Model scale versus domain knowledge in statistical forecasting of chaotic systems

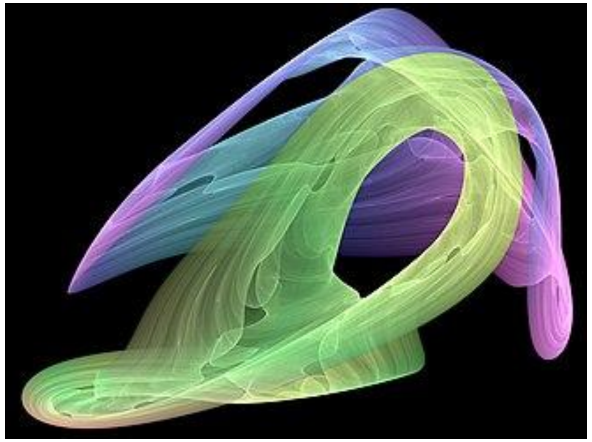
William Gilpin

(https://journals.aps.org/prresearch/abstract/10.1103/PhysRevResearch.5.043252)

ML Journal Club 2/21/2023

Background: Dynamical Systems

- System with evolution of state represented by differential equations
- Behavior described by chaotic attractor: path of system is state-space
- Chaotic: system in which a small difference in initial conditions grows exponentially (e.g. weather and climate)
- Characterized by *invariant properties*
  - e.g. Lyapunov time λ_{max}^{-1} : characteristic e-folding time on which system is chaotic
- Forecasting of such systems has improved… why? 2 types of models



Strange (fractally structured) attractor, often associated with chaotic systems

Background: Physics-based models

- “Domain Knowledge”
- Better represent system at hand/chaos
- Examples:
  - Reservoir computing (lift into higher dimension, making relationships more linear)
  - Neural ODEs
  - Physics-informed NNs
  - RNNs with domain-specific structural design
- Can we think of any examples within atmospheric science?

Background: Domain-agnostic models

- “Model Scale”
- Large, overparameterized not built with knowledge of field
- Examples:
    - Transformers
    - Hierarchical NNs
- Perform well with sufficient data
- Can we think of any examples in atmospheric science?

Methods

- Compared models across 135 dynamical systems

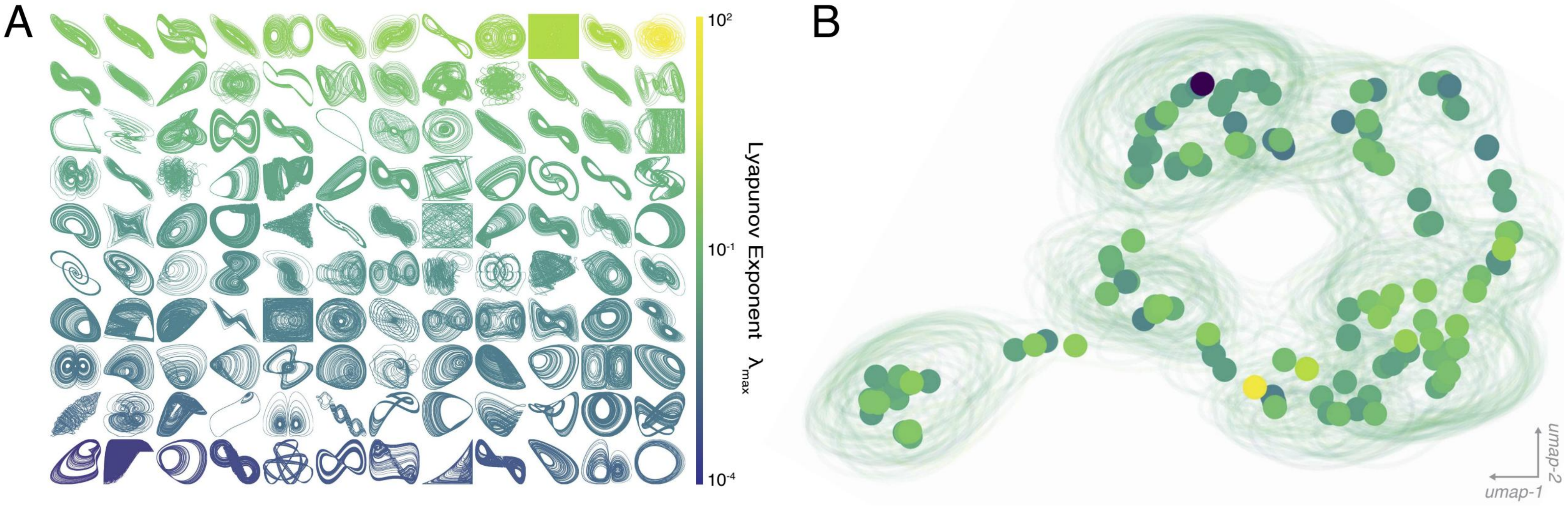


Figure 1. **A space of low-dimensional chaotic systems.** (A) A dataset of 135 distinct low-dimensional chaotic systems, colored by largest Lyapunov exponent ($\lambda_{\max}$). (B) A nonlinear embedding of the attractors. Each attractor is featurized using 747 invariant properties such as entropy, fractal dimension, et al., and then embedded in a two-dimensional vector space with UMAP. Contours denote 50% confidence intervals in each system's embedding across 500 random initial conditions and feature subsets; points denote centroids for each system.

Methods

- 24 models tested on each system (full list in Appendix D)
  - Physics-based
  - Domain-agnostic
  - Naive (e.g. regression/mean)
- Hyperparameters tuned for each model (different from previous experiments)
- Terms:
  - T_l: lookback window (i.e. input size)
  - t*: history length (i.e. training size)
  - t: forecast horizon
  - λ_{max}^{-1}: Lyapunov time

Results: Figure 2

- Discuss:
  - Which models performed well?
    - NBEATS, NHiTS, LSTM, transformer are large domain-agnostic
    - ESN, nVAR, nODE are physics-based
    - Compared to each other and to naïve methods?
  - Solid performance up to 14 Lyapunov times is better than historical results
  - What is forecast horizon scale?

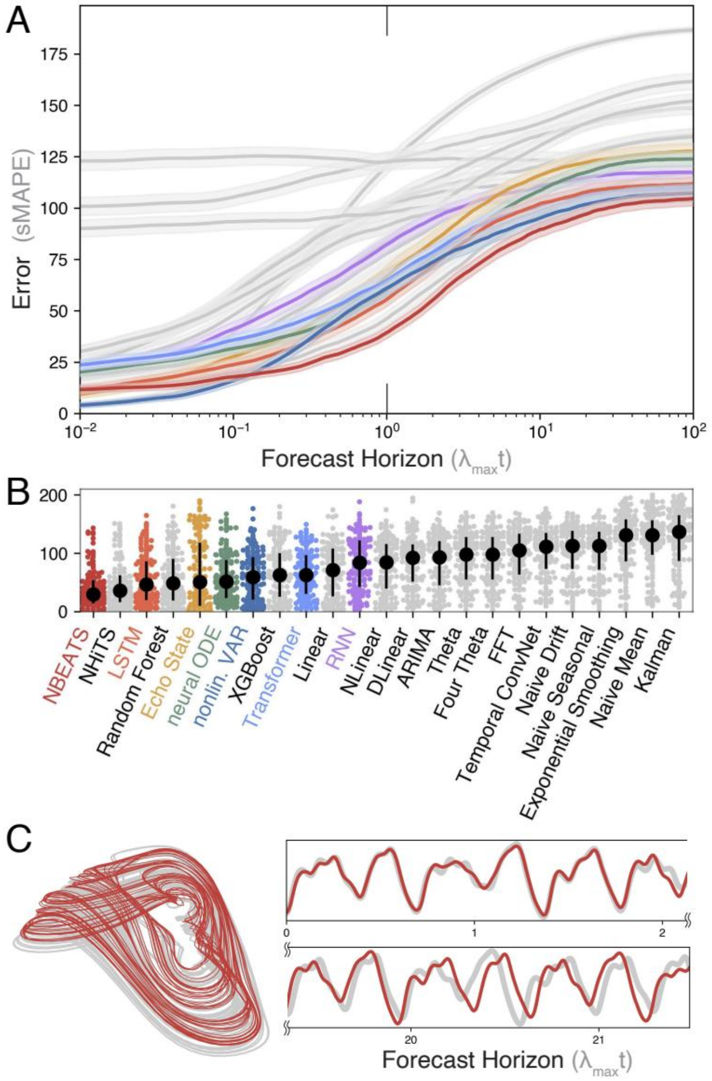


Figure 2. **Statistical forecasting across an ensemble of chaotic systems.** (A) The average error of 24 forecasting methods $\langle \epsilon_{ik}(t) \rangle_k$ as a function of Lyapunov time, averaged across 135 distinct chaotic systems. Colors denote high-performing models with properties of particular interest. (B) Distributions of the forecast errors when $t = \lambda_{\max}^{-1}$. (C) The predictions of the best-performing forecast model (red), relative to a held-out true trajectory from the Mackey-Glass model (gray) at short and long forecasting horizons.

Results: Figure 3

- Discuss:
  A. Larger models with good performance required longer training
  B. Best correlations at 1 λ_{max}^{-1}
     - Long enough for invariant properties to matter, not too long to be less predictable
     - Invariant properties not tell-all
  C. Physics-based models showed quicker initial improvement with limited training data
     - Potential strength of physics-based models

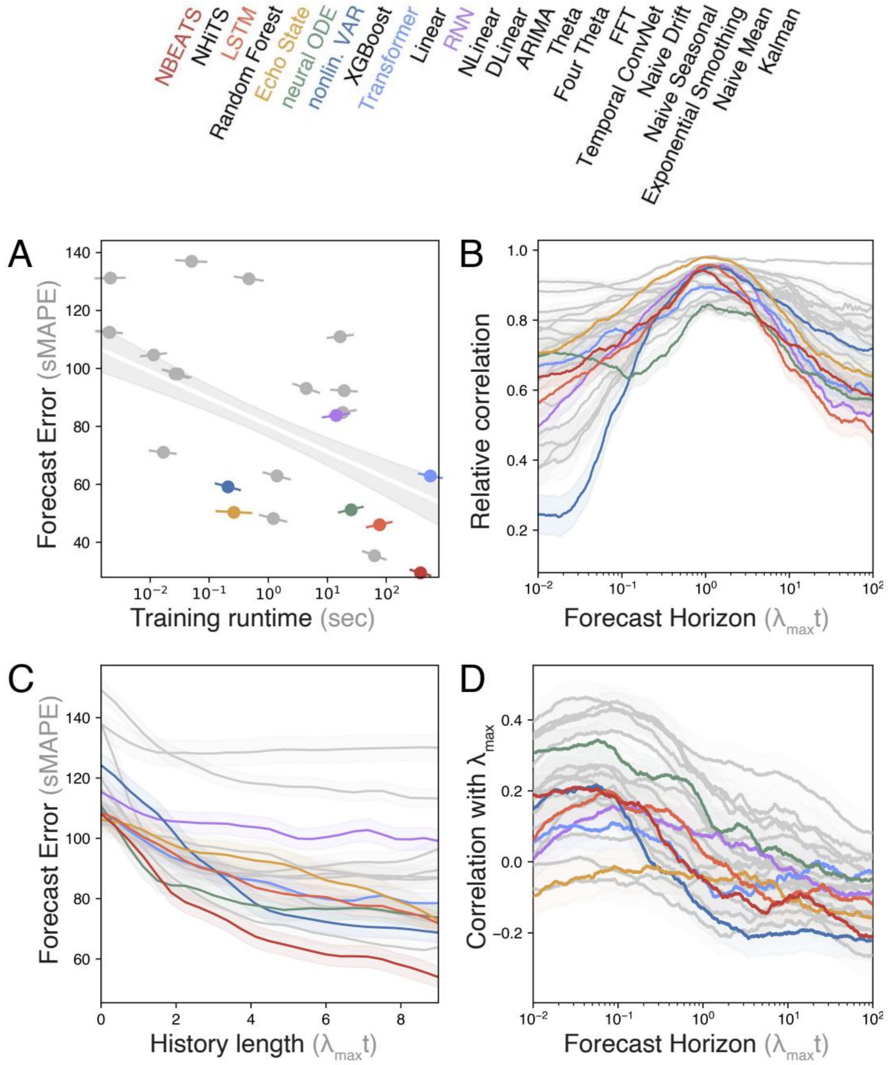


Figure 3. **Universal relationships among forecasting methods.** (A) Error versus training time at fixed forecast horizon $t = \lambda_{\max}^{-1}$ for all models. Bar lengths denote standard deviations along principal axes, with angle indicating Spearman correlation within each model group in order to detect Simpson's paradox. The underlaid linear fit indicates the overall correlation $\rho = -0.31 \pm 0.04$. (B) Median relative correlation of each forecasting method with its average prediction, across different forecast horizons. (C) Median model errors at $t = \lambda_{\max}^{-1}$ as the amount of history data increases. (D) Correlation of forecasting error with Lyapunov exponent $\lambda_{\max}$ as a function of forecasting horizon. All error bars correspond to 95% confidence intervals, and colors match methods from previous figures.

Conclusions

- Large domain-agnostic models better with extensive training data, physical models better with limited training data
- No free lunch
- Where do atmospheric science problems fit?
    - Lots of data: high frequency data in time and space, weather forecasting?
    - Limited data: extremes, seasonal/yearly/global averages
- Do we agree with their conclusions?