# Theoretical overview : Disentangled uncertainty quantification for regression

Kevin PASINI et ALL

December 18, 2023

## Contents



This document contains the theoretical sources related to the work on uncertainty quantification decomposed on ML/deep learning model for regression and anomaly detection.

## A Uncertainty in Regression task

Let us consider a modeling framework, in which random variables (denoted $\varepsilon$) are linked to specific uncertainty sources. Here, time series forecasting (link to the type of data on which the work was undertaken) is treated as a regression problem based on time-dependent features. In this context, a model $\widehat{f}$ aims to predict the nominal behavior for variables of interest represented by univariate/multivariate time series $Y = (y_1, \ldots, y_t, \ldots, y_T)$. The forecast at time step $t$ for the variable $y_t$ will be based on a vector of observed variables $\mathbf{x}_t$ composed of variables exogenous $\mathbf{c}_t$ and lagged response $\boldsymbol{Y}_t^{past} = (y_{t-lag}, \ldots, y_{t-1})$ variables, as well as some latent variables $\mathrm{h}_t$:

$$\mathrm{y_t} = \boldsymbol{f}(\mathbf{x}_t) + \varepsilon_t^{\boldsymbol{u}} \quad \text{with;} \quad \varepsilon_t^{\boldsymbol{u}} \sim \mathcal{N}(0, \sigma_t^u(\mathbf{x}_t, \mathrm{h}_t)) \quad ; \quad \mathbf{x}_t = \{\mathbf{c}_t, \boldsymbol{Y}_t^{past}\}$$

with $\boldsymbol{f}(\mathbf{x}_t)$ the average explainable signal, and $\varepsilon_t^u$ a time-dependent Gaussian noise (local homogeneity assumption). The latter is associated with upstream irreducible variability, encompassing both intrinsic, measurement noises and pre-modeling noise arising from limits of the modeling scope (e.g. due to the influence of hidden variables $\mathrm{h}_t$ that cannot be captured through lagged temporal variables $\boldsymbol{Y}_t^{past}$).

The ML model $\hat{f}_\theta$ aims to approximate the explainable part $f$ of the target $y$ using observed variables $x$ from a training set $D_\theta = (\mathbf{x}_1, y_1), .., (\mathbf{x}_n, y_n)$, a subset of the dataset $\mathscr{D}$. $\theta$ is the set of parameters obtained

using the training set $D_\theta$, over $\Theta$ indicating the whole set of parameters linked to all subsets of the dataset $\mathscr{D}$. According to the bias-variance trade-off (Eq.1), we decompose all error sources between $y_t$ and $\hat{f}_\theta(\mathbf{x}_t)$:

$$\begin{aligned}\boldsymbol{E}_{\boldsymbol{\Theta}}\left[(y_t-\hat{f}_\theta(\mathbf{x}_t))^2\right] &= \boldsymbol{E}_{\boldsymbol{\Theta}}\left[\hat{f}_\theta(\mathbf{x}_t)-f(\mathbf{x}_t)\right]^2+\boldsymbol{E}_{\boldsymbol{\Theta}}\left[\left(\boldsymbol{E}_{\boldsymbol{\Theta}}\left[\hat{f}_\theta(\mathbf{x}_t)\right]-\hat{f}_\theta(\mathbf{x}_t)\right)^2\right]+E_y\left[(y_t-f(\mathbf{x}_t))^2\right]\\ &= \underbrace{\left(f^*_\Theta(\mathbf{x}_t)-f(\mathbf{x}_t)\right)^2}_{\text{Bias}}+\underbrace{\boldsymbol{E}_{\boldsymbol{\Theta}}\left[\left(f^*_\Theta(\mathbf{x}_t)-\hat{f}_\theta(\mathbf{x}_t)\right)^2\right]}_{\text{Variance}}+\underbrace{\sigma^u_t}_{\text{Intrinsic variability}}\quad,\end{aligned} \tag{1}$$

with $f^*_\Theta=\boldsymbol{E}_{\boldsymbol{\Theta}}[\hat{f}_\theta(\mathbf{x}_t)]$, the average function given the distribution $\Theta$. Among the three above-mentioned error sources, the *variance* can be explained by a noise $\varepsilon^\theta_t$ which corresponds to the gap between the average function over $\Theta$ and the ML model: $\varepsilon^\theta_t=f^*_\Theta(\mathbf{x}_t)-\hat{f}_\theta(\mathbf{x}_t)$. This epistemic noise is related to insufficient observations and could be reduced by gathering more data. The *bias* requires another random variable $\varepsilon^\Theta_t$ linked to the gap between the average explainable signal and the average function over $\Theta$: $\varepsilon^\Theta_t=f(\mathbf{x}_t)-f^*_\Theta(\mathbf{x}_t)$. This noise, due to the modeling constraint over $\Theta$, is irreducible in the modeling scope. Finally, the *intrinsic variability* is related to the irreducible noise $\varepsilon^u_t$ that appears upstream of the modeling scope. It quantifies a lower bound for the expected error in the test data with both infinite data and unconstrained modeling. To show the relation between the introduced random variables and the epistemic/aleatoric concepts, we inject them into the total uncertainty law [8] in Eq.11. After simplification allowed by strong independence and zeros-mean assumptions:

$$\begin{aligned}&\text{With } y_t=f_\theta(\mathbf{x}_t)+\varepsilon^\theta_t+\varepsilon^\Theta_t+\varepsilon^u_t \quad\text{and}\quad \sigma(y_t|x_t;\theta)=\sigma_\Theta\left[E_y(y_t|x_t;\theta)\right]+\boldsymbol{E}_\Theta\left[\sigma_y(y_t|x_t;\theta)\right]=\boldsymbol{\sigma}^{\boldsymbol{E}}_t+\boldsymbol{\sigma}^{\boldsymbol{A}}_t\\ &\text{We obtain}\quad \boldsymbol{\sigma}^{\boldsymbol{E}}_t=\sigma_\Theta\left[\widehat{f_\theta}(\mathbf{x}_t)\right]=\boldsymbol{E}_{\boldsymbol{\Theta}}\left[\left(\varepsilon^\theta_t\right)^2\right]\quad,\quad \boldsymbol{\sigma}^{\boldsymbol{A}}_t=\sigma_y(\varepsilon^u_t)+\sigma_y(\varepsilon^\Theta_t)\end{aligned} \tag{2}$$

From these equations, we can see that the decomposition into epistemic and aleatoric components (denoted by $E$ and $A$ superscripts) requires the manipulation of the whole set of parameters $\Theta$. As expected, the epistemic part is essentially made up of the variance error caused by the sampling of the training set. However, the aleatoric part contains several quantities that are all irreducible in the modeling scope but may be associated with different sources: upstream modeling scope (intrinsic, measurement, and pre-modeling noise), and model constraints which also cause bias. When we move slightly outside the domain of validity of the assumptions (due to limited training data and approximate manipulation of $\Theta$), the previous negligible terms can then induce blurs into the uncertainty decomposition.

**View of an unified dUQ framework:** The functional scheme of the proposed dUQ framework incorporating various UQ paradigms is shown in Fig. 2. It is based on a metamodel $\mathscr{M}^\Theta$ that learns and manipulates diverse submodels ($\hat{f}_\theta$) to combine their inferences. The learning phase aims to capture the explainable variability and estimate irreducible variability while exploring a diversity of submodel candidates $\Theta$. To ensure diversity and avoid submodel redundancy, a variability infusion mechanism (depending on the UQ paradigm) is needed during the learning phase. The estimated submodels produce, at the inference step, a local regressor $\hat{f}_\theta(\mathbf{x}_t)$ and an estimation of aleatoric variability $\widehat{\sigma}^a(f_\theta(\mathbf{x}_t))$. Furthermore, an epistemic variability $\widehat{\sigma}^e_t$ is produced by computing the variability of the submodel regression $\widehat{y}_t$ (using for example a Gaussian assumption). Finally, the metamodel provides a *risk-aware forecast* comprising three indicators: $\overline{\mu}_t, \overline{\sigma^a}_t, \widehat{\sigma^e}_t$ expressing forecast, aleatoric and epistemic indicators. As can be seen in Fig. 1, these indicators correspond to three independent axes on how the model perceives the data regarding forecast and sources of uncertainty. We can use them to design confidence intervals, error margins, or warnings highlighting a lack of model confidence.

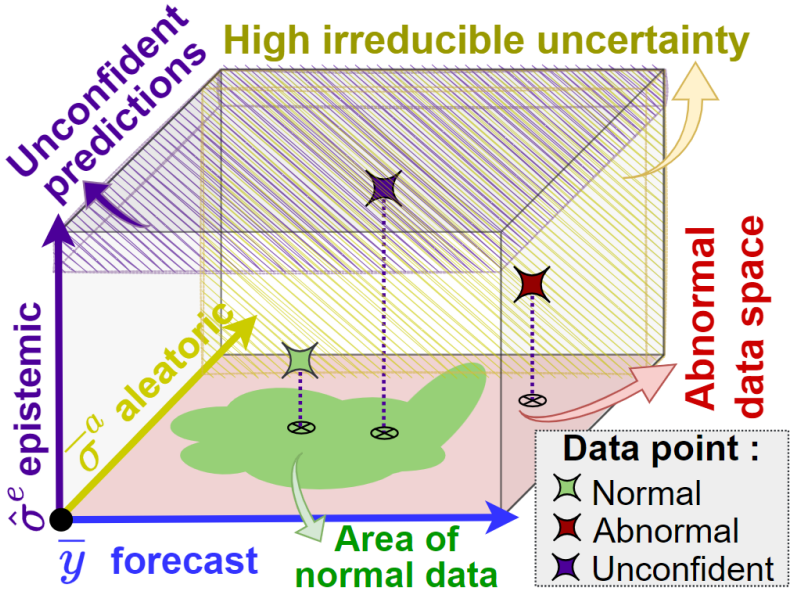


Figure 1: Theoretical *disentangled Uncertainty Quantification* (dUQ) indicators space

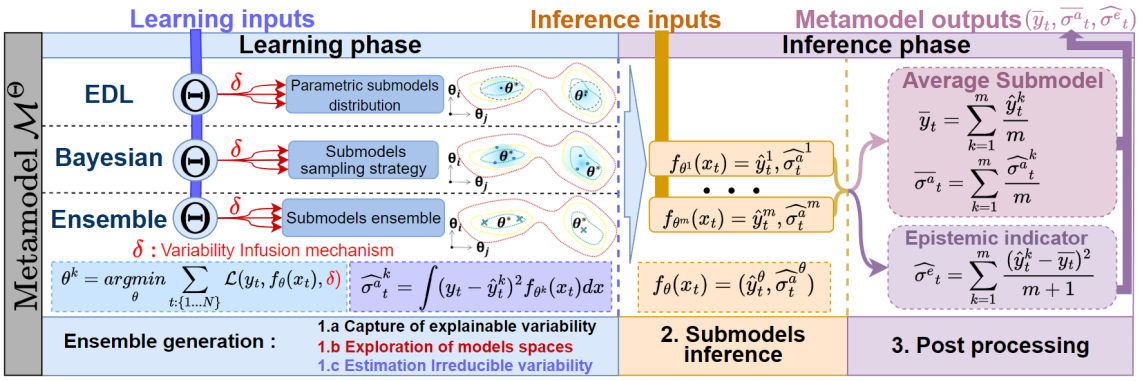


Figure 2: Illustration of a metamodel using Gaussian Aleatoric and Epistemic assumptions.

## A.1 Detailed explanation of a disentangled uncertainty quantification formalism :

**Submodels** are functions $f = \{f^{\theta^1}, \ldots, f^{\theta^m}\}$ parameterized by $\theta$ that allow the *local* estimation of a decision function and its irreducible (Aleatoric) variability $\varepsilon_t^{\boldsymbol{A}}$ around the time stamp $t$. We have seen that this quantity explains the inherent noise in the data, which could not be explained during the training of the models but could be measured at each time stamp $t$ for submodel $k$:

$$\varepsilon_t^{\boldsymbol{A}} \sim \mathcal{N}(0, \sigma_t^a) : \quad \sigma_t^a = \int (y_t - \mu_t)^2 f^\theta(\mathbf{x}_t) dy = \int (y_t - \hat{\mu}_t)^2 f^{\theta^k}(\mathbf{x}_t) dy = \widehat{\sigma^a}_t. \tag{3}$$

Such mechanisms take various forms in the literature (Bayesian model, Frequentist or Set-based approach) under different hypotheses (distribution assumption, moment estimation, quantiles, etc.). With a local Gaussian Aleatoric assumption justified by the application of central limit theorem on the observed variables, our submodels estimate the conditional probability distribution of the target variable from the observed variables.

$$\theta^k = \arg\min_{\theta} \sum_{t=1}^{T} Loss(y_t, \widehat{\mu}_t^{\theta}, \widehat{\sigma^a}_t^{\theta}) \quad , \quad P_{\theta^k}(y_t|x_t) \sim \mathcal{N}(\widehat{\mu}_t^k, \widehat{\sigma^a}_t^k)$$

**The metamodel $\mathcal{M}^{\boldsymbol{\theta}}$** is composed of a submodel ensemble (or submodel family). To specify a well-constructed metamodel, we have to introduce an abstract set $\Omega$: the family of reachable functions (that can be approximated by the type of selected model) and relevant submodels regarding a standard ML framework $S$. Then, the metamodel generates, manipulates, and enhances this ensemble of submodels through four following steps:

**1. The generation of the submodels ensemble** is realized by introducing variability during parameter optimization $\theta$ through a variability infusion mechanism $\delta$, which takes different forms depending on the UQ-paradigm. The mechanisms of our four evaluated approaches can be interpreted as bagging/bootstrap (Ensemble), local minimum convergence (Deep Ensemble), binomial probabilistic weight (MCDP), or parameterization evidence-based (EDL). Mathematically, this is often achieved by disturbing the learning process (eq. 4) to generate submodel variants. To avoid having redundant submodels (lack of diversity) or coarse estimations (lack of accuracy), the metamodel has to be well-defined meaning:

$$\{\mathcal{M}^{\theta} = \mathcal{M}_{\{\theta^1,\dots,\theta^m\}} \text{ with } \theta^k \in \Omega \mid \theta^k = \arg\min_{\theta} Loss(Y, f^{\theta}(\mathbf{x}), \delta(Y, X, f^{\theta}))\} \quad \text{With} \tag{4}$$

- Considerable diversity of the submodels: $\forall (k,k')$ with $k \neq k'$, $\quad d(\theta^k, \theta^{k'}) > \varepsilon$
- Acceptable accuracy and generalization capacity of submodels: $\forall k \in [1,m],\ f^{\theta^k} \in \Omega$

**2. Average model extrapolation** could be virtually computed through the average decision and the average aleatoric indicator of the submodels ensemble in a Gaussian aleatoric assumption.

$$\overline{\mu}_t = \frac{1}{m}\sum_{k=1}^{m} \widehat{\mu}_t^k, \qquad \overline{\sigma^a}_t = \sqrt{\frac{1}{m}\sum_{k=1}^{m} (\widehat{\sigma^a}_t^k)^2} \tag{5}$$

**3. Epistemic estimation $\varepsilon_t^{\boldsymbol{E}}$** allows for local estimation of uncertainty through the metamodel submodels. Epistemic confidence can then be interpreted [1] as the likelihood of the *average submodel decision* ($\overline{\mu}_t$) over the metamodel. Under the well-formed ensemble assumption for a regression task, local decision estimators $f^{\theta^k}$ are centred with respect to the local mean $\mu_t$ and have independent errors. Then, a Gaussian epistemic assumption[2] allows the use of the following unbiased empirical variance estimators.

$$\varepsilon_t^{\boldsymbol{E}} \sim \mathcal{N}(0, \sigma_t^e): \qquad \sigma_t^e \approx \widehat{\sigma^e}_t = \sqrt{\frac{1}{m+1}\sum_{k=1}^{m} (\widehat{\mu}_t^k - \overline{\mu}_t)^2} \tag{6}$$

**4. Metamodel output** can express a decision and estimate its forecast uncertainty, $\mathcal{M}_{\{\theta^k\}_{k=1}^m}(\mathbf{x}_t) \sim \mathcal{N}(\overline{\mu}_t, \widehat{\sigma}_t)$ corresponding to the combination of aleatoric and epistemic uncertainties $\widehat{\sigma}_t^2 = \overline{\sigma^a}_t^2 + \widehat{\sigma^e}_t^2$.

[1] More information in Section 3 of supplementary materials

[2] Thanks to unbiased (target-centred) models with independent errors

It should be noticed that there is a high correlation between the epistemic and aleatoric uncertainty indicators: the model error increases with high irreducible uncertainty. To better exploit local epistemic variability, we propose the dE-Indicator[1] corresponding to a Negative Ratio of Epistemic under Total Log-Likelihoods through Aleatoric and Epistemic Gaussian assumptions: $I_t^e = -ln(1 + \frac{\widehat{\sigma^a}^2_t}{\widehat{\sigma^e}^2_t})$

**Additional remarks** Two additional remarks should be made. First, the dUQ framework may be extended with other Aleatoric and Epistemic assumptions (distribution law, quantile, other kinds of moments). These extensions will lead to three main issues: submodel aleatoric estimation, submodel aggregation, and epistemic extraction from the ensemble. Secondly, this framework "contains" the standard *Machine Learning* (ML) approaches if they are formalized under the constant aleatoric assumption and total confidence in the unique submodel. It also "contains" other classical ML *Uncertainty Quantification* (UQ) focusing on Aleatoric or Epistemic if formalized under the assumption that neglecting the other part.

# B Anomaly score :

One standard way of characterizing abnormal behavior in a time series is to analyze the deviation, the difference between the observation and the behavior expected according to a model making a nominal prediction. It allows producing anomaly score based on prediction residuals. Then, it is possible to include consideration of an Uncertainty measure the score by creating a Z-score by normalizing the deviation based on the uncertainty to obtain a contextual anomaly score. In what follows, we propose a contextual anomaly score desgin including uncertainty, aléatorics and epistemics :

The design of the score anomaly ($\boldsymbol{S}_t$) is then based on three terms :

- The absolute deviation ($\boldsymbol{r}_t$) measures the normality of a sample.
- $\beta$-level confidence interval threshold $\widehat{\boldsymbol{\sigma}}_t^{\beta}$ discriminate local statistical anomalies using local variability estimated by the model.
- $\alpha$-level model risk penalization ($\boldsymbol{pen}_t^{E,\alpha}$) allows handling of predictions with low epistemic reliability.

$$\boldsymbol{S}_t = sign(\boldsymbol{r}_t) * \frac{|\boldsymbol{r}_t| + \boldsymbol{pen}_t^{E,\alpha}}{\widehat{\boldsymbol{\sigma}}_t^{\beta}} : \begin{cases} \boldsymbol{r}_t &= (\boldsymbol{y}_t - \widehat{\boldsymbol{y}}_t) \\ \widehat{\boldsymbol{\sigma}}_t^{\beta} &= \beta * \sqrt{\widehat{\boldsymbol{\sigma}}_t^{A^2} + \widehat{\boldsymbol{\sigma}}_t^{E^2}} \\ \boldsymbol{pen}_t^{E,\alpha} &= \alpha * \widehat{\boldsymbol{\sigma}}_t^{E} \end{cases} \tag{7}$$

$$\boldsymbol{S}_t = \frac{(\boldsymbol{y}_t - \boldsymbol{M}(x_t)) \pm \boldsymbol{pen}_t^{E,\alpha}}{\widehat{\boldsymbol{\sigma}}_t^{\beta}} = \underbrace{\frac{\boldsymbol{e}_t^a}{\widehat{\boldsymbol{\sigma}}_t^{\beta}}}_{Ctx\ deviation} \pm \underbrace{\frac{\varepsilon_t^A}{\widehat{\boldsymbol{\sigma}}_t^{\beta}}}_{Data\ Noise} \pm \underbrace{\frac{B_t + \varepsilon_t^E}{\widehat{\boldsymbol{\sigma}}_t^{\beta}}}_{Model\ Noise} \pm \underbrace{\frac{\boldsymbol{pen}_t^{E,\alpha}}{\widehat{\boldsymbol{\sigma}}_t^{\beta}}}_{Model\ risk} \tag{8}$$

The following 3 assumptions allow us to make optimal use of the score:

$$\begin{aligned} &(i) \quad \text{Rare and significant anomaly}: \quad \forall\ t_i\ with\ t_i\ is\ anom,\ \boldsymbol{e}_{t_i}^a \gg \widehat{\boldsymbol{\sigma}}_{t_i}^{\beta} \\ &(iI) \quad \text{good } \sigma\text{-estimator}: \quad \forall\ (t_i, t)\ with\ t_i\ is\ anom,\ \frac{\boldsymbol{e}_{t_i}^a}{\widehat{\boldsymbol{\sigma}}_{t_i}^{\beta}} \gg \frac{\varepsilon_t^A + \varepsilon_t^E}{\widehat{\boldsymbol{\sigma}}_t^{\beta}} \\ &(iii) \quad \text{good } \mu\text{-estimator}: \quad \widehat{\boldsymbol{\sigma}}_t^{\beta} > B_t \quad \text{and} \quad B_t \Rightarrow 0 \end{aligned} \tag{9}$$

Under the hypotheses described by formula 9, the model & data noises are negligible compared to the anomaly error in an abnormal context, while the noises are lower than the detection threshold in the nominal situation due to normalization of the variance. Therefore, the anomaly score $\boldsymbol{S}_t$ discriminates ($|\boldsymbol{S}_t| > 1$) statistical anomalies from a contextual confidence interval defined for a confidence threshold $\beta$ with particular consideration for the risk of prediction unreliability according to a specified critical value $\alpha$ (**misprediction risk aversion**).

# C Additional mathemathical justification :

## C.1 Notation

**NOTATIONS**

| | |
|---|---|
| Basic term | |
| $t$ | index associated with a time step of a time series |
| $k$ | index associated to a member (submodel) of a metamodel |
| $N$ | number of samples (time series) |
| $m$ | number of submodels in a metamodel |
| $\widehat{a}$ | Estimation of the quantity a |
| $\overline{a}$ | Average of the quantity a (often used on a set of estimations) |
| ML Modeling variable | |
| $y_t$ | the value of the series $y$ (ground truth) at time step $t$ |
| $f$ | function expressing the explainable part of the series $y$ |
| $\boldsymbol{c}_t, \boldsymbol{h}_t$ | exogenous variables (observed, hidden) impacting the time series $y_t$ |
| $Y_t^{past}$ | lagged endogenous variables from series $y$ used for forecasting of future values |
| $\mathbf{x}_t$ | the input data at time stamps $t$ composed in our framework by $(\boldsymbol{c}_t, Y_t^{past})$ |
| $D, D_\theta$ | dataset composed of pairs $(X_1, y_1), .., (X_N, y_N)$, and subset of dataset used to train a model |
| $\Theta$ | Ensemble of parameters set |
| $\theta, \theta^i, \theta^*$ | parameters set (resp one, the $i^{th}$, the "optimal") from $\Theta$ |
| $\hat{f}_\theta(\mathbf{x}_t) = \hat{y}_t$ | a ML model (function approximation) parametrized by $\theta$ with $x$ as inputs and providing $\hat{y}$ as output |
| UQ modeling term | |
| $\delta$ | abstract diversity infusion mechanism allowing the exploration of $\Theta$ space (ex bootstrap). |
| $\mathcal{M}, \mathcal{M}^c, \mathcal{M}^d$ | metamodel, control metamodel, degraded metamodel impacted by a variability injection |
| $\varepsilon^u, \varepsilon^\Theta, \varepsilon^\theta$ | Noise (uncertainty) link respectively to upstream sources, modeling constraint, and epistemic issues |
| $\sigma_t^A, \sigma_t^E$ | Variance (Aleatoric, Epistemic), at step $t$ |
| model and meta-model output | |
| $\widehat{f}_\theta(\mathbf{x}_t) = (\hat{y}_t^\theta, \hat{\sigma}_t^{A,\theta})$ | Outputs (Forecast and aleatoric estimation) corresponding to submodel $\theta$ at time step $t$ |
| $M(\mathbf{x}_t) = (\overline{y}, \overline{\sigma}_t^A, \widehat{\sigma}_t^E)$ | Outputs (Forecast, aleatoric, epistemic estimation) corresponding to metamodel M $\theta$ at time step $t$ |
| Experimental modeling term | |
| $I_t^e$ | disentangled Epistemic indicator (dE-Indicator) |
| $X_n, X_a$ | nominal query without changes, altered queries that are degraded by variability injection |

## C.2 Biais-variance trade-off and total uncertainty law:

Taking up the formalization introduced in the paper, we introduced three aleatoric variable that can each be associated with a specific source of errors in the Biais-variance trade off.

- $\varepsilon_t^u = y_t - f(\mathbf{x}_t)$ the noise upstream the modeling scope associated with intrinsic irreducible variability errors, that corresponds to the difference between the series Y, and its explainable part $f$ according to the modeling scope $S$.

- $\varepsilon_t^{\Theta} = f(\mathbf{x}_t) - f_{\Theta}^*(\mathbf{x}_t)$ the modeling-constraint noise, related to bias errors due to the exploration space $\Theta$, that correspond to the difference between the explainable part $f-$ of data and the optimal reachable function $f_{\Theta}^*$ in $\Theta$.
- $\varepsilon_t^{\theta} = f_{\Theta}^*(\mathbf{x}_t) - \widehat{f}_{\theta}(\mathbf{x}_t)$ the epistemic noise, related to variance errors, that correspond to the difference between the optimal reachable function $f_{\Theta}^*$ in $\Theta$, and the obtained model $\widehat{f}_{\theta}$ which is biased by limited dataset $D_{\theta}$.

This modeling frame provides us with an artificial way to decompose the series $y$ according to a model, and its different sources of errors :

$$\text{With:} \quad y_t = f(\mathbf{x}_t) + \varepsilon_t^u = f_{\Theta}^*(\mathbf{x}_t) + \varepsilon_t^{\Theta} + \varepsilon_t^u = \widehat{f}_{\theta}(\mathbf{x}_t) + \varepsilon_t^{\theta} + \varepsilon_t^{\Theta} + \varepsilon_t^u \tag{10}$$

This modeling frame provides us with an artificial way to decompose the series $y$ according to a model, and its different sources of errors. It can then be rejected in the total law of uncertainty decomposition to analyze how this term impacts an aleatoric versus epistemic decomposition :

$$\begin{aligned}
\text{With} \quad \sigma(y_t|x_t;\theta) =& \sigma_{\Theta}\left[E_y(y_t|x_t;\theta)\right] + \boldsymbol{E}_{\Theta}\left[\sigma_y(y_t|x_t;\theta)\right] = \boldsymbol{\sigma_t^E} + \boldsymbol{\sigma_t^A} \\
\sigma_t^E \overset{\text{if ind.}}{=}& \underbrace{\sigma_{\Theta}[E_y[\widehat{f}_{\Theta}^{\theta_k}(\mathbf{x}_t)]]}_{\boldsymbol{E_{\Theta}}\left[\left(f_{\Theta}^*(\mathbf{x}_t) - \hat{f}_{\theta}(\mathbf{x}_t)\right)^2\right]} + \sigma_{\Theta}[\underbrace{E_y[\varepsilon_t^{\Theta}]}_{0^*}] + \sigma_{\Theta}[\underbrace{E_y[\varepsilon_t^u]}_{0^*}] + \sigma_{\Theta}[\underbrace{E_y[\varepsilon_t^{\theta}]}_{0^*}] \\
\sigma_t^A \overset{\text{if ind.}}{=}& E_{\Theta}[\underbrace{\sigma_y(f^{\theta_k}(\mathbf{x}_t))}_{0}] + \underbrace{\sigma_y(\varepsilon_t^{\Theta})}_{modeling} + \underbrace{\sigma_y(\varepsilon_t^u)}_{upstream} + \underbrace{E_{\Theta}[E_y(\varepsilon_t^{\theta} - \underbrace{E_y[\varepsilon_t^{\theta}]}_{0^*})^2]}_{0^*}
\end{aligned} \tag{11}$$

Two types of assumption allow making simplification. We can consider that the random variables are independent (ind.), obtaining independent terms, then consider that the different random variables are centered (for different reasons), which makes the terms negligible.

## C.3 dE-Indicator development

By manipulating dUQ framework under assumptions about aleatoric and epistemic uncertainty, we can design predictive indicators from metamodel likelihood using the assumption A1: the submodels ensemble of metamodel is well-formed, which means rich in variability $\forall (k,k')\ with\ k \neq k' \quad d(\theta^k,\theta^{k'}) > \epsilon$ and composed of submodels with good predictions and with few overfitting $\forall\ k,\ f_{\theta^k} \in \Theta$. Starting from the normal log-likelihood using total variance $\sigma^{tot}$ :

$$\begin{aligned}
ln(L(\overline{y}_t, \widehat{\sigma^{tot}}_t; \Theta, y_t)) &= ln(P(y_t; \overline{y}_t, \widehat{\sigma^{tot}}_t|\Theta)) \\
&\approx ln(P(y_t; \overline{y}_t, \widehat{\sigma^{tot}}_t|\{\theta^k\}_{k\in[1,m]})) \quad \text{A1 submodel-ensemble approx} \\
&= cst - ln(\widehat{\sigma^{tot}}_t) - \frac{(y_t - \overline{y}_t)^2}{2 * \widehat{\sigma^{tot}}_t} \quad \text{NLL developpement}
\end{aligned} \tag{12}$$

A local likelihood of the metamodel on the training set can be approximated under:

- the normal-aleatoric assumption $y_t \sim \mathcal{N}(\overline{y}_t, \overline{\sigma^a})$

- the normal-epistemic assumptions $\mu_t^* \sim \mathcal{N}(\overline{y}_t, \overline{\sigma^e})$
- the A2 assumption of a local unbiased capture based on the ability of the submodels to perform local estimation in a standard ML-framework.

We can derive the A2 assumption of a "local unbiased capture performed by submodels" by combining a standard ML-framework and normal-aleatoric assumption.

$$\text{A2: Local neigboor observation approx: } y_t \approx \int \sum_k f_{\theta^k}(\mathbf{x}_t)dx \approx \int y_t dy$$

$$\begin{aligned} ln(L(\overline{y}_t, \widehat{\sigma^{tot}}_t; \Theta, y_t)) &\approx cst - ln(\widehat{\sigma^{tot}}_t) - \frac{(y_t - \overline{y}_t)^2}{2 * \widehat{\sigma^{tot}}_t} \\ &\approx cst - ln(\widehat{\sigma^{tot}}_t) - \frac{(\int y_t dy - \overline{y}_t)^2}{2 * \widehat{\sigma^{tot}}_t} \\ &\approx cst - ln(\overline{\sigma^{tot}}_t) + 0 \\ &\approx cst - ln(\overline{\sigma^a}_t + \widehat{\sigma^e}_t) + 0 \\ L(\overline{y}_t, \widehat{\sigma^{tot}}_t | \Theta)) &\propto \frac{1}{\overline{\sigma^a}_t + \widehat{\sigma^e}_t} \end{aligned} \tag{13}$$

*Formula 5 + $y_t \sim$ A2*

*$\overline{y}_t$: Local MLE by NLL loss minization*

A local epistemic likelihood of the “average-submodel” mean among the reachable and relevant models $\Theta$ can also be obtained under the A3 assumption of “local unbiased mean modelling approximation” stemming from the combination of the A1 assumption and normal-epistemic assumptions.

$$\text{A3 Local mean modelling approx: } \mu_t^* \approx \int_\Theta \frac{\widehat{y}_t^k}{m} \approx \sum_{k=1}^{m} \frac{\widehat{y}_t^k}{m}$$

$$\begin{aligned} ln(L(\overline{y}_t, \widehat{\sigma^e}_t | \Theta, \mu_t^*) &= cst - ln(\widehat{\sigma^e}_t) - \frac{(\mu_t^* - \overline{y}_t)^2}{2 * \widehat{\sigma^e}} \\ &\approx cst - ln(\widehat{\sigma^e}_t) - \frac{\left(\sum_{k=1}^{m} \frac{\widehat{y}_t^k}{m} - \overline{y}_t\right)^2}{2 * \widehat{\sigma^e}} \\ &= cst - ln(\widehat{\sigma^e}_t) \ - 0 \\ L(\overline{y}_t, \widehat{\sigma^e}_t | \Theta) &\propto \frac{1}{\widehat{\sigma^e}_t} \end{aligned} \tag{14}$$

*Formula 5 + $\mu_t^* \sim$ A3*

*$\overline{y}_t$: local MLE*

Finally, we are more specifically interested in the fluctuation of the local epistemic likelihood of the decision regarding local forecasting uncertainty to be more “invariant” to the magnitude of local forecasting uncertainty. Therefore, we propose to analyze the local epistemic likelihood indicator (Formula 14) normalized by the local total likelihood indicator (Formula 13).

$$\begin{aligned} \frac{L^{epi}(\overline{y}_t, \widehat{\sigma^e}_t | \Theta)}{L^{tot}(\overline{y}_t, \widehat{\sigma^{tot}}_t | \Theta)} &\propto \frac{\overline{\sigma^a}_t + \widehat{\sigma^e}_t}{\widehat{\sigma^e}_t} \\ ln(\frac{L^{epi}}{L^{tot}}) &\propto ln(\overline{\sigma^a}_t + \widehat{\sigma^e}_t) - ln(\widehat{\sigma^e}_t) \\ &\propto ln(1 + \frac{\widehat{\sigma^a}_t}{\overline{\sigma^e}_t}) \end{aligned} \tag{15}$$

# References


[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *Osdi*, volume 16, pp. 265–283. Savannah, GA, USA, 2016.

[2] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 2021.

[3] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33:14927–14937, 2020.

[4] Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. *arXiv preprint arXiv:2002.06470*, 2020.

[5] Sheraz Aslam, Herodotos Herodotou, Syed Muhammad Mohsin, Nadeem Javaid, Nouman Ashraf, and Shahzad Aslam. A survey on deep learning methods for power load and renewable energy forecasting in smart microgrids. *Renewable and Sustainable Energy Reviews*, 144:110992, 2021.

[6] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pp. 1613–1622. PMLR, 2015.

[7] Ian C. Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: a unified framework for model explanation. *The Journal of Machine Learning Research*, 22(1):209:9477–209:9566, January 2021. ISSN 1532-4435.

[8] Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, pp. 1184–1193. PMLR, 2018.

[9] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

[10] Timothy Dozat. Incorporating nesterov momentum into adam. 2016.

[11] Romain Egele, Romit Maulik, Krishnan Raghavan, Bethany Lusch, Isabelle Guyon, and Prasanna Balaprakash. Autodeuq: Automated deep ensemble with uncertainty quantification. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pp. 1908–1914. IEEE, 2022.

[12] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

[13] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*, 2021.

[14] Alex Graves. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24, 2011.

[15] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

[16] Giles Hooker, Lucas Mentch, and Siyu Zhou. Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31(6):82, October 2021. ISSN 1573-1375. doi: 10.1007/s11222-021-10057-z. URL `https://doi.org/10.1007/s11222-021-10057-z`.

[17] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.

[18] Thanasis Kotsiopoulos, Panagiotis Sarigiannidis, Dimosthenis Ioannidis, and Dimitrios Tzovaras. Machine learning and deep learning in smart manufacturing: The smart grid paradigm. *Computer Science Review*, 40:100341, 2021.

[19] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[20] Chan-Chiao Lin, Huei Peng, Soonil Jeon, and Jang Moo Lee. Control of a hybrid electric truck based on driving pattern recognition. In *Proceedings of the 2002 advanced vehicle control conference, Hiroshima, Japan*, pp. 9–13, 2002.

[21] Jeremiah Liu, John Paisley, Marianthi-Anna Kioumourtzoglou, and Brent Coull. Accurate uncertainty estimation and decomposition in ensemble learning. *Advances in neural information processing systems*, 32, 2019.

[22] David John Cameron Mackay. *Bayesian methods for adaptive models*. California Institute of Technology, 1992.

[23] Nis Meinert and Alexander Lavin. Multivariate deep evidential regression. *arXiv preprint arXiv:2104.06135*, 2021.

[24] Nicolai Meinshausen and Greg Ridgeway. Quantile regression forests. *Journal of machine learning research*, 7(6), 2006.

[25] Mouhcine Mendil, Marc Mossina, Lucaand Nabhan, and Kevin Pasini. Robust gas demand forecasting with conformal prediction. In *Conformal and Probabilistic Prediction with Applications*, pp. 169–187. PMLR, 2022.

[26] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[27] Giovanna Nicora, Miguel Rios, Ameen Abu-Hanna, and Riccardo Bellazzi. Evaluating pointwise reliability of machine learning prediction. *Journal of Biomedical Informatics*, 127:103996, 2022.

[28] Dongpin Oh and Bonggun Shin. Improving evidential deep learning via multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7895–7903, 2022.

[29] Kevin PASINI. *Forecast and anomaly detection on time series with dynamic context : Application to the mining of transit ridership data*. PhD thesis, Université Gustave Eiffel, 2021.

[30] Kevin Pasini, Mostepha Khouadjia, Allou Samé, Martin Trépanier, and Latifa Oukhellou. Contextual anomaly detection on time series: a case study of metro ridership analysis. *Neural Computing and Applications*, 34(2):1483–1507, 2022.

[31] Tim Pearce, Mohamed Zaki, Alexandra Brintrup, N Anastassacos, and A Neely. Uncertainty in neural networks: Bayesian ensembling. *stat*, 1050:12, 2018.

[32] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

[33] Yuan Wang, Dongxiang Zhang, Ying Liu, Bo Dai, and Loo Hay Lee. Enhancing transportation systems via deep learning: A survey. *Transportation research part C: emerging technologies*, 99:144–163, 2019.

[34] Florian Wenzel, Jasper Snoek, Dustin Tran, and Rodolphe Jenatton. Hyperparameter ensembles for robustness and uncertainty quantification. *Advances in Neural Information Processing Systems*, 33: 6514–6527, 2020.

[35] Lingxue Zhu and Nikolay Laptev. Deep and confident prediction for time series at uber. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 103–110. IEEE, 2017.