AI Weather Forecasting: Assessing Extrapolation and Physical Consistency

Sebastian Engelke
www.sengelke.com

Joint work with Nicola Gnecco, Manuel Hentschel, Marco Froelich,
Zhongwei Zhang, Erich Fischer, Jakob Zscheischler

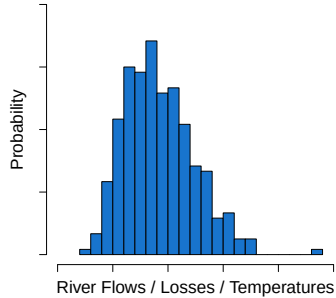
EXALT Workshop, Louvain-la-Neuve
May 27, 2026

Extreme Value Theory and Statistics

- Analysis of rare phenomena with small probabilities
- Impact on various risks (health, environment, economy,...)

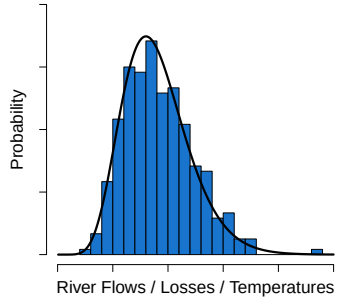
Extreme Value Theory and Statistics

- Analysis of rare phenomena with small probabilities
- Impact on various risks (health, environment, economy,...)



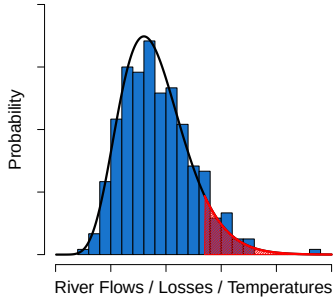
Extreme Value Theory and Statistics

- Analysis of rare phenomena with small probabilities
- Impact on various risks (health, environment, economy,...)

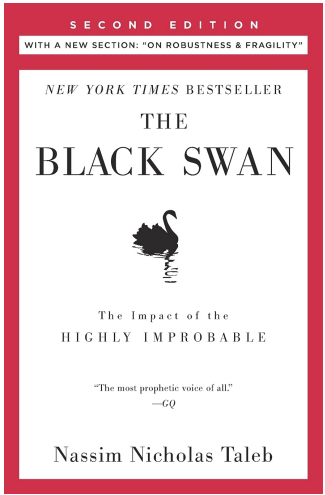


Extreme Value Theory and Statistics

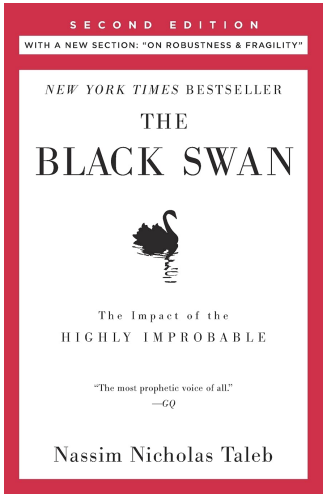
- Analysis of rare phenomena with small probabilities
- Impact on various risks (health, environment, economy,...)



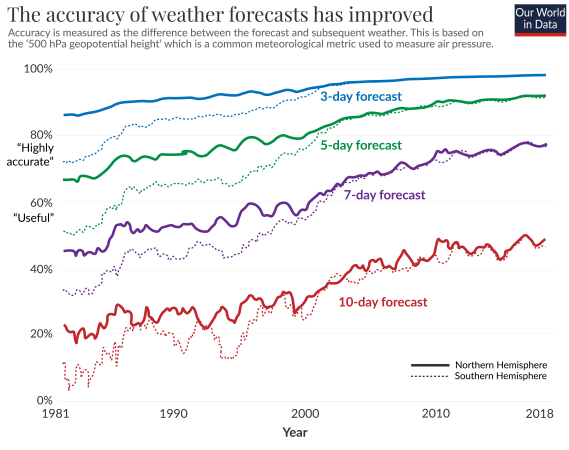
The "Black Swan"



The "Black Swan"



Evolution of weather forecasts



Physical and AI weather models

The earth system is a (complicated) dynamical system

$$Z_{t+1} = M(Z_t),$$

where $M : \mathbb{R}^p \to \mathbb{R}^p$ describes evolution of atmospheric states in dimension $p \approx 10^6$

Physical and AI weather models

The earth system is a (complicated) dynamical system

$$Z_{t+1} = M(Z_t),$$

where $M : \mathbb{R}^p \to \mathbb{R}^p$ describes evolution of atmospheric states in dimension $p \approx 10^6$

Traditional numerical models

- Describe M through PDEs
- Mostly based on physical laws

Physical and AI weather models

The earth system is a (complicated) dynamical system

$$Z_{t+1} = M(Z_t),$$

where $M : \mathbb{R}^p \to \mathbb{R}^p$ describes evolution of atmospheric states in dimension $p \approx 10^6$

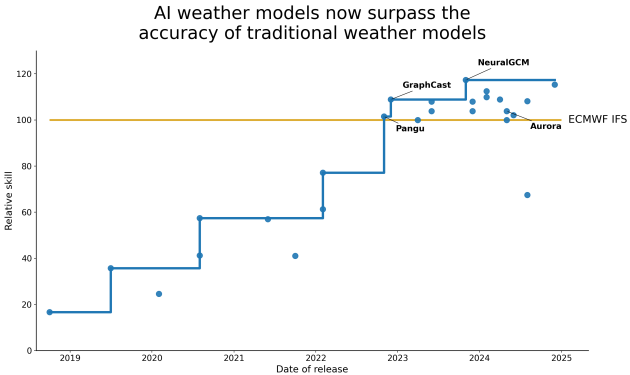
Traditional numerical models

- Describe M through PDEs
- Mostly based on physical laws

AI weather models

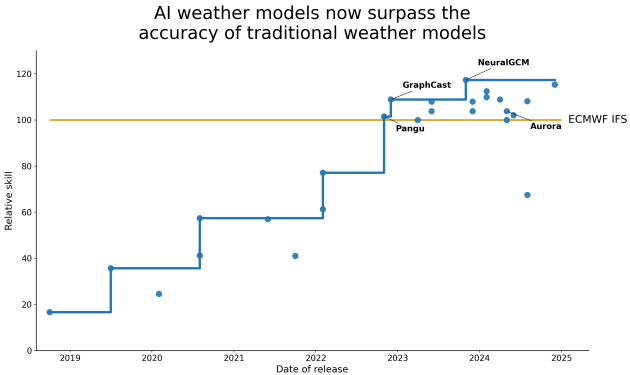
- Learn from past weather $(Z_t)_{t=1,\ldots,T}$ in a training period (usually 1979–2017) a parametric approximation $f_\theta \approx M$
- No physics involved – purely data driven

The AI revolution of weather forecasts



DeepMind	GraphCast, GenCast
Microsoft	ClimaX, Aurora
HUAWEI	Pangu-Weather
NVIDIA	FourCastNet

The AI revolution of weather forecasts

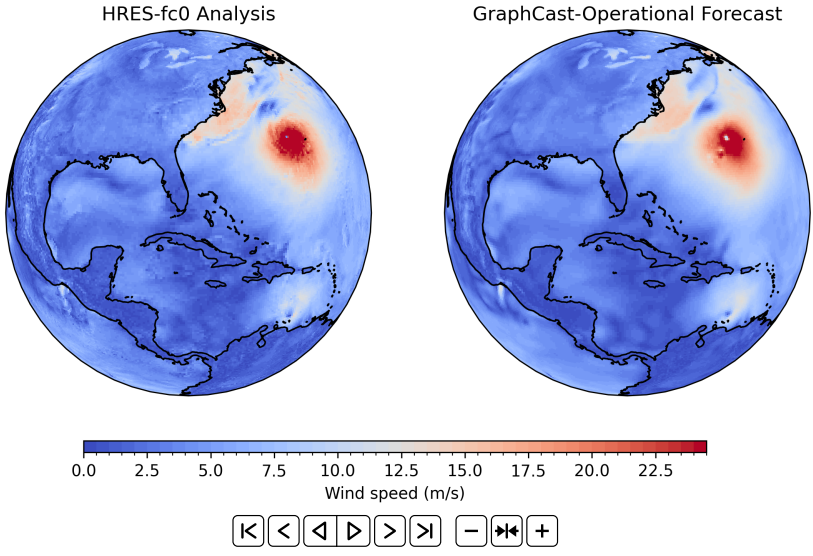


DeepMind	GraphCast, GenCast
Microsoft	ClimaX, Aurora
HUAWEI	Pangu-Weather
NVIDIA	FourCastNet

Advantages of AI weather models

- Extremely efficient compared to physical models
- Models are open-source and differentiable, which allows for generation of tailor-made events, fine-tuning to particular tasks, etc.

Hurricane-Ian, September 2022



Research questions

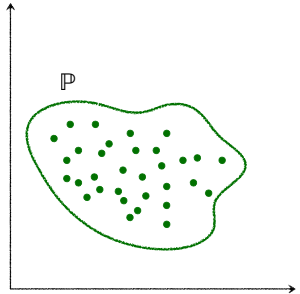
1. Can AI weather models extrapolate to unseen extreme events? (climate science)

2. Are AI forecasts physically realistic? (machine learning)

3. How to improve extrapolation in AI methods? (statistics)

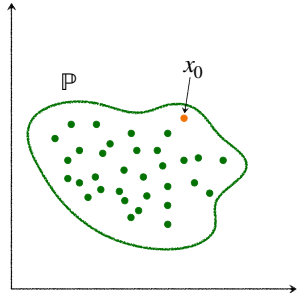
Can AI weather models extrapolate to unseen extreme events?

AI training data $\mathbb{P}$



[1] O. C. Pasche et al. "Validating Deep Learning Weather Forecast Models on Recent High-Impact Extreme Events". In: *Artificial Intelligence for the Earth Systems* 4.1 (2025), e240033.

AI training data $\mathbb{P}$

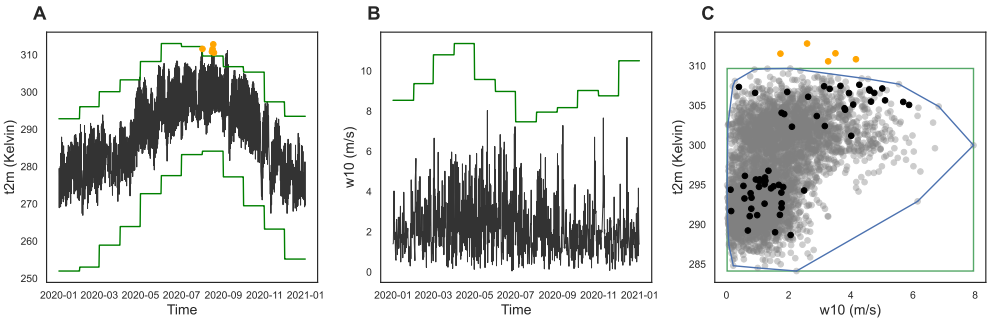


– In [1] we study performance on well-known high-impact events

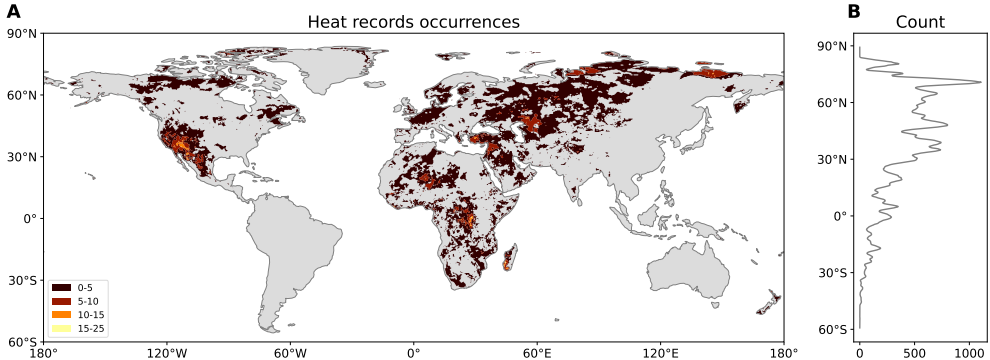
[1] O. C. Pasche et al. "Validating Deep Learning Weather Forecast Models on Recent High-Impact Extreme Events". In: *Artificial Intelligence for the Earth Systems* 4.1 (2025), e240033.

Record definition and extrapolation

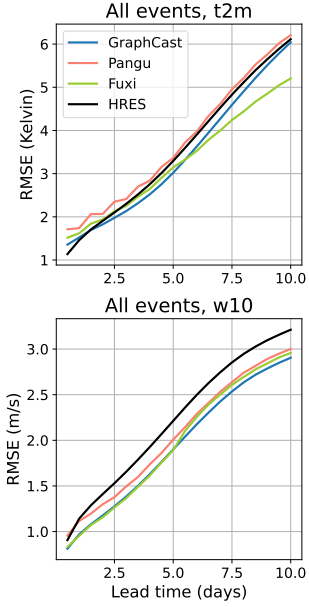
- Define monthly record on training period (1979–2017)
- Select record-breaking events per month (here August) in test period (2020)



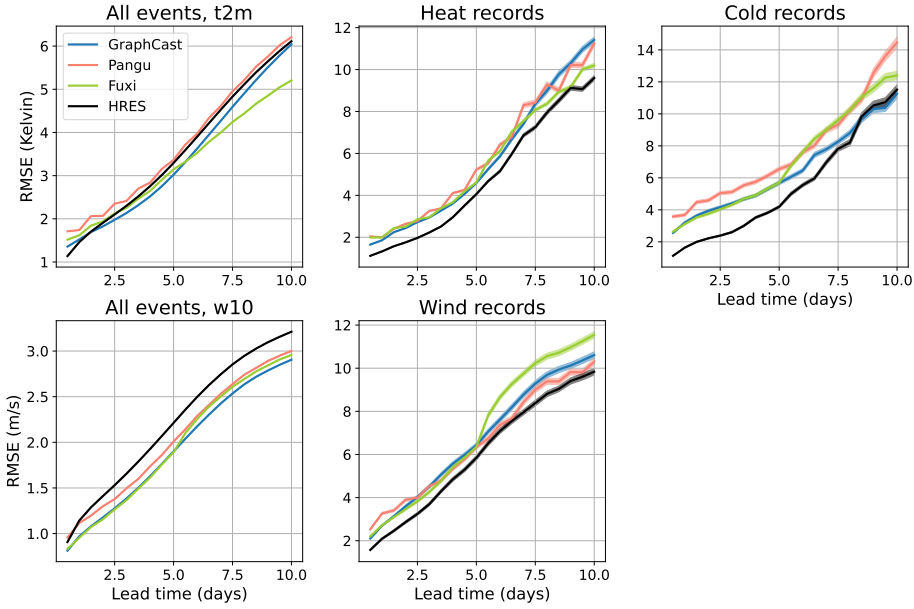
Heat records in 2020



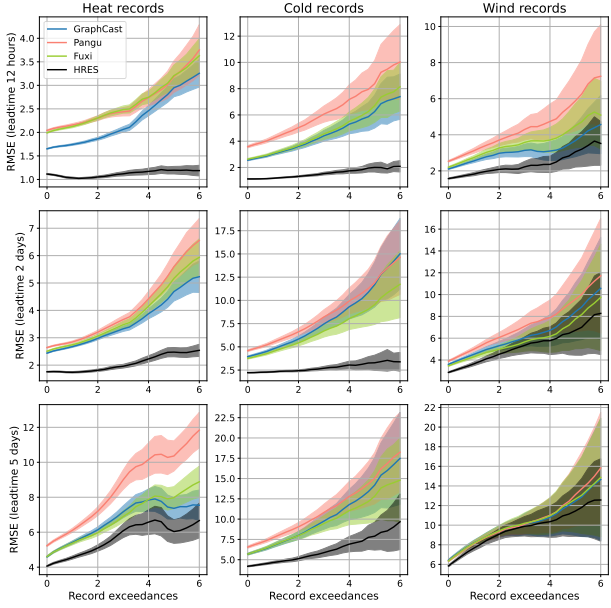
Quantitative comparison



Quantitative comparison



Record exceedances



Extrapolation of AI forecasts

– In [2] we find that AI models do not (yet) outperform physical models for record-breaking events

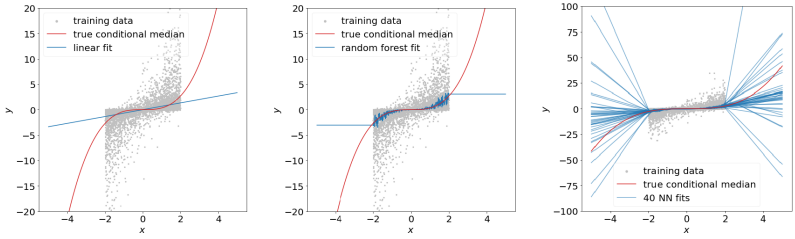
[2] Z. Zhang, E. Fischer, J. Zscheischler, and S. Engelke. "Physics-based models outperform AI weather forecasts of record-breaking extremes". In: *Science Advances* 12.18 (2026), eaec1433.

[3] X. Shen and N. Meinshausen. "Engression: extrapolation through the lens of distributional regression". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* (2024). Forthcoming.

[4] S. Engelke, N. Gnecco, and A. Sabourin. *Extrapolation in Statistical Learning with Extreme Value Theory*. 2026. URL: https://arxiv.org/abs/2605.01909.

Extrapolation of AI forecasts

- In [2] we find that AI models do not (yet) outperform physical models for record-breaking events
- This phenomenon is related to the lack of extrapolation of statistical/AI models (e.g., [3])
- Recent review on this topic on solutions with extreme value theory [4]



(a) linear model (b) tree-based model (c) neural network

[2] Z. Zhang, E. Fischer, J. Zscheischler, and S. Engelke. "Physics-based models outperform AI weather forecasts of record-breaking extremes". In: *Science Advances* 12.18 (2026), eaec1433.

[3] X. Shen and N. Meinshausen. "Engression: extrapolation through the lens of distributional regression". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* (2024). Forthcoming.

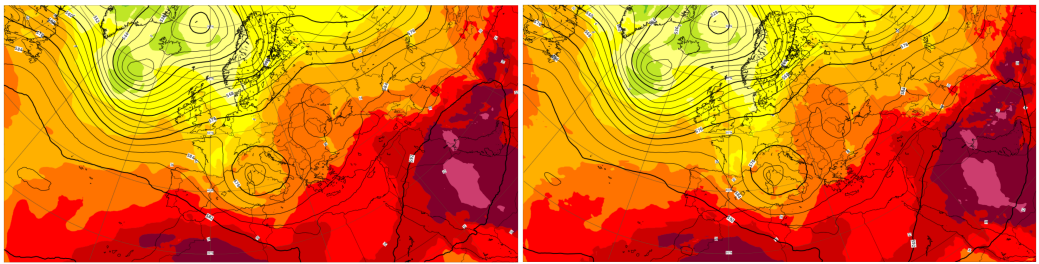
[4] S. Engelke, N. Gnecco, and A. Sabourin. *Extrapolation in Statistical Learning with Extreme Value Theory*. 2026. URL: https://arxiv.org/abs/2605.01909.

Are AI forecasts physically realistic?

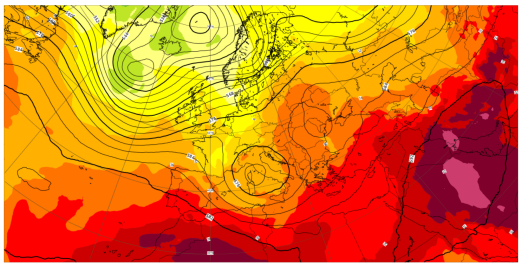
Are AI forecasts physically realistic?

A Turing test for extrapolation

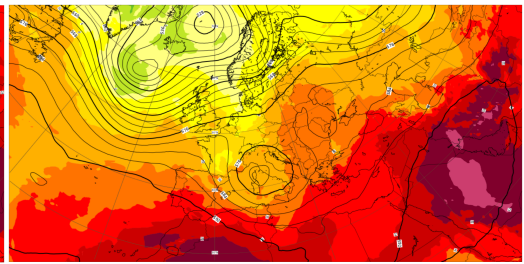
Turing test: Which of the two is the AI forecast?



Turing test: Which of the two is the AI forecast?

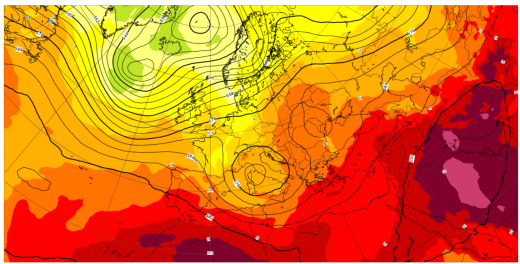


(AI)

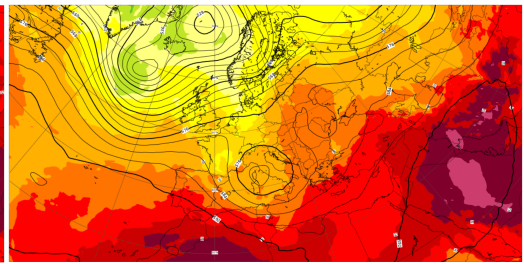


(physical)

Turing test: Which of the two is the AI forecast?



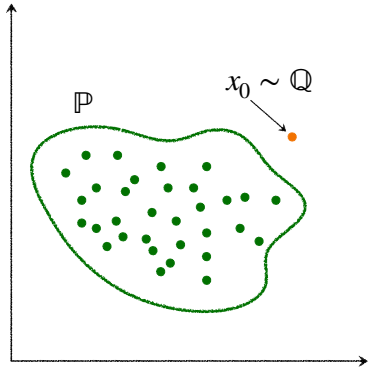
(AI)



(physical)

- Ad-hoc verification of physicality of AI outputs is difficult/impossible
- We formalize the Turing test to detect extrapolation and non-physical outputs

Non-physical states



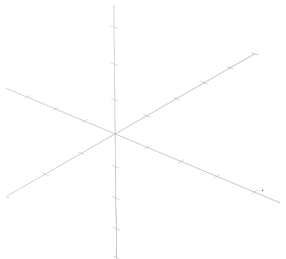
- Can we detect non-physical points $x_0 \sim \mathbb{Q}$ outside of the training distribution?
- Here, $\mathbb{P}$ denotes physical states from the training distribution (1979–2017)

Lorenz attractor

- As toy model, we use Lorenz 63 model [5]
- It is a chaotic dynamical system $z(t) \in \mathbb{R}^3$ described by the ODEs

$$\frac{dz_1}{dt} = 10(z_2 - z_1), \quad \frac{dz_2}{dt} = z_1(28 - z_3) - z_2, \quad \frac{dz_3}{dt} = z_1 z_2 - 8/3 z_3$$

  with an invariant ergodic measure $\mathbb{P}$
- Can be seen as simplified model of atmospheric convection



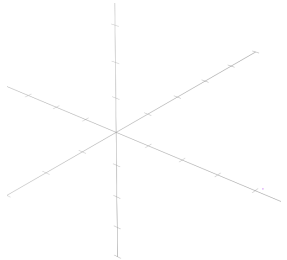
[5] E. N. Lorenz. "Deterministic Nonperiodic Flow". In: *Journal of Atmospheric Sciences* 20.2 (1963), pp. 130–141.

AI Lorenz model

- Let $Z_1, \ldots, Z_n$ be a discrete trajectory of the Lorenz attractor (can be seen as sample of $\mathbb{P}$)
- Train a neural network $\hat{f} : \mathbb{R}^3 \to \mathbb{R}^3$ to the Lorenz dynamics with training data

$$\mathcal{T}_n = \{(X_t, Y_t)\}_{t=1}^{n-1}, \quad X_t = Z_t,\ Y_t = Z_{t+1}$$

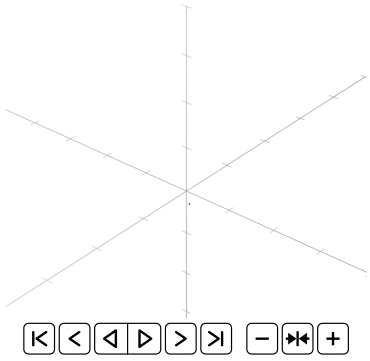
- Then $\hat{f}$ approximates the discretized Lorenz update $Z_{t+1} = f(Z_t)$ well for $Z_t \in \mathbb{R}^3$ sampled from $\mathbb{P}$ (e.g., [6])

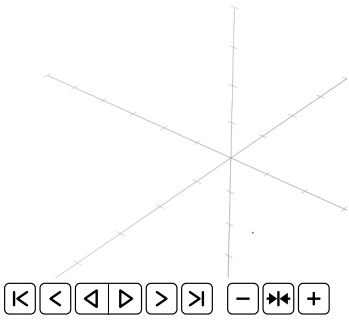


[6] J. Schmidt-Hieber. "Nonparametric regression using deep neural networks with ReLU activation function". In: *The Annals of Statistics* 48.4 (2020), pp. 1875–1897.

Epistemic uncertainty

– We train an ensemble of neural networks $\hat{f}^{(1)}, \ldots, \hat{f}^{(B)}$ on $\mathcal{T}_n$

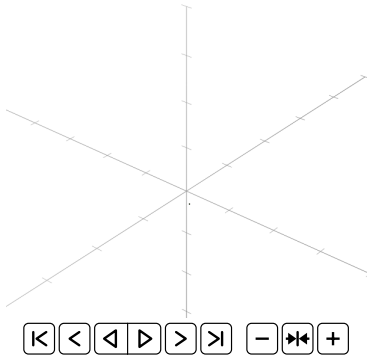


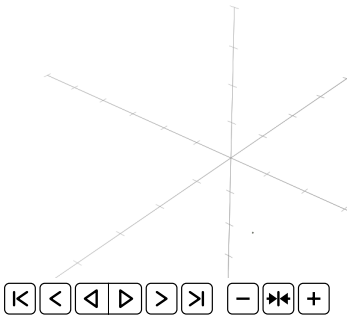


[7] B. Lakshminarayanan, A. Pritzel, and C. Blundell. "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles". In: *Advances in Neural Information Processing Systems (NeurIPS)*. vol. 30. 2017.

Epistemic uncertainty

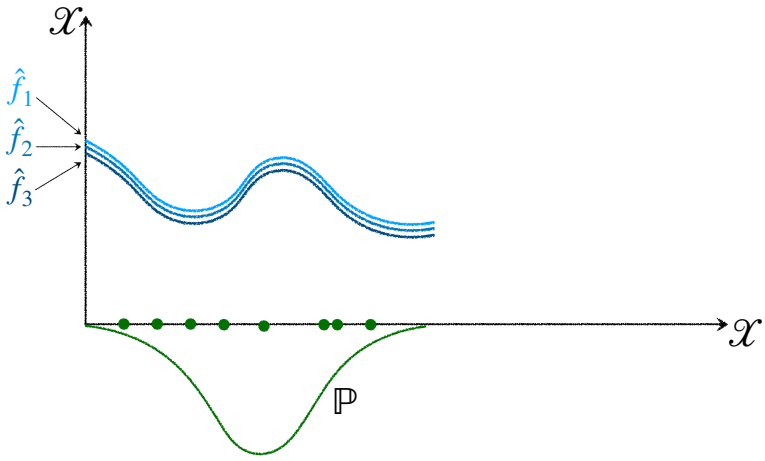
- We train an ensemble of neural networks $\hat{f}^{(1)}, \ldots, \hat{f}^{(B)}$ on $\mathcal{T}_n$
- Epistemic uncertainty of neural networks is leveraged to assess prediction uncertainty in deep ensembles [7]



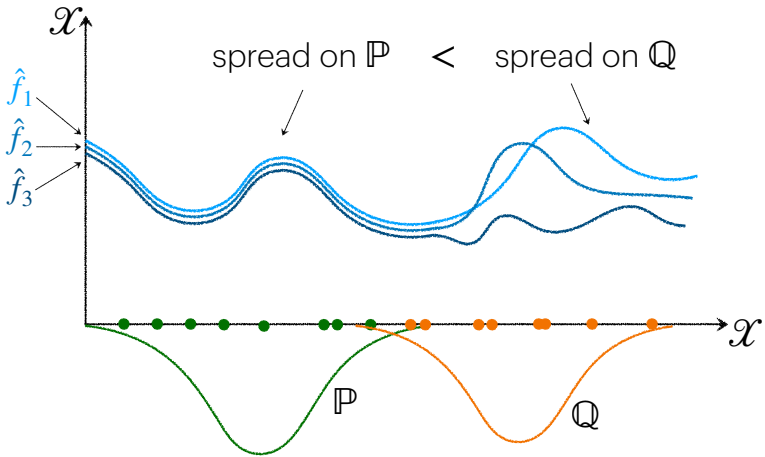


[7] B. Lakshminarayanan, A. Pritzel, and C. Blundell. "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles". In: *Advances in Neural Information Processing Systems (NeurIPS)*. vol. 30. 2017.

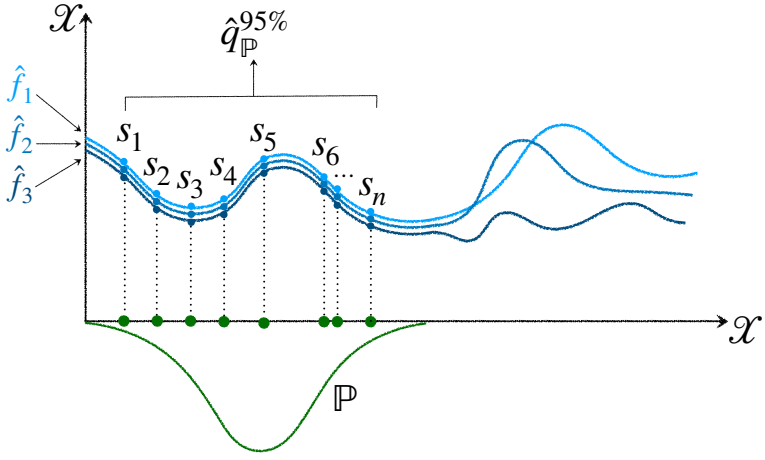
A Turing test for extrapolation via conformal prediction



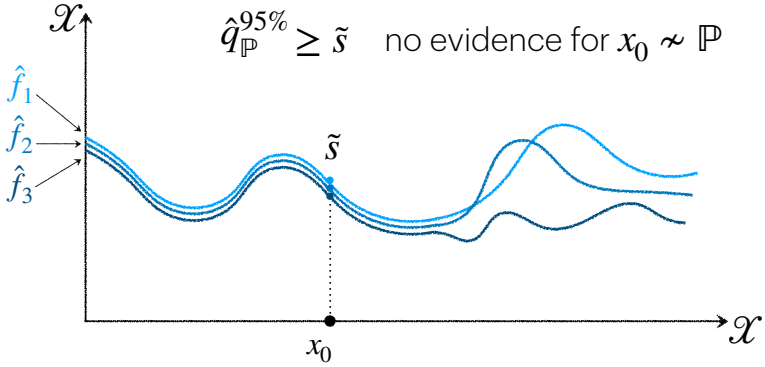
A Turing test for extrapolation via conformal prediction



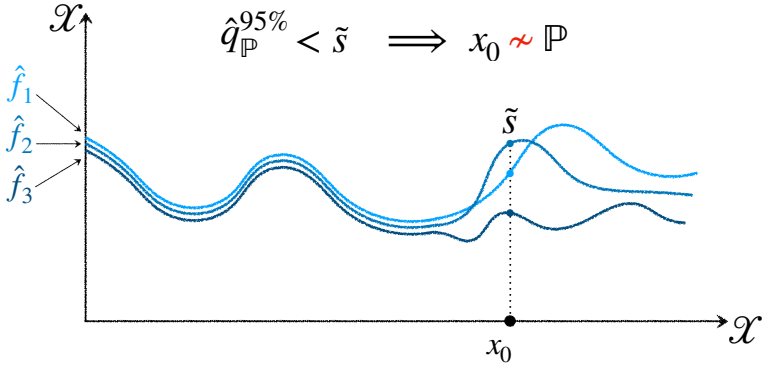
A Turing test for extrapolation via conformal prediction



A Turing test for extrapolation via conformal prediction



A Turing test for extrapolation via conformal prediction



A Turing test for extrapolation

- Let $\hat{f}^{(1)}, \ldots, \hat{f}^{(B)}$ be black-box models (pre-)trained on (possibly the same) data from $\mathbb{P}$
- They form the panel (or ensemble) that decides whether a new point is OOD and requires extrapolation
- Statistically, given an i.i.d. calibration data set of predictors $X'_1, \ldots, X'_m$ from $\mathbb{P}$, we test the null hypothesis

$$H_0 : X'_{m+1} \sim \mathbb{P},$$

where X'_{m+1} is an independent sample

Conformal prediction intervals

- Recall i.i.d. calibration data set $X'_1, \ldots, X'_m$ (independent of training data)
- Define non-conformity scores

$$S_t = \frac{1}{2B^2} \sum_{i=1}^{B} \sum_{j=1}^{B} \|\hat{f}^{(i)}(X'_t) - \hat{f}^{(j)}(X'_t)\|^2, \quad t = 1, \ldots, m$$

[8] H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman. "Inductive Confidence Machines for Regression". In: *Mach. Learn.: ECML 2002*. Springer, 2002, pp. 345–356.

Conformal prediction intervals

- Recall i.i.d. calibration data set $X'_1, \ldots, X'_m$ (independent of training data)
- Define non-conformity scores

$$S_t = \frac{1}{2B^2} \sum_{i=1}^{B} \sum_{j=1}^{B} \|\hat{f}^{(i)}(X'_t) - \hat{f}^{(j)}(X'_t)\|^2, \quad t = 1, \ldots, m$$

- For a nominal type-I error $\alpha \in (0, 1)$, define

$$\hat{q} := \text{Quantile}\left(\frac{\lceil (1-\alpha)(m+1) \rceil}{m}; \frac{1}{m} \sum_{t=1}^{m} \delta_{S_t} \right)$$

[8] H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman. "Inductive Confidence Machines for Regression". In: *Mach. Learn.: ECML 2002*. Springer, 2002, pp. 345–356.

Conformal prediction intervals

- Recall i.i.d. calibration data set $X'_1, \ldots, X'_m$ (independent of training data)
- Define non-conformity scores

$$S_t = \frac{1}{2B^2} \sum_{i=1}^{B} \sum_{j=1}^{B} \|\hat{f}^{(i)}(X'_t) - \hat{f}^{(j)}(X'_t)\|^2, \quad t = 1, \ldots, m$$

- For a nominal type-I error $\alpha \in (0, 1)$, define

$$\hat{q} \coloneqq \text{Quantile}\left(\frac{\lceil (1-\alpha)(m+1) \rceil}{m}; \frac{1}{m} \sum_{t=1}^{m} \delta_{S_t} \right)$$

- For a new predictor value X'_{m+1}, define prediction interval $C(X'_{m+1}) = [0, \hat{q})$ and reject $H_0 : X'_{n+1} \sim \mathcal{P}_0$ if

$$\{S_{m+1} \notin C(X'_{m+1})\} = \{S_{m+1} > \hat{q}\}$$

[8] H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman. "Inductive Confidence Machines for Regression". In: *Mach. Learn.: ECML 2002*. Springer, 2002, pp. 345–356.

Conformal prediction intervals

- Recall i.i.d. calibration data set $X'_1, \ldots, X'_m$ (independent of training data)
- Define non-conformity scores

$$S_t = \frac{1}{2B^2} \sum_{i=1}^{B} \sum_{j=1}^{B} \|\hat{f}^{(i)}(X'_t) - \hat{f}^{(j)}(X'_t)\|^2, \quad t = 1, \ldots, m$$

- For a nominal type-I error $\alpha \in (0, 1)$, define

$$\hat{q} \coloneqq \text{Quantile}\left(\frac{\lceil (1-\alpha)(m+1) \rceil}{m}; \frac{1}{m} \sum_{t=1}^{m} \delta_{S_t} \right)$$

- For a new predictor value X'_{m+1}, define prediction interval $C(X'_{m+1}) = [0, \hat{q})$ and reject $H_0 : X'_{n+1} \sim \mathcal{P}_0$ if

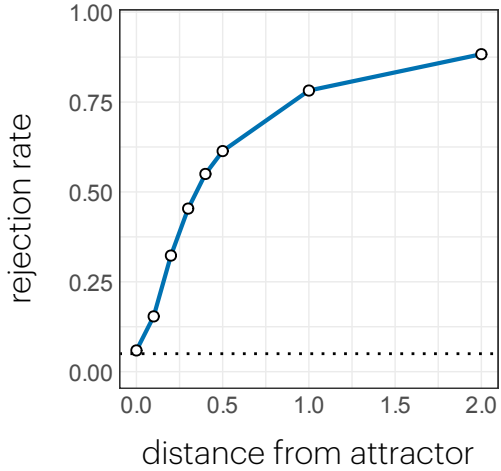
$$\{S_{m+1} \notin C(X'_{m+1})\} = \{S_{m+1} > \hat{q}\}$$

- Conformal inference [8] guarantees this test to have correct type-I error (marginal coverage)

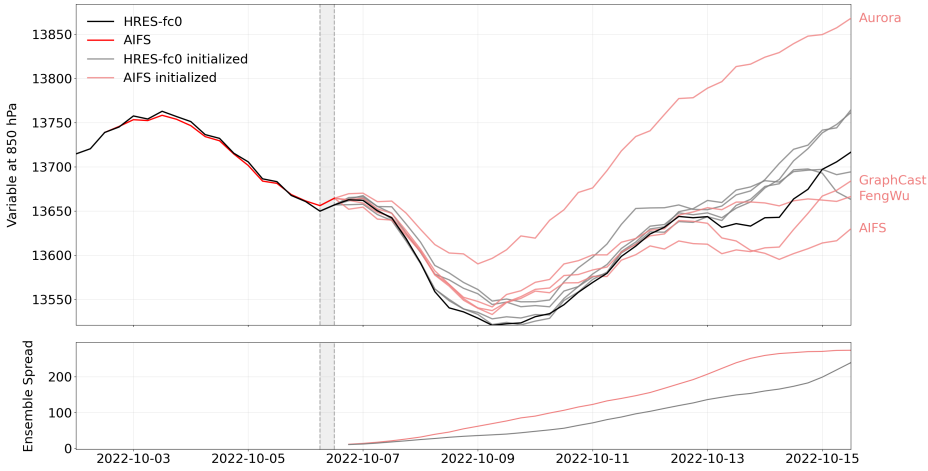
$$\mathbb{P}(S_{m+1} \in C(X'_{m+1})) \geq 1 - \alpha, \qquad \text{for all } X'_{m+1} \sim \mathbb{P}$$

[8] H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman. "Inductive Confidence Machines for Regression". In: *Mach. Learn.: ECML 2002*. Springer, 2002, pp. 345–356.

Detection power on Lorenz attractor



Outlook: Are AI forecasts physically realistic?



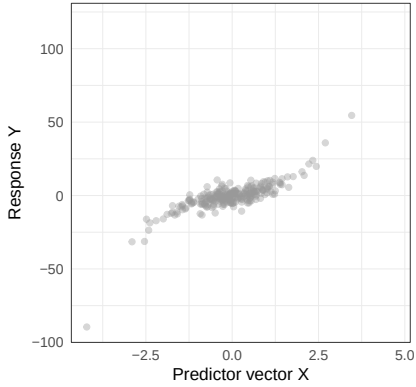
- Initialize the jury/ensemble with output of 3-day forecast of AIFS (orange)
- Compare the ensemble spread with the one of jury initialized with ground truth (grey)

How to improve extrapolation in AI methods?

How to improve extrapolation in AI methods?

The progression method

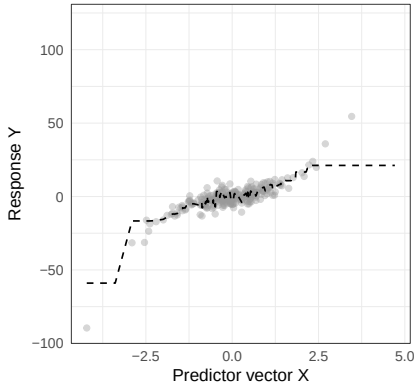
Beyond data range: out-of-distribution generalization



– Prediction of conditional mean $f(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$ for $\mathbf{x} \in \mathbb{R}^p$

[9] G. Buriticá and S. Engelke. *Progression: an extrapolation principle for regression*. 2024. URL: https://arxiv.org/abs/2410.23246.

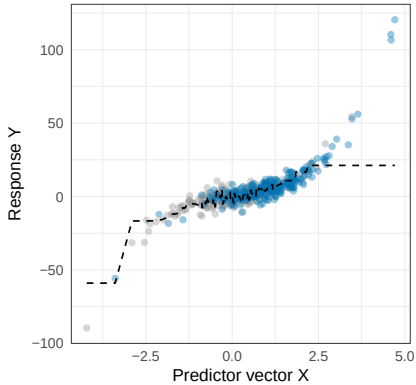
Beyond data range: out-of-distribution generalization



- Prediction of conditional mean $f(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$ for $\mathbf{x} \in \mathbb{R}^p$
- Machine learning methods typically perform poorly outside of the range of training distribution

[9] G. Buriticá and S. Engelke. *Progression: an extrapolation principle for regression*. 2024. URL: https://arxiv.org/abs/2410.23246.

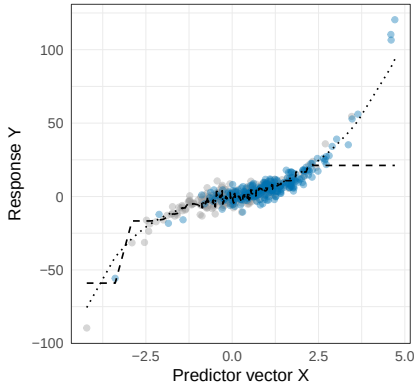
Beyond data range: out-of-distribution generalization



- Prediction of conditional mean $f(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$ for $\mathbf{x} \in \mathbb{R}^p$
- Machine learning methods typically perform poorly outside of the range of training distribution
- If test distribution of $\mathbf{X}$ is different, out-of-distribution generalization is needed

[9] G. Buriticá and S. Engelke. *Progression: an extrapolation principle for regression*. 2024. URL: https://arxiv.org/abs/2410.23246.

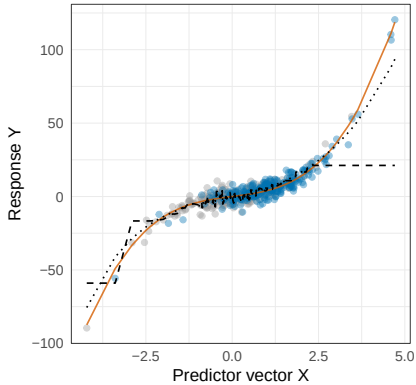
Beyond data range: out-of-distribution generalization



- Prediction of conditional mean $f(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$ for $\mathbf{x} \in \mathbb{R}^p$
- Machine learning methods typically perform poorly outside of the range of training distribution
- If test distribution of $\mathbf{X}$ is different, out-of-distribution generalization is needed
- We work on methods based on extreme value theory

[9] G. Buriticá and S. Engelke. *Progression: an extrapolation principle for regression*. 2024. URL: https://arxiv.org/abs/2410.23246.

Beyond data range: out-of-distribution generalization



- Prediction of conditional mean $f(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$ for $\mathbf{x} \in \mathbb{R}^p$
- Machine learning methods typically perform poorly outside of the range of training distribution
- If test distribution of $\mathbf{X}$ is different, out-of-distribution generalization is needed
- We work on methods based on extreme value theory

[9] G. Buriticá and S. Engelke. *Progression: an extrapolation principle for regression*. 2024. URL: https://arxiv.org/abs/2410.23246.

Beyond data range: out-of-distribution generalization

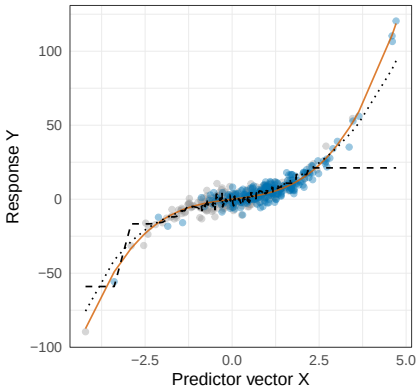
– Assume the regression model

$$Y = f(\mathbf{X}, \varepsilon),$$

with regression function $f : \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}$

– Training data: n independent observations

$$(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n) \sim \mathbb{P}_{\text{train}}$$



Beyond data range: out-of-distribution generalization

- Assume the regression model

$$Y = f(\mathbf{X}, \varepsilon),$$

  with regression function $f : \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}$
- Training data: n independent observations

$$(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n) \sim \mathbb{P}_{\text{train}}$$

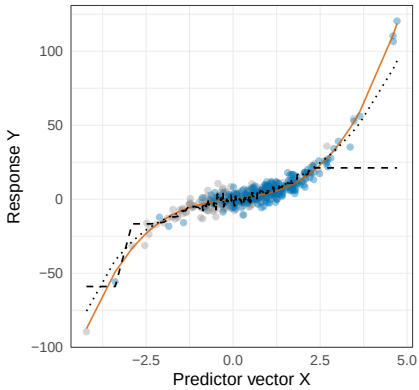
- At test time, predict response at a new point $\mathbf{x} \sim \mathbb{P}_{\text{test}}$ by conditional mean or median

$$m(\mathbf{x}) := \text{median}(Y \mid \mathbf{X} = \mathbf{x})$$

- Example: for additive noise

$$Y = f(\mathbf{X}) + \varepsilon$$

  with $\text{median}(\varepsilon) = 0$ we have $f(\mathbf{x}) = m(\mathbf{x})$



Beyond data range: out-of-distribution generalization

– Machine learning methods are strong at interpolating if $\mathbf{x}$ is inside the data range, but bad at extrapolating if $\mathbf{x}$ is beyond the training data range

[10] Z. Zhang, E. Fischer, J. Zscheischler, and S. Engelke. "Physics-based models outperform AI weather forecasts of record-breaking extremes". In: *Science Advances* 12.18 (2026), eaec1433.

Beyond data range: out-of-distribution generalization

- Machine learning methods are strong at interpolating if $\mathbf{x}$ is inside the data range, but bad at extrapolating if $\mathbf{x}$ is beyond the training data range
- The latter happens in many practical cases, e.g.,
  - if $\mathbb{P}_{\text{test}} = \mathbb{P}_{\text{train}}$ by stochasticity, $\mathbf{x}$ may be outside of training data range;
  - if $\mathbb{P}_{\text{test}} \neq \mathbb{P}_{\text{train}}$ because of a mean/variance shift in the predictors, then $\mathbf{x}$ is likely to be far from training predictors.

[10] Z. Zhang, E. Fischer, J. Zscheischler, and S. Engelke. "Physics-based models outperform AI weather forecasts of record-breaking extremes". In: *Science Advances* 12.18 (2026), eaec1433.

Beyond data range: out-of-distribution generalization

- Machine learning methods are strong at interpolating if $\mathbf{x}$ is inside the data range, but bad at extrapolating if $\mathbf{x}$ is beyond the training data range
- The latter happens in many practical cases, e.g.,
  - if $\mathbb{P}_{\text{test}} = \mathbb{P}_{\text{train}}$ by stochasticity, $\mathbf{x}$ may be outside of training data range;
  - if $\mathbb{P}_{\text{test}} \neq \mathbb{P}_{\text{train}}$ because of a mean/variance shift in the predictors, then $\mathbf{x}$ is likely to be far from training predictors.
- Example:
  - $\mathbf{X}$ climatological variables (e.g., temperature, precipitation, etc.)
  - Y an impact (e.g., crop yield, heat wave) [10]
  - $\mathbb{P}_{\text{train}}$: current climate; $\mathbb{P}_{\text{test}}$: future climate, where $\mathbf{X}$ has distributional shift (climate change)

[10] Z. Zhang, E. Fischer, J. Zscheischler, and S. Engelke. "Physics-based models outperform AI weather forecasts of record-breaking extremes". In: *Science Advances* 12.18 (2026), eaec1433.

Existing work

Parametric methods

- Put structure on function class (e.g., polynomials)
- Assume causal structure between predictors and response (e.g., [11])
- This requires domain knowledge!

[11] N. Gnecco, J. Peters, S. Engelke, and N. Pfister. *Boosted Control Functions: Distribution generalization and invariance in confounded models*. 2024. URL: https://arxiv.org/abs/2310.05805.
[12] N. Pfister and P. Bühlmann. "Extrapolation-Aware Nonparametric Statistical Inference". In: *arXiv:2402.09758* (2024).
[13] X. Shen and N. Meinshausen. "Engression: extrapolation through the lens of distributional regression". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* (2024). Forthcoming.

Existing work

Parametric methods

- Put structure on function class (e.g., polynomials)
- Assume causal structure between predictors and response (e.g., [11])
- This requires domain knowledge!

Non-parametric methods

- Extrapolation-aware bounds outside of training range by Taylor expansion [12]
- Assumption of pre-additive noise models

$$Y = g(X + \eta)$$

  in combination with distributional regression [13]

---

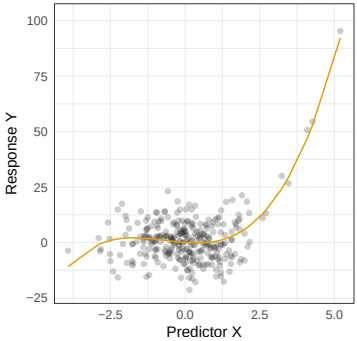
[11] N. Gnecco, J. Peters, S. Engelke, and N. Pfister. *Boosted Control Functions: Distribution generalization and invariance in confounded models*. 2024. URL: https://arxiv.org/abs/2310.05805.

[12] N. Pfister and P. Bühlmann. "Extrapolation-Aware Nonparametric Statistical Inference". In: *arXiv:2402.09758* (2024).

[13] X. Shen and N. Meinshausen. "Engression: extrapolation through the lens of distributional regression". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* (2024). Forthcoming.

Extrapolation principle for regression

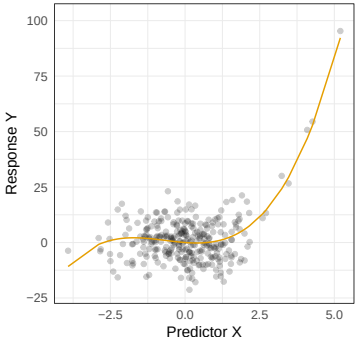
- Assume for now that $p = 1$, i.e., there is only one predictor X
- The relationship between X and Y can be arbitrarily complex on the original data scale



[14] J. E. Heffernan and J. A. Tawn. "A conditional approach for multivariate extreme values (with discussion)". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 66.3 (2004), pp. 497–546.

Extrapolation principle for regression

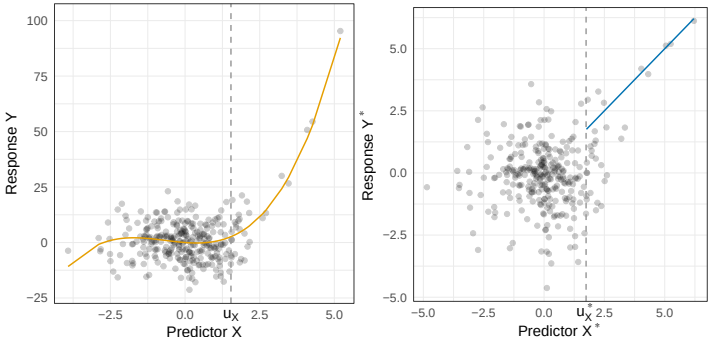
- Assume for now that $p = 1$, i.e., there is only one predictor X
- The relationship between X and Y can be arbitrarily complex on the original data scale
- Extreme value theory: dependence in the tails simplifies in suitable margins [14]



[14] J. E. Heffernan and J. A. Tawn. "A conditional approach for multivariate extreme values (with discussion)". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 66.3 (2004), pp. 497–546.

Extrapolation principle for regression

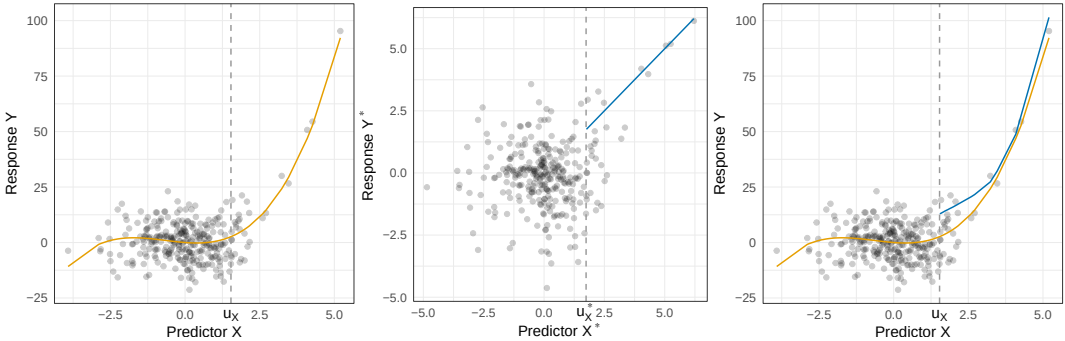
- Assume for now that $p = 1$, i.e., there is only one predictor X
- The relationship between X and Y can be arbitrarily complex on the original data scale
- Extreme value theory: dependence in the tails simplifies in suitable margins [14]
- If predictor and response are transformed to Laplace margins (X^*, Y^*), then the conditional median is approximately linear for large predictor values



[14] J. E. Heffernan and J. A. Tawn. "A conditional approach for multivariate extreme values (with discussion)". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 66.3 (2004), pp. 497–546.

Extrapolation principle for regression

- Assume for now that $p = 1$, i.e., there is only one predictor X
- The relationship between X and Y can be arbitrarily complex on the original data scale
- Extreme value theory: dependence in the tails simplifies in suitable margins [14]
- If predictor and response are transformed to Laplace margins (X^*, Y^*), then the conditional median is approximately linear for large predictor values



[14] J. E. Heffernan and J. A. Tawn. "A conditional approach for multivariate extreme values (with discussion)". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 66.3 (2004), pp. 497–546.

Extrapolation principle for regression

– Consider the additive noise model

$$Y = f(X) + \varepsilon, \qquad \text{median}(\varepsilon) = 0$$

Extrapolation principle for regression

- Consider the additive noise model

$$Y = f(X) + \varepsilon, \qquad \text{median}(\varepsilon) = 0$$

- We transform predictor and response to Laplace margins:

$$X^* = Q_L \circ F_X(X), \qquad Y^* = Q_L \circ F_Y(Y)$$

Extrapolation principle for regression

- Consider the additive noise model

$$Y = f(X) + \varepsilon, \qquad \text{median}(\varepsilon) = 0$$

- We transform predictor and response to Laplace margins:

$$X^* = Q_L \circ F_X(X), \qquad Y^* = Q_L \circ F_Y(Y)$$

- The conditional median can be written as

$$f(x) = \text{median}(Y \mid X = x) = Q_Y \circ F_L\{\text{median}(Y^* \mid X^* = x^*)\},$$

where $x^* = Q_L \circ F_Y(x)$ is the predictor of interest on Laplace scale

Extrapolation principle for regression

Main Assumption

For $a \in [-1, 1]$, $\beta \in [0, 1)$, $b \in \mathbb{R}$, the conditional median on Laplace scale is first-order linear

$$\text{median}(Y^* \mid X^* = x^*) = ax^* + (x^*)^\beta b + o(1), \quad x^* \to \infty$$

[15] G. Buriticá and S. Engelke. *Progression: an extrapolation principle for regression*. 2024. URL: https://arxiv.org/abs/2410.23246.

Extrapolation principle for regression

Main Assumption

For $a \in [-1, 1]$, $\beta \in [0, 1)$, $b \in \mathbb{R}$, the conditional median on Laplace scale is first-order linear

$$\text{median}(Y^* \mid X^* = x^*) = ax^* + (x^*)^\beta b + o(1), \quad x^* \to \infty$$

This motivates `progression` approximation of $m(x) = \text{median}(Y \mid X = x)$ above high quantile:

$$\tilde{m}_{\tau_0}(x) = Q_Y \circ F_L\{ax^* + (x^*)^\beta b\}, \quad x > Q_X(\tau_0),$$

where $x^* = Q_L \circ F_X(x)$

[15] G. Buriticá and S. Engelke. *Progression: an extrapolation principle for regression*. 2024. URL: https://arxiv.org/abs/2410.23246.

Extrapolation principle for regression

Main Assumption

For $a \in [-1, 1]$, $\beta \in [0, 1)$, $b \in \mathbb{R}$, the conditional median on Laplace scale is first-order linear

$$\text{median}(Y^* \mid X^* = x^*) = ax^* + (x^*)^\beta b + o(1), \quad x^* \to \infty$$

This motivates `progression` approximation of $m(x) = \text{median}(Y \mid X = x)$ above high quantile:

$$\tilde{m}_{\tau_0}(x) = \tilde{Q}_Y \circ F_L\{a\tilde{x}^* + (\tilde{x}^*)^\beta b\}, \quad x > \tilde{Q}_X(\tau_0),$$

where $\tilde{x}^* = Q_L \circ \tilde{F}_X(x)$ and $\tilde{F}_X$ and $\tilde{Q}_Y$ denote the GPD approximations of F_X and Q_Y

[15] G. Buriticá and S. Engelke. *Progression: an extrapolation principle for regression*. 2024. URL: https://arxiv.org/abs/2410.23246.

Extrapolation principle for regression

Main Assumption

For $a \in [-1, 1]$, $\beta \in [0, 1)$, $b \in \mathbb{R}$, the conditional median on Laplace scale is first-order linear

$$\text{median}(Y^* \mid X^* = x^*) = ax^* + (x^*)^\beta b + o(1), \quad x^* \to \infty$$

This motivates `progression` approximation of $m(x) = \text{median}(Y \mid X = x)$ above high quantile:

$$\tilde{m}_{\tau_0}(x) = \tilde{Q}_Y \circ F_L\{a\tilde{x}^* + (\tilde{x}^*)^\beta b\}, \quad x > \tilde{Q}_X(\tau_0),$$

where $\tilde{x}^* = Q_L \circ \tilde{F}_X(x)$ and $\tilde{F}_X$ and $\tilde{Q}_Y$ denote the GPD approximations of F_X and Q_Y

Theorem

Under the above assumption (+tail regularity of X and Y) we uniformly control the relative approximation error

$$\lim_{\tau_0 \to 1} \sup_{Q_X(\tau_0) < x < \nu(\tau_0)} \left| \frac{\tilde{m}_{\tau_0}(x)}{m(x)} - 1 \right| = 0,$$

where $\nu(\tau_0) \in [Q_X(\tau_0), \infty]$ is the limit of extrapolation.

---

[15] G. Buriticá and S. Engelke. *Progression: an extrapolation principle for regression*. 2024. URL: https://arxiv.org/abs/2410.23246.

When is 'Main Assumption' satisfied?

– Consider the additive noise model

$$Y = f(X) + \varepsilon, \qquad \text{median}(\varepsilon) = 0$$

– Validity of extrapolation principle depends on
  - Regularity of regression function $f(x)$
  - Training distribution $\mathbb{P}_X$ of predictors
  - Training distribution $\mathbb{P}_\varepsilon$ of noise

[16] G. Buriticá and S. Engelke. *Progression: an extrapolation principle for regression*. 2024. URL: https://arxiv.org/abs/2410.23246.

When is 'Main Assumption' satisfied?

– Consider the additive noise model

$$Y = f(X) + \varepsilon, \qquad \text{median}(\varepsilon) = 0$$

– Validity of extrapolation principle depends on
  - Regularity of regression function $f(x)$
  - Training distribution $\mathbb{P}_X$ of predictors
  - Training distribution $\mathbb{P}_\varepsilon$ of noise

– Intuitively, it suffices that noise ε has lighter tail than signal $f(X)$

– For wide range of examples (also beyond additive noise), see Section 4.1 in [16]

[16] G. Buriticá and S. Engelke. *Progression: an extrapolation principle for regression*. 2024. URL: https://arxiv.org/abs/2410.23246.

Comments

- Progression = Principle of regression extrapolation
- Contrary to original regression [17], progression can extrapolate towards values more extreme than in the training data

[17] F. Galton. "Regression Towards Mediocrity in Hereditary Stature.". In: *The Journal of the Anthropological Institute of Great Britain and Ireland* 15 (1886), pp. 246–263.

[18] G. E. P. Box and D. R. Cox. "An Analysis of Transformations". In: *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 26.2 (1964), pp. 211–252.

Comments

- Progression = Principle of regression extrapolation
- Contrary to original regression [17], progression can extrapolate towards values more extreme than in the training data
- In the spirit of Box–Cox [18], relation between predictor and response simplifies after transformation
- We transform both response and predictor, and only assume parametric form in the tails

[17] F. Galton. "Regression Towards Mediocrity in Hereditary Stature.". In: *The Journal of the Anthropological Institute of Great Britain and Ireland* 15 (1886), pp. 246–263.

[18] G. E. P. Box and D. R. Cox. "An Analysis of Transformations". In: *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 26.2 (1964), pp. 211–252.

Random forest progression

- Transform data (X_i, Y_i) to Laplace scale (X_i^*, Y_i^*), $i = 1, \ldots, n$ using GPD approximations
- Let $w_i(x^*)$ be the localizing weights from a regression random forest at x^*

[19] R. Friedberg, J. Tibshirani, S. Athey, and S. Wager. "Local Linear Forest". In: *Journal of computational and graphical statistics* 30.2 (2021), pp. 503–517.

Random forest progression

- Transform data (X_i, Y_i) to Laplace scale (X_i^*, Y_i^*), $i = 1, \ldots, n$ using GPD approximations
- Let $w_i(x^*)$ be the localizing weights from a regression random forest at x^*
- Define local linear median regression [19] by

$$(\widehat{a}(x^*), \widehat{c}(x^*)) = \underset{a \in [-1,1], c \in \mathbb{R}}{\arg\min} \sum_{i=1}^{n} w_i(x^*) \left| Y_i^* - c - (X_i^* - x^*)a \right|,$$

and note that $\widehat{c}(x^*) \approx \text{median}(Y^* \mid X^* = x^*)$

[19] R. Friedberg, J. Tibshirani, S. Athey, and S. Wager. "Local Linear Forest". In: *Journal of computational and graphical statistics* 30.2 (2021), pp. 503–517.

Random forest progression

- Transform data (X_i, Y_i) to Laplace scale (X_i^*, Y_i^*), $i = 1, \ldots, n$ using GPD approximations
- Let $w_i(x^*)$ be the localizing weights from a regression random forest at x^*
- Define local linear median regression [19] by

$$(\widehat{a}(x^*), \widehat{c}(x^*)) = \underset{a \in [-1,1], c \in \mathbb{R}}{\arg\min} \sum_{i=1}^{n} w_i(x^*) \left| Y_i^* - c - (X_i^* - x^*)a \right|,$$

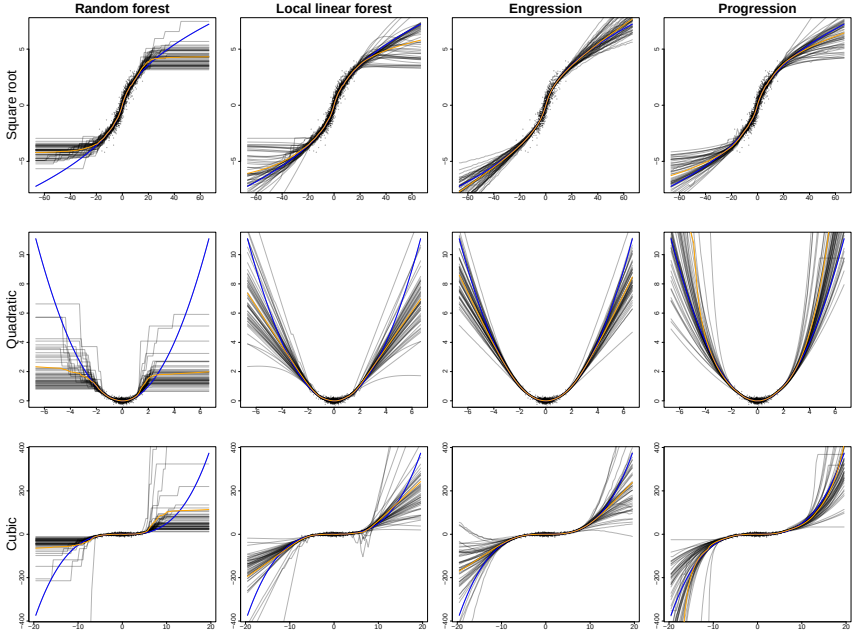
  and note that $\widehat{c}(x^*) \approx \text{median}(Y^* \mid X^* = x^*)$
- For any $x^* > X_{n:n}^*$ above the maximum training predictor we have

$$\widehat{c}(x^*) = \widehat{c}(X_{n:n}^*) + (x^* - X_{n:n}^*)\widehat{a}(X_{n:n}^*)$$

- This is linear extrapolation on Laplace scale!

[19] R. Friedberg, J. Tibshirani, S. Athey, and S. Wager. "Local Linear Forest". In: *Journal of computational and graphical statistics* 30.2 (2021), pp. 503–517.

Experiments



Additive model progression

– For multivariate predictor $\mathbf{X} \in \mathbb{R}^p$, consider the non-parametric additive model

$$Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \varepsilon$$

with regression functions $f_j : \mathbb{R} \to \mathbb{R}$ for $j = 1, \ldots, p$ and zero-median noise ε

Additive model progression

– For multivariate predictor $\mathbf{X} \in \mathbb{R}^p$, consider the non-parametric additive model

$$Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \varepsilon$$

with regression functions $f_j : \mathbb{R} \to \mathbb{R}$ for $j = 1, \ldots, p$ and zero-median noise ε

Backfitting Algorithm for Progression

1: **Input:** Independent samples $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$.
2: Initialize $\widehat{f}_j(x) = 0$ for all $x \in \mathbb{R}$, $j = 1, \ldots, p$, and $\widehat{\alpha} = \frac{1}{n} \sum_{i=1}^{n} Y_i$.
3: Cycle through $j = 1, 2, \ldots, p, 1, 2, \ldots, p, 1, 2, \ldots$
4: Compute residuals:

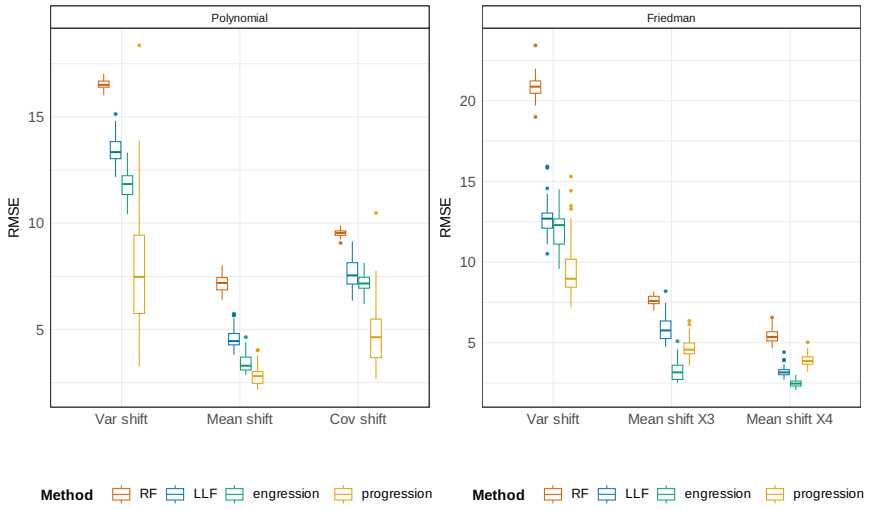
$$R_{ij} = Y_i - \widehat{\alpha} - \sum_{k \neq j} \widehat{f}_j(X_{ik})$$

5: Fit `progression smoother` S to dataset $(R_{1j}, X_{1,j}), \ldots, (R_{nj}, X_{n,j})$.
6: Update:

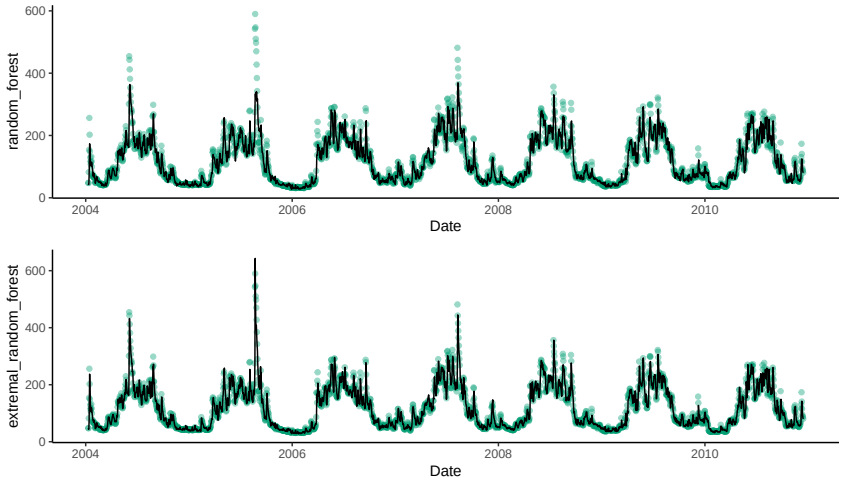
$$\widehat{f}_j(X_{ij}) = S(R_{ij}).$$

7: Until convergence of the functions $\widehat{f}_j$.

Experiments



Extrapolation for peak flow prediction



Thank you!