and moments. Body6 contains local translational velocity and world-frame angular velocity rather than full pose, acceleration, and moment reconstruction.

WB-to-WFB comparisons measure incremental predictive information and do not establish a causal force-to-body mechanism. Angular-velocity improvement should not be interpreted as direct aerodynamic-moment identification.

Long-horizon PySINDy degradation is specific to recursive rollout with zero-order-held controls and does not imply that force is intrinsically harmful. Likewise, lack of improvement under the evaluated GRU configuration does not imply that recurrent models cannot use force information.

The external experimental comparison is restricted to forward prediction over shared wing-motion and aerodynamic variables. It does not validate the complete 14-dimensional state, establish same-platform physical fidelity, or constitute zero-shot simulation-to-real transfer.

Table 17: Channel-wise PySINDy MSE for state-only and force-conditioned dynamics.

	$H = 1$		$H = 16$		$H = 32$	
Channel	WB	WFB	WB	WFB	WB	WFB
v_x	0.01488	**0.01473**	**0.80168**	0.82874	**1.53063**	2.98819
v_y	0.01889	**0.01283**	**0.72307**	1.69956	**1.47145**	4.32387
v_z	0.00820	**0.00801**	**0.23373**	0.23471	**0.83160**	0.93191
ω_x	0.03087	**0.02662**	0.30049	**0.29794**	0.58152	**0.57370**
ω_y	0.02131	**0.01722**	**0.24160**	0.43492	**0.33478**	0.80511
ω_z	0.05674	**0.04149**	**1.12165**	4.52013	**3.05264**	21.38888

Table 18: Forward prediction on FlapKAD-State using seed 42. Errors are in normalized target space.

	$H = 1$			$H = 16$			$H = 32$		
Model	MSE	MAE	P95	MSE	MAE	P95	MSE	MAE	P95
DLinear	0.850	0.683	1.786	0.861	0.694	1.770	0.808	0.679	1.694
NLinear	0.852	0.684	1.787	0.862	0.695	1.770	0.808	0.679	1.694
GRU	0.741	0.614	**1.740**	0.703	0.607	1.673	0.668	0.598	1.594
TCN	**0.734**	**0.611**	1.745	**0.703**	**0.606**	**1.666**	**0.649**	**0.584**	**1.580**
PatchTST	0.754	0.619	1.753	0.718	0.610	1.670	0.656	0.585	1.582
TimesNet	0.764	0.625	1.761	0.724	0.607	1.715	0.666	0.586	1.601
iTransformer	0.765	0.623	1.784	0.719	0.610	1.707	0.662	0.588	1.606

Table 19: Forward prediction on the independent experimental dataset using seed 42. Errors are in normalized target space.

	$H = 1$			$H = 16$			$H = 32$		
Model	MSE	MAE	P95	MSE	MAE	P95	MSE	MAE	P95
DLinear	0.381	0.476	1.225	0.401	0.497	1.233	0.490	0.545	1.418
NLinear	0.387	0.491	1.204	0.413	0.502	1.265	0.444	0.524	1.315
GRU	0.146	0.278	0.827	**0.170**	**0.307**	**0.858**	**0.187**	**0.323**	**0.916**
TCN	**0.142**	**0.273**	**0.801**	0.185	0.317	0.898	0.209	0.336	0.938
PatchTST	0.294	0.418	1.105	0.289	0.414	1.083	0.276	0.403	1.064
TimesNet	0.211	0.342	0.966	0.231	0.361	1.011	0.228	0.360	1.011
iTransformer	0.224	0.356	0.989	0.215	0.346	0.968	0.217	0.348	0.990

Table 20: Forward MSE-ranking agreement between simulation and experiment.

H	Best simulation	Best experiment	Top-2	ρ	p
1	TCN	TCN	2/2	0.893	0.007
16	TCN/GRU	GRU	2/2	0.883	0.008
32	TCN	GRU	1/2	0.523	0.229

Table 21: Reference models used in Benchmarks 2 and 3.

Model	Benchmark role	Seeds	Selection criterion
Persistence	Body-state continuity	Deterministic	None
Ridge	Linear history reference	Deterministic	Validation EB-MSE over α
GRU	Recurrent sequence reference	42, 1024, 3407	Validation EB-MSE
Grouped-State MLP	Separate wing/body encoding	42, 1024, 3407	Validation EB-MSE
TSMixer	Temporal-channel mixing	42, 1024, 3407	Validation EB-MSE
TimeXer	Target-history augmentation	42, 1024, 3407	Validation EB-MSE
Neural ODE	Continuous-time body dynamics	42, 1024, 3407	Validation EB-MSE
PySINDy	Sparse controlled dynamics	Deterministic	Validation EB-MSE over sparse grid