Probabilistic predictions with Fourier neural operators

Anonymous Author(s)
Affiliation
Address
email

Abstract

Neural networks have been successfully applied in modeling partial differential equations, especially in dynamical systems. Commonly used models, such as neural operators, are performing well at deterministic prediction tasks, but lack a quantification of the uncertainty inherent in many complex systems, for example weather forecasting. In this paper, we explore a new approach that combines Fourier neural operators with generative modeling based on strictly proper scoring rules in order to create well-calibrated probabilistic predictions of dynamical systems. We demonstrate improved predictive uncertainty for our approach, especially in settings with very high inherent uncertainty.

1 Introduction

Many complex phenomena in the sciences are described via time-dependent partial differential equations (PDEs), making their study a crucial research topic. Recent developments in machine learning led to an effective class of neural networks for solving PDEs, called neural operators [Kovachki et al., 2023]. In dynamical systems, these models aim to learn the operator that maps an initial system state to the corresponding solution across time and have been applied to problems such as weather forecasting [Pathak et al., 2022] or fluid dynamics [Renn et al., 2023]. However, neural operators are usually studied in the context of deterministic predictions, not accounting for the inherent uncertainty in complex and chaotic dynamical systems. Several methods have been proposed to enhance neural network architectures to quantify uncertainty. While some approaches focus on perturbing initial conditions [Pathak et al., 2022], many approaches are applied to the network post-hoc. These include statistical post-processing [Bülte et al., 2024] or Bayesian methods [Gal and Ghahramani, 2016]. For neural operators, which learn an output in function space, uncertainty quantification can be more complex. Gal and Ghahramani [2016] propose to generate samples from a posterior predictive distribution by utilizing dropout in the model inference phase. Weber et al. [2024] and Magnani et al. [2024] use a Laplace approximation for the Fourier neural operator, which utilizes a linearized neural network to generate a tractable posterior distribution in function space.

In the context of spatial predictions, especially weather forecasting, methods based on the notion of proper scoring rules are commonly applied and have shown to work very well in combination with neural networks [Pacchiardi et al., 2024, Chen et al., 2024]. However, this has not yet been transferred to the setting of operator learning, which requires additional analysis of the corresponding scoring rules in infinite dimensional spaces. In this paper, we utilize proper scoring rules in separable Hilbert spaces in order to train a neural operator to estimate a predictive distribution over functions. We theoretically prove that this is well-motivated by showing that the energy score is strictly proper in infinite dimensional spaces. Our primary aim is to demonstrate the advantages of the approach in predicting dynamical systems. Our approach, which we refer to as probabilistic Fourier neural operator (PFNO), leads to better-calibrated predictive distributions and adequate uncertainty representations even for long dynamical trajectories.

Submitted to Workshop on Bayesian Decision-making and Uncertainty, 38th Conference on Neural Information Processing Systems (BDU at NeurIPS 2024). Do not distribute.

2 Neural operators

The aim of operator learning is to utilize a neural network to learn a mapping between two function spaces from a finite collection of input-output pairs. Consider an operator $\mathcal{G} : \mathcal{A} \to \mathcal{U}$, acting on two separable Banach spaces of functions. A neural operator is a map $\mathcal{G}_\theta : \mathcal{A} \to \mathcal{U}$, that is parametrized by finitely many weights $\theta \in \mathbb{R}^p$ and trained on observational data $\{(a_n, u_n)\}_{n=1}^N$, which aims to approximate the operator $\mathcal{G}$. Here, a_n is usually an initial system state and u_n is the solution state of the PDE after some time T. The most commonly used architecture is the Fourier neural operator (FNO) [Li et al., 2021], which acts on an input function a by specifying several layers of integral kernels that are parametrized in Fourier space. By utilizing the convolution theorem, one so-called Fourier block is given as

$$G_i v_i(x) = \sigma\left(\mathcal{F}^{-1}(R_i \cdot \mathcal{F}(v_i))(x) + W_i v_i(x)\right), \tag{1}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ are the Fourier transform and its inverse. Here, the matrix-valued functions R_i are directly parametrized in the Fourier domain as neural network weights. The whole network is specified as a combination of several Fourier blocks with some additional lifting and projection functions.

3 Probabilistic predictions using neural operators

The method we propose is based on generating a predictive distribution via a sample-based empirical distribution. We denote the initial condition and the solution of the PDE as a and u, respectively, the spatio-temporal domain as $\mathcal{D}$ and the empirical predictive distribution, for a fixed initial condition a, as $\hat{P}_\theta^M = \{\hat{u}^m\}_{m=1}^M$ with samples $\hat{u}^m$ for $m = 1, ..., M$. For this analysis, we restrict ourselves to data from separable Hilbert spaces, which includes most solution spaces of PDEs, such as the Sobolev space $H^k, k \in \mathbb{N}$.

Scoring rule minimization A scoring rule S is a function that assigns a real-valued score to the fit between a probability distribution and a corresponding observation [Gneiting and Raftery, 2007]. Define the so-called *expected score* as $S(Q, P) := \mathbb{E}_{X\sim P}[S(Q, X)]$. The scoring rule is called *proper* with respect to a class of probability measures $\mathcal{P}$ if $S(P, P) \leq S(Q, P), \forall P, Q \in \mathcal{P}$ and it is called *strictly proper* if equality implies $P = Q$. In other words, a scoring rule is strictly proper if the true distribution of the observation uniquely minimizes the expected score. More details on proper scoring rules can be found in Appendix A. Here, we focus mainly on the well-known *energy score* [Gneiting and Raftery, 2007], which is defined as

$$\text{ES}(P, x) := \mathbb{E}_P[\|X - x\|_{\mathcal{H}}] - \frac{1}{2}\mathbb{E}_P[\|X - X'\|_{\mathcal{H}}], \tag{2}$$

where $X, X' \overset{\text{i.i.d}}{\sim} P, x \in \mathcal{H}$ and $\mathcal{H}$ is a separable Hilbert space. Pacchiardi et al. [2024] show how generative neural networks can be trained via scoring rule minimization in the finite-dimensional setting. Consider data observation pairs of the form $(a_i, u_i)_{i=1}^n$, where $a_i \sim P_\mathcal{A}$ and $u_i \sim P_\mathcal{U}$ follow some distributions over a separable Hilbert space and let $P_\theta(\cdot \mid a)$ denote an approximate posterior generated by the network. In the conditional data setting, we assume that $u_i \sim P^*(\cdot \mid a_i)$. For a (strictly) proper scoring rule, the minimization objective is given as

$$\underset{\theta}{\operatorname{argmin}}\, \mathbb{E}_{a\sim P_\mathcal{A}} \mathbb{E}_{u\sim P^*(\cdot|a)} S(P_\theta(\cdot \mid a), u)$$

and leads to $P_\theta(\cdot \mid a) = P^*(\cdot \mid a)$ almost everywhere. In the finite data setting, this objective is approximated with a Monte Carlo estimator. While closed-form expressions of S are not always available, the energy score has a representation that admits an unbiased estimator, which requires the output from our neural network to consist of multiple samples of the predictive distribution, e.g. $(\hat{u}^m)_{m=1}^M \sim P_\theta(\cdot \mid a)$. In our case, the minimization objective for the neural network with the energy score then becomes

$$\underset{\theta}{\operatorname{argmin}}\, \frac{1}{n}\sum_{i=1}^n \left(\frac{1}{M}\sum_{m=1}^M \|\hat{u}_i^m - u_i\|_{\mathcal{H}} - \frac{1}{2M(M-1)} \sum_{\substack{m,h=1\\ m\neq h}}^M \|\hat{u}_i^m - \hat{u}_i^h\|_{\mathcal{H}} \right). \tag{3}$$

In this paper, we extend this approach to the infinite-dimensional setting of neural operators. This is mathematically motivated, as we prove in Appendix A that the energy score is strictly proper in separable Hilbert spaces, which allows network training via scoring rule minimization. As the base architecture we utilize FNOs, and refer to our approach as probabilistic Fourier neural operator (PFNO). In order to create an empirical distribution as the network output, we focus on utilizing stochastic forward passes via dropout [Gal and Ghahramani, 2016]. To further account for the structure of the FNO, we apply additional dropout over the parameters in Fourier space, which then acts globally on the network prediction. The PFNO has the advantage that it is easy to implement in an existing architecture and creates a nonparametric predictive distribution, allowing for more flexibility. We compare the PFNO against the MCDropout and Laplace approximation as baselines.

Baselines: Gal and Ghahramani [2016] show that a neural network with dropout before each layer is mathematically equivalent to a variational approximation of a Gaussian process. This leads to a simple and efficient way of creating a predictive distribution, referred to as MCDropout. In our setting, the predictive distribution is given as

$$\hat{P}_\theta^M = \{\mathcal{G}_\theta^*(a, \boldsymbol{\omega}_m)\}_{m=1}^M, \tag{4}$$

where $\boldsymbol{\omega}_m$ is the random dropout variable and $\mathcal{G}_\theta^*$ denotes a neural operator trained.

Weber et al. [2024] propose to utilize the Laplace approximation (LA) for FNOs, which is based on building a second-order approximation of the weights around the maximum a posteriori (MAP) estimate. By assuming a Gaussian weight prior, the weight-space uncertainty of the LA is given by

$$p(\theta, \mathcal{C}) \approx \mathcal{N}(\theta; \theta_{\mathrm{MAP}}, \Sigma), \quad \Sigma := -(\nabla_\theta^2 \mathcal{L}(\mathcal{C}; \theta)|_{\theta_{\mathrm{MAP}}})^{-1}, \tag{5}$$

where $\mathcal{C} = \{(a_n, u_n)\}_{n=1}^N$. The corresponding predictive distribution in function space is an analytically available Gaussian and is used to generate M predictive samples. In contrast to the PFNO, both methods quantify uncertainty for an already trained neural operator.

4 Experimental results

We analyze our methods on two highly uncertain dynamical systems. First, the Kuramoto-Sivashinsky (KS) equation, which is a one-dimensional chaotic fourth-order parabolic PDE, described by

$$\partial_t u(x,t) + u\partial_x u(x,t) + \partial_x^2 u(x,t) + \partial_x^4 u(x,t) = 0, \quad u(x,0) = u_0(x). \tag{6}$$

We generate 10.000 samples (10% for validation/evaluation) over the domain $\mathcal{D} = [0, 100] \times [0, 300]$. In addition, we evaluate the models on a 2-meter surface temperature prediction task, where we utilize the ERA5 dataset, provided via the WeatherBench benchmark [Rasp et al., 2024] with a spatial resolution of $0.25°$ across Europe and a time resolution of $6h$. We use data from 2011-2022 with the last two years as validation and test data respectively. For the KS data the model takes and predicts 20 timesteps, for the ERA5 it takes and predicts 10 timesteps (60 hours). For evaluation, we consider the following metrics:

$$\mathrm{RMSE}(\hat{P}_\theta^M, u) = \|\overline{u}_M - u\|_{L^2}, \tag{7}$$

$$\mathrm{ES}(\hat{P}_\theta^M, u) = \frac{1}{M}\sum_{m=1}^M \|\hat{u}^m - u\|_{L^2} - \frac{1}{2M(M-1)}\sum_{m \neq h}^M \|\hat{u}^m - \hat{u}^h\|_{L^2}, \tag{8}$$

$$\mathcal{C}_\alpha(\hat{P}_\theta^M, u) = \int_{\mathcal{D}} \mathbb{1}\left\{u(x,t) \in [\hat{q}_\theta^{\alpha/2}(x,t), \hat{q}_\theta^{1-\alpha/2}(x,t)]\right\} dx\, dt, \tag{9}$$

Table 1: Evaluation metrics on the Kuramoto-Sivashinsky equation and the 2-meter surface temperature. The best model is highlighted in bold.

		Validation data			Test data		
		RMSE	ES	$\mathcal{C}_{0.05}$	RMSE	ES	$\mathcal{C}_{0.05}$
KS	PFNO	0.8674	**0.6108**	**0.8781**	0.8677	**0.6110**	**0.8774**
	MCDropout	0.8446	0.7298	0.3595	0.8457	0.7310	0.3580
	Laplace	**0.8352**	0.8247	0.0197	**0.8359**	0.8250	0.0207
T2M	PFNO	**0.1677**	**0.1182**	**0.9427**	0.3291	**0.2427**	**0.7865**
	MCDropout	0.1834	0.1427	0.6325	**0.3035**	0.2535	0.4284
	Laplace	0.1910	0.1421	0.3231	0.3145	0.2491	0.2387

where $\overline{u}_M$ denotes the mean prediction, and $\hat{q}_\theta^\alpha$ denotes the empirical α quantile of the predictive distribution. All metrics are then averaged over the validation or test dataset. The RMSE evaluates the match between the mean of the predictive distribution and the observation, while the energy score evaluates the match for the predictive distribution as a whole. The coverage $\mathcal{C}_\alpha$ is calculated pointwise and describes, whether the predictive $1-\alpha$ interval entails the true value. It should be close to $1-\alpha$ for a well-calibrated prediction.

For a fair comparison, all methods use the same architecture, namely an FNO with 20 hidden channels, 10 modes in the time dimension, and 12 modes in the spatial dimension (2d or 3d respectively), and generate a predictive distribution of size $M = 100$. Furthermore, we tune the dropout for all methods separately via grid search, as they highly depend on this parameter. The results for both experiments can be found in Table 1, while Figure 1 shows an additional analysis of the temporal behavior of the estimations for the Kuramoto-Sivashinsky equation.

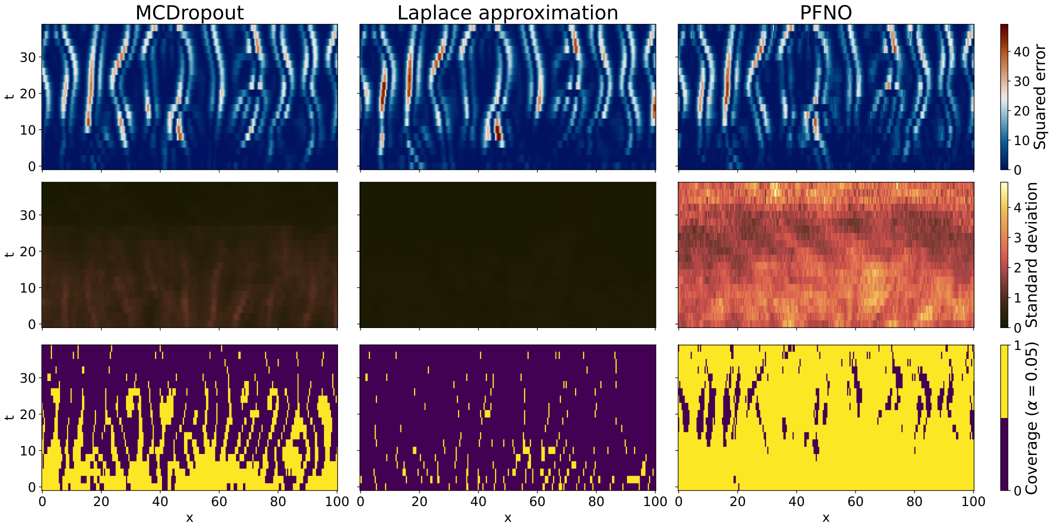

Figure 1: The figure shows the squared error, standard deviation, and 95%-coverage for the different methods on a random test sample of the Kuramoto-Sivashinsky equation. The mean coverage values from left to right are 30.90%, 4.82% and 89.32%.

5 Conclusion

The PFNO obtains the best fit between the observation and the predictive distribution in terms of the energy score. Furthermore, for the temperature prediction task, the RMSE is comparable to or even lower than the other methods, although it is not explicitly minimized by the network. Finally, the PFNO provides the best-calibrated prediction intervals, although the coverage is generally below the optimal value. For the temperature prediction task, the performance is significantly worse on the test set for all models and future research might revolve around making the approaches more robust against out-of-distribution data. While the Laplace approximation is very easy to use and admits an approximate analytically available Gaussian distribution, this might not be flexible enough for complex dynamical systems, if, for example, the uncertainty does not follow a symmetric distribution. Although it leads to a better mean estimation, as it is based on the MAP estimate, the predictive uncertainty is not as adequate. The MCDropout method lacks calibration in terms of coverage but also generally provides a good mean prediction.

While our approach shows improved performance, is easy to implement, and can be used with basically any architecture, it requires an additional training step and more computational power, as multiple samples are necessary to calculate the loss function. Still, these findings are very promising and encourage further analysis of neural operators trained with scoring rule minimization for complex dynamical systems. Some aspects to investigate are different suitable scoring rules, different ways of generating the samples, as well as ways to improve coverage and calibration of the prediction.

References

C. Bülte, N. Horat, J. Quinting, and S. Lerch. Uncertainty quantification for data-driven weather models. Number arXiv:2403.13458. arXiv, Mar. 2024. doi:10.48550/arXiv.2403.13458.

J. Chen, T. Janke, F. Steinke, and S. Lerch. Generative machine learning methods for multivariate ensemble postprocessing. *The Annals of Applied Statistics*, 18(1):159–183, Mar. 2024. ISSN 1932-6157, 1941-7330. doi:10.1214/23-AOAS1784.

Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 2016. PMLR.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. ISSN 0162-1459. doi:10.1198/016214506000001437.

N. Kovachki, Z. Li, B. Liu, K. Azizzadenesheli, K. Bhattacharya, A. Stuart, and A. Anandkumar. Neural operator: Learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar. Fourier Neural Operator for Parametric Partial Differential Equations. Number arXiv:2010.08895. arXiv, May 2021.

R. Lyons. Distance covariance in metric spaces. *The Annals of Probability*, 41(5):3284–3305, Sept. 2013. ISSN 0091-1798, 2168-894X. doi:10.1214/12-AOP803.

E. Magnani, M. Pförtner, T. Weber, and P. Hennig. Linearization Turns Neural Operators into Function-Valued Gaussian Processes. Number arXiv:2406.05072. arXiv, June 2024.

L. Pacchiardi, R. A. Adewoyin, P. Dueben, and R. Dutta. Probabilistic forecasting with generative networks via scoring rule minimization. *Journal of Machine Learning Research*, 25(45):1–64, 2024.

J. Pathak, S. Subramanian, P. Harrington, S. Raja, A. Chattopadhyay, M. Mardani, T. Kurth, D. Hall, Z. Li, K. Azizzadenesheli, P. Hassanzadeh, K. Kashinath, and A. Anandkumar. FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators. Number arXiv:2202.11214. arXiv, Feb. 2022. doi:10.48550/arXiv.2202.11214.

S. Rasp, S. Hoyer, A. Merose, I. Langmore, P. Battaglia, T. Russel, A. Sanchez-Gonzalez, V. Yang, R. Carver, S. Agrawal, M. Chantry, Z. B. Bouallegue, P. Dueben, C. Bromberg, J. Sisk, L. Barrington, A. Bell, and F. Sha. WeatherBench 2: A benchmark for the next generation of data-driven global weather models. Number arXiv:2308.15560. arXiv, Jan. 2024. doi:10.48550/arXiv.2308.15560.

P. I. Renn, C. Wang, S. Lale, Z. Li, A. Anandkumar, and M. Gharib. Forecasting subcritical cylinder wakes with Fourier Neural Operators. Number arXiv:2301.08290. arXiv, Jan. 2023.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, Oct. 2013. ISSN 0090-5364, 2168-8966. doi:10.1214/13-AOS1140.

I. Steinwart and J. F. Ziegel. Strictly proper kernel scores and characteristic kernels on compact spaces. *Applied and Computational Harmonic Analysis*, 51:510–542, Mar. 2021. ISSN 1063-5203. doi:10.1016/j.acha.2019.11.005.

T. Weber, E. Magnani, M. Pförtner, and P. Hennig. Uncertainty quantification for fourier neural operators. In *ICLR 2024 Workshop on AI4DifferentialEquations in Science*, 2024.

J. Ziegel, D. Ginsbourger, and L. Dümbgen. Characteristic kernels on Hilbert spaces, Banach spaces, and on sets of measures. *Bernoulli*, 30(2):1441–1457, May 2024. ISSN 1350-7265. doi:10.3150/23-BEJ1639.

A Proper scoring rules in separable Hilbert spaces

This section aims to provide more detailed insights into proper scoring rules and generalizations over separable Hilbert spaces. The results and notations draw mainly on Ziegel et al. [2024], Steinwart and Ziegel [2021]. Let $(\mathcal{X}, \mathcal{A})$ be a measurable space and let $\mathcal{M}_1(\mathcal{X})$ denote the class of all probability measures on $\mathcal{X}$. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric and positive function referred to as *kernel* and define $\mathcal{M}_1^k(\mathcal{X}) := \left\{ P \in \mathcal{M}_1(X) \mid \int_X \sqrt{k(x,x)} dP(x) < \infty \right\}$. A measurable and bounded kernel is called *characteristic* if the kernel embedding defined by $\Phi(P) := \int k(\cdot, \omega) dP(\omega)$, with $P \in \mathcal{M}_1^k(\mathcal{H})$ is injective [Steinwart and Ziegel, 2021].

For $\mathcal{P} \subseteq \mathcal{M}_1(\mathcal{X})$, a *scoring rule* is a function $S : \mathcal{P} \times \mathcal{X} \to [-\infty, \infty]$ such that the integral $\int S(P,x) dQ(x)$ exists for all $P, Q \in \mathcal{P}$. Define the so-called *expected score* as $S(Q,P) := \int S(Q,x)\, dP(x) = \mathbb{E}_{X \sim P}[S(Q,X)]$. Then S is called *proper* with respect to $\mathcal{P}$ if

$$S(P,P) \leq S(Q,P), \quad \text{for all } P, Q \in \mathcal{P}, \tag{10}$$

and it is called *strictly proper* if equality in (10) implies $P = Q$. For a more detailed overview compare Gneiting and Raftery [2007]. Proper scoring rules are closely connected to characteristic kernels, in fact Steinwart and Ziegel [2021] showed that the *kernel score* is strictly proper if and only if the underlying kernel is characteristic. The kernel score S_k associated with a measurable kernel k on $\mathcal{X}$ is the scoring rule $S_k : \mathcal{M}_1^k \times \mathcal{X} \to \mathbb{R}$ defined by

$$S_k(P,x) = \frac{1}{2}\mathbb{E}_P[k(X,X')] - \mathbb{E}_P[k(X,x)] + \frac{1}{2}k(x,x),$$

where $x \in \mathcal{X}$ and $X, X' \overset{i.i.d}{\sim} P \in \mathcal{M}_1^k$.

We will use the notion of the kernel score to derive a functional version of the commonly used energy score [Gneiting and Raftery, 2007] and show that it is a strictly proper scoring rule in separable Hilbert spaces.

Theorem A.1 (Energy score)**.** *Let $\mathcal{H}$ denote a separable Hilbert space. The energy score* ES : $\mathcal{M}_1^k(\mathcal{H}) \times \mathcal{H} \to \mathbb{R}$ *defined as*

$$\mathrm{ES}(P,x) := \mathbb{E}_P[\|X - x\|_{\mathcal{H}}] - \frac{1}{2}\mathbb{E}_P[\|X - X'\|_{\mathcal{H}}],$$

with $x \in \mathcal{H}$ and $X, X' \overset{i.i.d}{\sim} P \in \mathcal{M}_1^k$ is strictly proper relative to the class $\mathcal{M}_1^k(\mathcal{H})$.

Proof. Lyons [2013, Theorem 3.25] states that every separable Hilbert space is of so-called strong negative type, which implies the existence of a positive definite *distance kernel* induced by the metric $\|\cdot\|_{\mathcal{H}}$ given as $k(z,z') = \|z - z_0\|_{\mathcal{H}} + \|z' - z_0\|_{\mathcal{H}} - \|z - z'\|_{\mathcal{H}}$ for some fixed $z_0 \in \mathcal{H}$. Furthermore, Sejdinovic et al. [2013, Proposition 29] state that the corresponding kernel k is characteristic. Defining $k_d(z,z') := d(z,z_0) + d(z',z_0) - d(z,z')$, with $d(x,x') := \|x - x'\|_{\mathcal{H}}$ and $z_0 \in \mathcal{H}$ leads to

$$\begin{aligned}
S_{k_d}(P,x) &= -\mathbb{E}_P[k_d(X,x)] + \frac{1}{2}\mathbb{E}_P[k_d(X,X')] + \frac{1}{2}k_d(x,x) \\
&= -\mathbb{E}_P[d(X,z_0) + d(x,z_0) - d(X,x)] + \frac{1}{2}\mathbb{E}_P[d(X,z_0) + d(X',z_0) - d(X,X')] \\
&\quad + \frac{1}{2}(d(x,z_0) + d(x,z_0) - d(x,x)) \\
&= \mathbb{E}_P[d(X,x)] - \frac{1}{2}\mathbb{E}_P[d(X,X')] - \frac{1}{2}\mathbb{E}_P[d(X,z_0)] + \frac{1}{2}\mathbb{E}_P[d(X',z_0)] - \mathbb{E}_P[d(x,z_0)] + d(x,z_0) \\
&= \mathbb{E}_P[d(X,x)] - \frac{1}{2}\mathbb{E}_P[d(X,X')] = \mathrm{ES}(P,x)
\end{aligned}$$

Since k_d is characteristic, by Steinwart and Ziegel [2021] the energy score is strictly proper relative to the class $\mathcal{M}_1^{k_d}(\mathcal{H})$. □