NAIST-IS-MT1351119

Master's Thesis

Robust parameter estimation in wind power forecasting

Matthew James Holland

March 12, 2015

Department of Information Science
Graduate School of Information Science
Nara Institute of Science and Technology

A Master's Thesis
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
MASTER of ENGINEERING

Matthew James Holland

Thesis Committee:

Professor Kazushi Ikeda	(Supervisor)
Professor Yuji Matsumoto	(Co-supervisor)
Assistant Professor Takatomi Kubo	(Co-supervisor)

Robust parameter estimation in wind power forecasting*

Matthew James Holland

Abstract

In this research, we consider the task of short-term (<6h horizon) and very short-term (<10s horizon) prediction of wind speeds using parametric probabilistic models, with a particular focus on parameter estimation methods. For short-term forecasts, relatively accurate predictive distributions are well-known, though performance tends to vary significantly across seasons and locations. At higher temporal resolutions on the order of a few seconds, the literature becomes sparse and effective methods remain to be proposed.

Our first contribution is a systematic method for deriving families of estimators which have theoretical properties making them suitable for automated estimation tasks via minimization of tractable loss functions. We provide examples derived using this framework which are parsimonious, can be used for real-time forecasting tasks, and are computable for many important classes of parametric models. The second key contribution is a thorough empirical evaluation of the predictive accuracy and robustness of forecasters as a function of estimators used, with the aim of elucidating the potential for improving upon standard methods. We report positive results showing superior performance across all forecast horizons, giving evidence for the robustness of the proposed forecasters to spatio-temporal condition changes, and suggesting a more broad utility beyond the domain considered.

Keywords:

Parameter estimation, wind power forecasting, density estimation, proper loss functions, minimum divergence estimators

* Master's Thesis, Department of Information Science, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-MT1351119, March 12, 2015.

風況の確率モデルにおける頑健な推定方法と風力発電への応用*

Matthew James Holland

内容梗概

極端気象の事前予知や風力発電所の風車制御等において重要な役割を果たすのは、風速の確率的予測モデルである。約 6 時間超の長期予測ならば安定した性能をもつ物理モデルは確立しているが、短期予測（6 時間未満）では位置や季節に敏感な局所的な方法しか提案されていない。さらに、超短期予測（10 秒未満）では文献上の知見がほとんどなく、予測地点ごとの気候や地形の違いに対応できる頑健な方法は依然として提案されていない。

この現状および既存の知見を踏まえて、本研究では先行研究で重視されてきたモデル構築方法ではなく、モデルのパラメータ推定方法に主眼をおき、最適化問題として定式化しやすい推定量を体系的に導出する新しい方法を提案している。実用性を確認するために、風速予測でよく用いられるモデルを所与として、提案方法を用いて得られた推定量を既存の手法と比較し、推定量が予測誤差、分布の適合性、また予測器の時空間的頑健性に及ぼす影響を定量的に評価した。その結果、気象データを用いた実験課題では、提案手法が予測期間によらず比較対象より優れた成績を残し、推定方法と頑健性の関係を示唆しつつ、有望な予測方法が示された。

キーワード

パラメータ推定, 風力発電量予測, 密度推定, 正当なロス関数, ダイバージェンス最小化推定量

* 奈良先端科学技術大学院大学 情報科学研究科 情報科学専攻 修士論文, NAIST-IS-MT1351119, 2015 年 3 月 12 日.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis is primarily concerned with the problem of methodically designing and automatically optimizing mathematical models of wind velocity at a future point in time, expressed as a function of observations available at the present. As a broad context for this work, it naturally falls within the realm of problems dealt with by the large body of techniques explicitly designed for dealing with time-series data (Box et al., 2008). More specifically however, we look at a probabilistic representation of the system of interest, wherein future wind speed is a random variable, whose stochastic traits are completely described by a function called a probability distribution. As is often the case with probabilistic models, the methods proposed in Chapter 2 are quite general and indeed may readily be applied to different domains of interest. In any case, our focus here is on the efficacy of the proposed methods with respect to wind power forecasting, wind farm operations management, and wind turbine control, among other applications related to wind-based energy. As such we begin by giving a background on the role of forecasting in wind energy.

1.1 Background on forecasting in wind energy

To motivate the need for our research, we begin by looking at wind-based energy generation, transmission, and supply. Winds are most simply observed by humans as spatial non-uniformities in the velocity (speed and direction) of particles which compose the gases of the lower atmosphere. The most fundamental origins of wind are thermal effects (uneven heating of the atmosphere) and Coriolis forces. With respect to the former, temperatures of atmospheric gases are

non-uniform across the surface of the earth; as regions of relatively high and low air pressure form, the pressure gradient force is directed from high to low pressure regions, and naturally acts on air particles. Coriolis forces are due to the rotation of the earth; a particle moving from the equator towards either of the poles gets nearer to the axis of rotation, and thus the Coriolis force acts in an eastward direction ("right" when moving north, "left" when moving south). A useful introduction to these topics is from Wallace and Hobbs (2006).

Given the fundamental forces that give rise to winds, it is clear that winds shall continue to blow for the foreseeable future, and thus the appeal of wind as a source of power is indeed intuitive. The task is then to convert the kinetic energy of the wind into electricity. Considering a horizontal-axis wind turbine (Fig. 1.1) with radius (m) $r > 0$ and surrounding air density (m/kg^3) ρ, a simple standard model (Slootweg et al., 2003) for instantaneous turbine power output at time t is given by

$$P(x_t, \lambda_t, \beta_t) = \frac{1}{2}\pi r^2 x_t^3 C(\lambda_t, \beta_t), \tag{1.1}$$

where x_t is scalar wind speed (m/s), β_t is blade pitch (rad), $\lambda_t = \omega_t r / x_t$ is tip-speed ratio with ω_t being rotation speed (rad/s), and C is a performance coefficient modelled as a function of pitch and tip-speed ratio. Seminal work on the capacity of wind turbines to extract power from the wind dates back to a 1919 paper from Albert Betz, who gave an upper bound on the ratio $2P(x_t, \lambda_t, \beta_t)/\pi r^2 x_t^3$ and thus on the function C of $16/27 \approx 0.59$. The important fact here is that power output is proportional to x_t^3, and while the model is indeed simplistic, even in more sophisticated approaches the sensitivity of power to wind speed remains, implying that in order to reliably forecast future power output from a given turbine, reliable forecasts of wind speed will be required. A detailed discussion of the aerodynamics of wind turbines is out of scope, but we refer the interested reader to any of the standard introductory texts, such as Burton et al. (2011).

We now begin a more in-depth literature review, beginning with wind-related forecasting in general, and then shifting our focus explicitly to the wind power literature in order to fully establish the context for our work.

The first critical element of the forecasting problem is the *horizon* length (Soman et al., 2010). If time is given in discrete steps, a forecast $\widehat{x}_{t+k}$ using data collected up to time t is a forecast with horizon $k > 0$. Short-term forecasts are typically less than 6 hours, and forecasts on the order of a few seconds up to a few minutes may be described with an additional qualifier as "very short-term" forecasts. For our purposes, forecasts with horizons beyond 6–12 hours may be

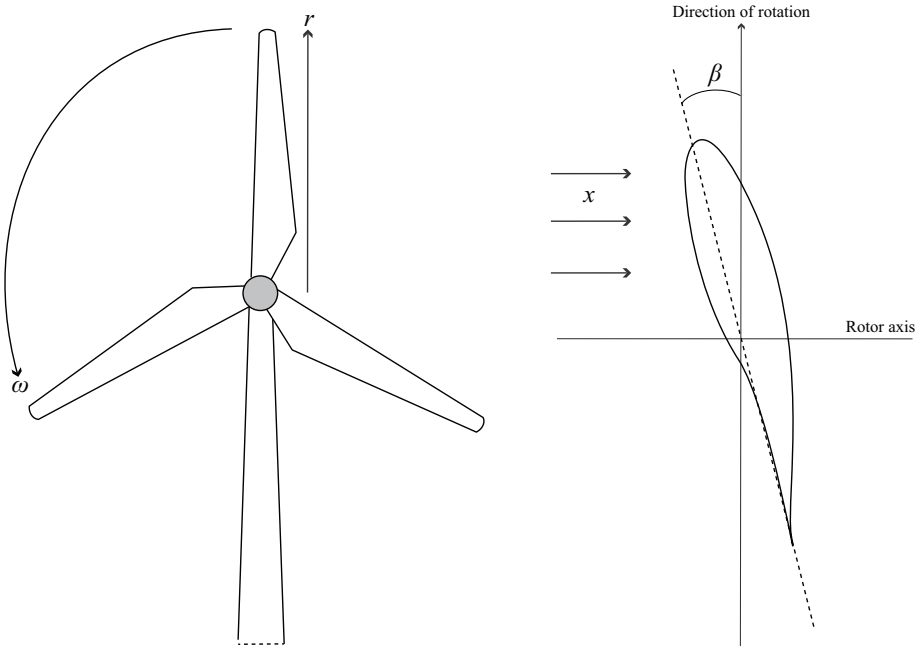

Figure 1.1: A schematic of a typical horizontal-axis wind turbine.

grouped together and simply labelled "long-term." For the reader interested in notions of complexity and non-linear systems, a classic study by Lorenz (1969) gives an upper bound on the forecast horizon within which deterministic forecasts can achieve a minimum required level of performance. The important modern dichotomy is between short and long horizons (Pinson, 2012), since the standard methods used differ significantly.

Let us now consider the value that sufficiently accurate, reliable predictions of wind speed (and/or resulting wind power output) have in industrial applications. A prominent role for very short-term forecasts exists in anticipatory control schemes, where generator torque and blade pitch are controlled so as to minimize nacelle vibration and maximize power output (Kusiak et al., 2010; Boukhezzar and Siguerdidjane, 2011; Kusiak and Zhang, 2012). We note that excess or irregular vibration of components shortens the useful life of a turbine (Walford, 2006), and as the automatic control strategies depend on accurate wind velocity forecasts, there is a clear economic incentive to develop reliable, local short-term forecasting technology. Increasing the forecast horizon slightly to the order of several hours, from the perspective of power suppliers selling energy in deregulated power markets, accurate forecasts of output over a relatively short horizon

can be used to optimize both supply contract volumes committed and the timing of said commitments (Pinson et al., 2007; Zugno et al., 2013). Numerous case studies which reflect region-specific characteristics in the relevant market models have been carried out, for regions including the Netherlands (Hutting and Cleijne, 1999), Denmark (Nielsen et al., 1999; Sørensen and Meibom, 1999), Greece (Kariniotakis et al., 1999), the USA (Makarov et al., 2009), UK (Bathurst et al., 2002), and Spain (Fabbri et al., 2005). A more comprehensive literature review is provided by Giebel et al. (2011).

Now that we have established the role that wind speed and power forecasting can play in a variety of applications, we examine the methods by which this is typically done. We start with long-term forecasts, and then consider the short-term case.

For longer-range forecasts, *numerical weather prediction* (NWP) is the name attached to the collection of methods typically used (Lynch, 2008). Major developments in physics in the 19th century paved the way for seminal work at the start of the 20th century. A key work is from Bjerknes (1904), who introduced a formulation of the problem of predicting future atmospheric states; namely, this was an introduction of systems of differential equations that allowed the analyst to associate measurable quantities such as pressure, density, and velocity with key equations expressing hydrodynamic motion and thermodynamic laws. The key issue then was intractability of the expressions given a set of observations, and the work of Lewis Fry Richardson (Lynch, 2006) during the same period contributed to the development of an experimental methodology in which real data could be used with the formulation due to Bjerknes by means of numerically derived approximate solutions. The true utility of NWP was not realized until the advent of the modern computer, and the birth of reliable NWP technology can be traced back to the work of John von Neumann's computing machinery team at Princeton University's Institute for Advanced Study (Charney et al., 1950). As computing power has exponentially increased over the past half-century, so has the utility of NWP (Kalnay et al., 1998), though for forecasts (often called "nowcasts" in the NWP context) of high spatial resolution at short time horizons, the body of physical models used in NWP cannot be used effectively, even on the order of a few hours, much less minutes or seconds (Lynch, 2008; Pinson, 2012).

For short-range forecasts, the class of tools considered broadens dramatically compared with the long-term case, though several major categories can be identified. Typical methods used for time-series data such as autoregressive (AR) models (Brown et al., 1984; Baïle et al., 2011), AR models incorporating a moving average (Kamal and Jafri, 1997), AR models which methodically incorporate

non-stationarity (ARIMA) (Kavasseri and Seetharaman, 2009), and Kalman filters (Bossanyi, 1985) all have seen numerous applications in case studies with varying degrees of success. A similar statement can be made for neural network (NN) based methods (Alexiadis et al., 1998; Bechrakis and Sparis, 1998; Potter and Negnevitsky, 2006), which allow for arbitrarily complex expressions of future wind speed by typically considering linear combinations of smooth functions which are non-linear in the inputs of interest. Models of the distribution of wind speed have been studied for many years, with Weibull, log-Normal, Rayleigh, and Gamma distributions being standard choices in locations around the world (Justus et al., 1976; Conradsen et al., 1984), with recent studies confirming in particular the utility of the bi-modal Weibull (Morgan et al., 2011). In what may be considered a natural development, this knowledge of the stochastic character of wind speeds has been applied to create successful forecasters by explicitly formulating the forecasting task as a density estimation problem, wherein a parametric distribution for future wind is specified and returned to the user as final output (Gneiting et al., 2006; Hering and Genton, 2010; Pinson, 2012).

The application of our work is most closely related to the above density forecasters, and in the following section we shall make more precise the issues this research seeks to provide solutions to, and summarize the degree to which we have been successful in doing so.

1.2 Contributions of this research

For very short-term forecasts (here < 10s horizon), the literature remains exceedingly sparse (Jiang et al., 2013), and the problem of a probabilistic forecaster which yields real-time forecasts superior to trivial benchmarks can safely be considered an open problem. For short-term forecasts (< 6h horizon), while there have been numerous successful probabilistic forecasters proposed (Brown et al., 1984; Gneiting et al., 2006; Pinson, 2012), even comparatively simple methods which have minimal site dependence have been shown to have performance which varies greatly by location and season (Hering and Genton, 2010). That is, while the stochastic traits of wind speed are quite well-understood (Hennessey, 1977), there is no method (either using this insight or not) which uniformly outperforms standard references across locations and seasons. This naturally implies that no non-trivial benchmark method exists. This would not be a problem if for any arbitrary site a reasonably good method could be found, even in an *ad hoc* manner, but in fact the sites selected in the above-referenced literature are those which were

particularly amenable to *a priori* analysis or for which valuable meteorological insights were available prior to model design. Given the evidence in the literature at present then, it is difficult to consider the methods proposed thus far sufficient for the general problem of forecasting at an arbitrary site, even under the constraint that wind speed observations are available.

We contend that a major factor which has been overlooked in the applied literature is the role of the parameter estimation applied. That is to say, it may be that the predictive distribution models considered in the literature are in fact viable at arbitrary locations, with the true issue being a lack of robustness in the estimator used to determine the model. We focus on this point explicitly, and one of the main contributions of this research is a systematic method for constructing new estimators by defining them in terms of members of a class of functions with desirable and easy to prove properties. Several pertinent estimators are derived as examples using no information specific to wind speed. As a competitive reference for Weibull-based models, a closed-form CRPS estimator is also derived. Another of the chief contributions lies in a thorough empirical evaluation of the predictive performance realized by the proposed estimation method compared with several competitive reference methods. For the very short-term task we make use of high-frequency (7.3Hz) anemometer data from the Google Heliostat project, and for the short-term task we utilize a large subset of the nationwide AMeDAS observation network operated by the Japan Meteorology Agency. Details of the references and proposed methods are given in Chapter 2, while experimental results and subsequent discussion is given in Chapter 3. The thesis closes with a summary of the insights gained and conclusions drawn from this work, as well as a look ahead at potential lines of related work in Chapter 4.

Chapter 2

Methods

2.1 Preliminaries

In this sub-section we establish notation, define some key basic concepts, give some intuitive background and discuss the literature related to the methods which will be proposed in proceeding sub-sections.

For our purposes, we will be interested in the stochastic behaviour of some observable variables $\mathbf{x}$ taking values in $\mathcal{X}$ which we assume to be some normed linear space, typically a subset of $\mathbb{R}^d$ for integer $d \geq 1$ (we shall re-use d frequently, but it should not be taken as common). Any distinction between response variables and covariates will be made clear given a particular modelling context. Formally, the description of this stochastic behaviour of interest is characterized by real functions called probability measures, defined on a set whose elements are subsets of $\mathcal{X}$. Let $\mathcal{A}$ be such a class, where any $E \in \mathcal{A}$ is $E \subset \mathcal{X}$. If $\mathcal{A}$ is closed under countable unions as well as complements of its elements, we call $\mathcal{A}$ a *σ-algebra*, and call $(\mathcal{X}, \mathcal{A})$ a *measurable space*. A non-negative real function $\rho : \mathcal{A} \to \mathbb{R}$ is called a *measure* if it satisfies $\rho(\emptyset) = 0$ and countable additivity, i.e., for any disjoint sequence $\{E_n\}, E_n \in \mathcal{A}, n = 1, 2, \ldots$ we have $\rho(\bigcup E_n) = \sum_n \rho(E_n)$. A measurable space $(\mathcal{X}, \mathcal{A})$ equipped with a valid measure ρ is called a *measure space*. Let $\mathbb{M}(\mathcal{X}, \mathcal{A})$ denote the set of all measures on $(\mathcal{X}, \mathcal{A})$. A *probability measure* on $(\mathcal{X}, \mathcal{A})$ is any member of $\mathbb{P} \coloneqq \{P \in \mathbb{M}(\mathcal{X}, \mathcal{A}) : P(\mathcal{X}) = 1\}$, and $(\mathcal{X}, \mathcal{A}, P)$ for any $P \in \mathbb{P}$ is typically called a *probability space*. We shall use the term (probability) *distribution* interchangeably with probability measure. The basic properties of measures on σ-algebras are described lucidly by Halmos (1974).

C2: *METHODS*

We want to describe unambiguously the probabilistic characteristics of $\mathbf{x}$ as accurately as possible. The range $[0, 1]$ and the monotonicity of any P on $\mathcal{A}$ intuitively suggests its suitability for quantifying the likelihood of stochastic events, specified by $E \in \mathcal{A}$. That is $P(E) \in [0, 1]$ gives us a real number representation of the probability that a particular observation satisfies $\mathbf{x} \in E$. If there is no randomness at all (over time or simply across distinct observations), this should be apparent in the measurements taken, and the problem becomes trivial. If uncertainty exists, then we make the assumption that there exists *some* appropriate distribution $P \in \mathbb{P}$ which accurately describes the system of interest, with the caveat that in general, this P will be unknown. The task then is to approximate the true distribution given observations. To do this, we often limit our focus to particular subsets of $\mathbb{P}$, which we will denote $\mathcal{P} \subset \mathbb{P}$ and call (probabilistic) *models*. Should this model include the true distribution (or even a sufficiently good approximation), then to automate the process of finding this approximation, it is often useful to evaluate a given estimate P by assigning to it a real number which essentially evaluates how good/bad the estimate is to the true distribution, an evaluation of their dissimilarity. Stated very informally, by some algorithm we then adjust our estimate such that when we recalculate the dissimilarity metric, it comes close to some optimal value; in many cases, this naive approach indeed brings us close to the true distribution, since while unknown, the observations we have naturally are generated by this distribution.

The focus of this thesis is on how to evaluate the above dissimilarity. Given some $P, Q \in \mathcal{P}$, equality is simple as $P = Q$ is equivalent to $P(E) = Q(E)$ for all $E \in \mathcal{A}$, though if $P \neq Q$ and $P' \neq Q$, which of P and P' is less similar, that is to say, more "divergent" from Q? To answer this question quantitatively, it will often be convenient to make use of density functions associated with probability measures. If ρ is a σ-finite measure on $\mathcal{A}$, and $P \ll \rho$ (absolute continuity with respect to ρ), then the Radon-Nikodym theorem gives us a measurable function $p : \mathcal{X} \to \mathbb{R}$ such that

$$P(E) = \int_E p \, d\rho, \quad \forall \, E \in \mathcal{A}$$

where p is unique up to sets of zero measure $[\rho]$. The modulo-ρ "almost everywhere" qualifier $[\rho]$ should be assumed and will not be repeated below. Henceforth, for arbitrary $P, Q \in \mathbb{P}$ we shall assume the existence of ρ such that $P \ll \rho$ and $Q \ll \rho$, with densities $p = dP/d\rho, q = dQ/d\rho$. Informally we may use the terms measurable function and random variable interchangeably. It should be noted that with respect to the former, given σ-algebra $\mathcal{A}$, if $f^{-1}(S) \in \mathcal{A}$ for all

Borel subsets of $\mathbb{R}$ we say f is $\mathcal{A}$-measurable, and while the latter is certainly a measurable function, notions of integrability are generally implied. When definable, we shall denote the mean vector and covariance matrix of $\mathbf{x}$ by $\boldsymbol{\mu}_P \coloneqq \mathbb{E}_P[\mathbf{x}]$ and $\mathbf{V}_P = \mathbb{E}_P[\mathbf{x}\mathbf{x}^T] - \mathbb{E}_P[\mathbf{x}]\mathbb{E}_P[\mathbf{x}]^T$. Let $\mathbb{S}^d$ denote the set of all symmetric real $d \times d$ matrices, and $\mathbb{S}_+^d \subset \mathbb{S}^d$ all those $\mathbf{A} \in \mathbb{S}^d$ which are positive definite, i.e., $\langle \mathbf{x}, \mathbf{A}\mathbf{x} \rangle > 0, \forall\, \mathbf{x} \in \mathcal{X}, \mathbf{x} \neq 0$. Notation for some models of interest will be

$$\begin{aligned}\mathcal{P}^1(\Theta_a) &= \{P \in \mathbb{P} : \boldsymbol{\mu}_P \in \Theta_a\} \\ \mathcal{P}^2(\Theta_b) &= \{P \in \mathbb{P} : \mathbb{E}_P[\mathbf{x}\mathbf{x}^T] \in \Theta_b\} \\ \mathcal{P}(\Theta_a, \Theta_b) &= \{P \in \mathcal{P}^1(\Theta_a) : \mathbf{V}_P \in \Theta_b\}.\end{aligned}$$

For $\delta > 0$ we denote the δ-radius open ball $B_\delta(\mathbf{x}_0) \coloneqq \{\mathbf{x} : |\mathbf{x} - \mathbf{x}_0| < \delta\}$. If $S \subset \mathcal{X}$, the interior and boundary of S are defined in the usual way as $\text{int}(S) = \{\mathbf{x} : \exists \delta > 0, B_\delta(\mathbf{x}) \subset S\}$ and $\text{bd}(S) = \overline{S} \setminus \text{int}(S)$ respectively, where $\overline{S}$ denotes the closure of S. We call $d : V \times V \to \mathbb{R}$ a *divergence* on V, where V is an arbitrary set, if $d(u, v) \geq 0$ for all $u, v \in V$ and $d(u, v) = 0$ iff $u = v$. Since d is not necessarily assumed to equal a valid norm of $u - v$, symmetry and the triangle inequality for d on $\mathcal{X}$ are also not assumed. While clear from the definition, we note this concept is entirely distinct from the divergence operator used in the analysis of vector fields. This general concept is natural as a means to quantify dissimilarity in the case $V = \mathcal{P}$, and we now consider some of the important literature related to divergences and statistical inference.

The theory of information provides the stage for many foundational developments; additivity of information metrics with respect to stochastically independent events is considered an axiomatic requirement (Barnard, 1951), naturally leading to logarithmic representations of information (Kullback, 1968). The best-known is assuredly the "directed divergence"

$$d_I(P, Q) = \int p(x) \log \frac{p(x)}{q(x)} d\mu(x) = \mathbb{E}_P \left[\log \frac{p(x)}{q(x)} \right],$$

called the *I-divergence* or *Kullback-Leibler divergence*. A related, family of divergences which is very general comes from Csiszár (1972), called *f-divergences* and defined

$$d_f(P, Q) = \int q(x) f\left(\frac{p(x)}{q(x)} \right) d\mu(x),$$

with the constraint that f be convex on $(0, \infty)$ and strictly convex at 1. Note that $d_f(P, Q) \geq f(1)$ and equality holds iff $P = Q$ can be readily verified. Some of the

other useful properties of d_f include convexity in both arguments and invariance to bijective transformations of random variables (Csiszár, 2008). If $f(x) = -\log(x)$ then clearly $d_I(Q, P) = d_f(P, Q)$.

Unsurprisingly, d_I and d_f are often used to explicitly look at dissimilarity on $\mathbb{P}$. Another class of divergences, called ***Bregman divergences*** are a broad class of divergences usually defined on open subsets of $\mathbb{R}^d$, and characterized by a convex, differentiable function f. Denoted d_B, the Bregman divergence with respect to such an f is defined

$$d_B(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle.$$

Effecient algorithms for computing "distances" which can be expressed by d_B for some f were originally proposed by Bregman (1967). In an axiomatic characterization of logically consistent methods for selecting elements from a given set by minimizing some form of divergence, Csiszár (1991) showed that only d_f and d_B are permissible given a set of natural, intuitive assumptions. Bregman divergences naturally appear when certain regularity conditions are imposed to large classes of loss metrics defined using convex functions, and as a result the related literature is broad and spans many fields. A detailed discussion in the statistical decision theory framework given by Grünwald and Dawid (2004). Recently, attention has been paid to d_B in the field of machine learning. Particularly noteworthy works includes early papers from Lafferty et al. (1997) and Collins et al. (2002), and interesting work from Banerjee et al. (2005) who showed that under certain conditions, Bregman divergences $d_B(\mathbf{x}, \mathbf{y})$ may be characterized as being the set of measurable functions for which given $\mathbf{x} \sim P$ the infimum of $\mathbb{E}_P[d_B(\mathbf{x}, \mathbf{y})]$ as a function of $\mathbf{y}$ is uniquely given by the conditional expectation of $\mathbf{x}$ given $\mathbf{y}$.

We also note a great deal of work has been done looking at theoretical unifications of seemingly disparate divergences. A comprehensive overview from Cichocki and Amari (2010) gives a detailed discussion of the relationships between α-divergences (Amari, 2009), β-divergences (Minami and Eguchi, 2002), γ-divergences (Fujisawa and Eguchi, 2008), d_f and d_B as given above. Many well-known metrics have been shown to be special cases of d_f, d_B, or both (this applies to both d_I and d_α). Related work in the differential geometry context is from Zhang (2004), and Hein and Bousquet (2005) focus on positive definite kernels defined on sets of probability measures.

2.2 Minimum divergence estimators

Functions with divergence-like properties have natural applications in problems of statistical inference. Most generally, given $\mathcal{P} \subset \mathbb{P}$ and a valid divergence $d : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$, the model is then trivially identifiable (Wooldridge, 2010) as $Q = \arg\min_{P \in \mathcal{P}} d(P, Q)$ uniquely achieves the minimum. Letting $Q \in \mathcal{P}$ be the unknown true distribution with observable random variable $\mathbf{x} \sim Q$ on $\mathcal{X}$, then for the practitioner a far more useful sub-class of divergences on $\mathcal{P}$ are those equivalent to the expectation of some measurable function D on $\mathcal{X}$, here denoted $d(P, Q) = \mathbb{E}_Q[D(\mathbf{x}; P, Q)]$. A particularly nice case is when we have

$$\arg\min_{P \in \mathcal{P}} \mathbb{E}_Q[D(\mathbf{x}; P, Q)] = \arg\min_{P \in \mathcal{P}} \mathbb{E}_Q[\lambda(\mathbf{x}, P)]$$

for all $Q \in \mathcal{P}$. That is we need only minimize the expectation of some "loss" function λ, where the expectation is taken with respect to Q. The reason for the utility of this form is intuitive; given observations $\mathbf{x}_1, \ldots, \mathbf{x}_N$, one might instinctively consider the estimator

$$P_N = \arg\min_{P \in \mathcal{P}} \sum_{n=1}^{N} \lambda(\mathbf{x}_n, P)$$

which is a particular M-estimator (Huber, 1981). Given regularity conditions on λ, intuition turns out to be correct as the law of large numbers gives us convergence of sequence (P_N) to Q as $N \to \infty$, which is to say that P_N is a consistent estimator of Q. Similarly, the behaviour of $\mathrm{Var}_Q[P_N]$ as a function of N is a means of evaluating the efficiency of the estimator.

Should we start with a valid divergence on $\mathcal{P} \times \mathcal{P}$ of the desirable sub-class noted above, the model is identifiable and we need only concern ourselves with the properties of the expected λ-minimizing estimator. In this case, we would likely evaluate the estimator as a function of loss function λ, comparing estimators $P_N(\lambda)$ over possible λ. In this research however, we are interested in the generation of *new* estimators, with no convenient divergence to begin with. Thus, taking a bottom-up approach, we start with only a loss function $\lambda : \mathcal{X} \times \mathcal{P} \to \mathbb{R}$, and by specifying its desirable form and properties, we shall systematically seek out divergence-like functions. To motivate the subsequent discussion, note that in the case of I-divergence d_I, we have

$$\begin{aligned} \arg\min_{P \in \mathcal{P}} \mathbb{E}_Q[\lambda(\mathbf{x}, P)] &= \arg\min_{P \in \mathcal{P}} \mathbb{E}_Q[\Delta_\lambda(P, Q)] \\ &= \arg\min_{P \in \mathcal{P}} d_I(Q; P) \end{aligned}$$

where $\Delta_\lambda(P,Q) \coloneqq \lambda(\mathbf{x},P) - \lambda(\mathbf{x},Q)$ and here $\lambda(\mathbf{x},P) = -\log p(\mathbf{x})$. This particular case of D, that is $D(\mathbf{x};P,Q) = \Delta_\lambda(P,Q)$, suggests a very appealing and simple way to attempt to construct divergences. Ideally one would like to show that for some λ, that $\mathbb{E}_Q[\Delta_\lambda(P,Q)]$ is a valid divergence on $\mathcal{P}\times\mathcal{P}$. We note that we shall consider a larger class of loss functions and thus a larger class of estimators as well.

We begin by noting an important property that we shall seek in loss functions.

Definition 1. Given $\lambda : \mathcal{X}\times\mathcal{P} \to \mathbb{R}$, we shall say that λ is $\mathcal{P}$*-proper* if

$$\mathbb{E}_Q[\Delta_\lambda(P,Q)] \geq 0, \quad \forall\, P,Q \in \mathcal{P} \tag{2.1}$$

and similarly, λ is *strictly* $\mathcal{P}$*-proper* if (2.1) holds and

$$\mathbb{E}_Q[\Delta_\lambda(P,Q)] = 0 \iff P = Q. \tag{2.2}$$

Clearly, $d_\lambda(P,Q) \coloneqq \mathbb{E}_Q[\Delta_\lambda(P,Q)]$ is a valid divergence if and only if λ is strictly $\mathcal{P}$-proper. If only (2.1) holds for λ, then we shall call d_λ a *quasi-divergence* on $\mathcal{P}$. In this paper we seek a systematic means of deriving quasi-divergences, and wherever possible, divergences. We note that while the term quasi-divergence has appeared in classical fluid dynamics literature with relevant physical interpretations, our definition here is plainly a weak version of the divergence defined above.

Next we consider the characteristics of the models of explicit interest to us. Informally, a parametric model is a subset of $\mathbb{P}$ specified by some finite-length $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$ for integer $d \geq 1$. A more formal formulation may begin with Θ, and consider a *parametrization* denoted $u : \Theta \to \mathbb{P}$, and defined $u(\boldsymbol{\theta}) \coloneqq \sqrt{p}$ for some $p = dP/d\mu \in L_2(\mu)$. A rigorous look at estimation on parametric models is given by Bickel et al. (1993), with particular importance being placed on the Fréchet differentiability of u. The corresponding parametric model is then simply range$(u) \subset \mathbb{P}$. This approach extends naturally to semi-parametric models for which assumptions of finite-dimensional Θ are relaxed.

In this work, we take a reverse approach, specifying a parameter space $\Theta \subset \mathbb{R}^d$ and a surjective (onto) map $v : \mathcal{P} \to \Theta$ first, the latter of which we shall refer to as a (reverse) parametrization. Defining

$$v^{-1}(\boldsymbol{\theta}) = \{P \in \mathcal{P} : v(P) = \boldsymbol{\theta}\},$$

instead of λ on $\mathcal{P}$, we may work with a function L on d-dimensional Euclidean space to derive quasi-divergences on $\mathcal{P}$. Now for an analogous definition of propriety.

Definition 2. We say that $L : \mathcal{X} \times \Theta \to \mathbb{R}$ is Θ-*proper* if

$$\mathbb{E}_Q[\Delta_L(\boldsymbol{\theta}', \boldsymbol{\theta})] \geq 0, \quad \forall\, Q \in v^{-1}(\boldsymbol{\theta}) \tag{2.3}$$

holds for every $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$. As well, if

$$\mathbb{E}_Q[\Delta_L(\boldsymbol{\theta}', \boldsymbol{\theta})] = 0, \forall\, Q \in v^{-1}(\boldsymbol{\theta}) \iff \boldsymbol{\theta} = \boldsymbol{\theta}' \tag{2.4}$$

holds, then L is *strictly* Θ-*proper*.

With these definitions in hand we confirm some basic facts.

Proposition 1. Given a surjective $v : \mathcal{P} \to \Theta$, L on $\mathcal{X} \times \Theta$, and $\lambda(v) \coloneqq \lambda_v(\mathbf{x}, P) \coloneqq L(\mathbf{x}, v(P))$, we have

$$L \text{ is } \Theta\text{-proper} \iff d_{\lambda(v)} \text{ is a quasi-divergence.}$$

Proof. Using surjectivity of v, have $v^{-1}(\boldsymbol{\theta}) \neq \emptyset$ on Θ. Let L be Θ-proper. Then

$$\begin{aligned} 0 &\leq \mathbb{E}_Q[L(\mathbf{x}, v(P)) - L(\mathbf{x}, v(Q))], \quad \forall\, P, Q \in \mathcal{P} \\ &= \mathbb{E}_Q[\lambda_v(\mathbf{x}, P) - \lambda_v(\mathbf{x}, Q)], \end{aligned}$$

since the definition of $\lambda(v)$ implies the latter inequality.

For the converse, first note that

$$\mathbb{E}_R[L(\mathbf{x}, v(R))] = \mathbb{E}_R[L(\mathbf{x}, v(Q))] = \mathbb{E}_R[L(\mathbf{x}, \boldsymbol{\theta})],$$

for all $R \in v^{-1}(\boldsymbol{\theta})$. Choose some $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$. For $P \in v^{-1}(\boldsymbol{\theta}')$ and $Q, R \in v^{-1}(\boldsymbol{\theta})$ then, we have that

$$\begin{aligned} \mathbb{E}_R[\Delta_L(\boldsymbol{\theta}', \boldsymbol{\theta})] &= \mathbb{E}_R[L(\mathbf{x}, v(P))] - \mathbb{E}_R[L(\mathbf{x}, v(Q))] \\ &= \mathbb{E}_R[\lambda_v(\mathbf{x}, P)] - \mathbb{E}_R[\lambda_v(\mathbf{x}, R)] \\ &\geq 0, \end{aligned}$$

where $\mathcal{P}$-propriety implies the final inequality. □

If the map v is bijective, which occurs if and only if $v^{-1}(\boldsymbol{\theta})$ is a singleton set for all $\boldsymbol{\theta} \in \Theta$, we may note the natural equivalence of strict propriety in this special case.

Proposition 2. Let $v : \mathcal{P} \to \Theta$ be bijective, with L and $\lambda(v)$ as in Prop. 1. Then we have

$$L \text{ is strictly } \Theta\text{-proper} \iff d_{\lambda(v)} \text{ is a divergence.} \tag{2.5}$$

Proof. Strict $\mathcal{P}$-propriety of $\lambda(v)$ is implied if $d_{\lambda(v)}$ is a divergence. Note $\boldsymbol{\theta} = \boldsymbol{\theta}'$ trivially implies $\mathbb{E}_Q[\Delta_L(\boldsymbol{\theta}', \boldsymbol{\theta})] = 0$. Conversely, assume $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$. As $v^{-1}(\boldsymbol{\theta}) \cap v^{-1}(\boldsymbol{\theta}') = \emptyset$, we have for $P \in v^{-1}(\boldsymbol{\theta}')$ and $Q \in v^{-1}(\boldsymbol{\theta})$ that $P \neq Q$. Then using $\mathcal{P}$-propriety, we have that

$$0 \neq \mathbb{E}_Q[\Delta_{\lambda(v)}(P, Q)] = \mathbb{E}_Q[\Delta_L(\boldsymbol{\theta}', \boldsymbol{\theta})]$$

which is to say the expected difference does not vanish for some $Q \in v^{-1}(\boldsymbol{\theta})$. The sufficiency result follows by contrapositive.

For the other direction, let strict Θ-propriety of L hold. $P = Q$ implies $\mathbb{E}_Q[\Delta_{\lambda(v)}(P, Q)] = 0$ trivially. Conversely, let $P \neq Q$. Then injectivity of v gives us $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$, and thus $\mathbb{E}_Q[\Delta_{\lambda(v)}(P, Q)] = \mathbb{E}_Q[\Delta_L(\boldsymbol{\theta}', \boldsymbol{\theta})] \neq 0$ which follows by strict Θ-propriety. □

Given a reverse parametrization from $\mathcal{P}$ to $\Theta \subset \mathbb{R}^d$ then, the propriety of a loss function on Θ and $\mathcal{P}$ indeed are equivalent. Should $\mathcal{P}$ and Θ be isomorphic, strict propriety is also equivalent on both sets. In this case, regularity conditions typically applied to parametrizations on Θ in formal asymptotic analyses would be applied to v^{-1}.

2.3 Competitive reference estimators

There are two major components to this section. The first component is a derivation of a closed-form expression for the minimum continuous ranked probability score which may be used in the case of a Weibull model. Strictly speaking, this CRPS derivation and application to estimation of a univariate Weibull model does not appear to be present in the literature and may be considered novel, given works such as that of Friederichs and Thorarinsdottir (2012), the possibility of such an application is almost certainly known. The second component of this section is a brief introduction of key estimators which may be used as competitive references in the performance evaluation tasks described in Chapter 3, compared against the estimators proposed in the following section 2.4. The minimum (Weibull) CRPS estimator discussed here shall also be used as a reference to beat.

A natural benchmark reference is the minimum KL-divergence estimator, or equivalently the minimum negative log-likelihood (NLL) estimator defined

$$\widehat{\boldsymbol{\theta}}_{NLL} \coloneqq \arg\min_{\boldsymbol{\theta}} \sum_{t=1}^{T} (-1) \log p(\mathbf{x}_t; \boldsymbol{\theta}).$$

Indeed, the NLL estimator is ubiquitous in applications of probabilistic wind speed modelling, in both tasks requiring an optimal fit (Seguro and Lambert, 2000; Morgan et al., 2011; Zhang et al., 2013) and tasks focused solely on forecasting (Taylor et al., 2009; Friederichs and Thorarinsdottir, 2012).

Another viable reference method is the continuous ranked probability score (CRPS), defined for CDF P as

$$\text{CRPS}(P(\boldsymbol{\theta}), x) \coloneqq \int_{-\infty}^{\infty} |P(u; \boldsymbol{\theta}) - \mathbf{1}[u \geq x]|^2 du. \tag{2.6}$$

While the term "score" is used, we naturally treat this as a loss function and the corresponding estimator is

$$\widehat{\boldsymbol{\theta}}_{CRP} \coloneqq \arg\min_{\boldsymbol{\theta}} \sum_{t=1}^{T} \text{CRPS}(P(\boldsymbol{\theta}), x_t).$$

The CRPS quantity dates back to work from Brown (1974), and in recent years has seen significant attention in the meteorological community, with detailed investigations carried out by Hersbach (2000) and Gneiting et al. (2005), among others. As noted by Gneiting and Raftery (2007), one may show that

$$\text{CRPS}(P(\boldsymbol{\theta}), x) = \frac{1}{2}\mathbb{E}_P[|X - X'|] - \mathbb{E}_P[|X - x|],$$

where $X \sim P$ and $X' \sim P$ are independent. This result is useful when closed-form expressions for the integral in (2.6) are not available. In the Normal or truncated Normal distribution cases, readily computable functional forms for (2.6) exist (Gneiting et al., 2005). For the more general forecasting task considered in our work, Gaussian assumptions can be difficult to make, and as such we would like to use the CRPS in the case of a model which is more generally applicable. A natural example of such a model is the two-parameter Weibull distribution with scale $\lambda > 0$ and shape $\kappa > 0$, distribution function $W(x; \lambda, \kappa) \coloneqq 1 - \exp(-(x/\lambda)^\kappa)$ and density function

$$w(x; \lambda, \kappa) = \frac{\kappa}{\lambda} \left(\frac{x}{\lambda}\right)^{\kappa - 1} \exp\left(-\left(\frac{x}{\lambda}\right)^{\kappa}\right).$$

The Weibull is a univariate, lower-bounded continuous probability distribution which arises naturally in the theory of extreme value distributions, and has seen extensive use by practitioners for modelling systems as diverse as survival rates, resource allocation, and wind velocity (Rinne, 2010, pp. 4–26). Studies focusing on the utility of the Weibull for wind-related forecasting date back to work from Hennessey (1977), Brown et al. (1984), and (Conradsen et al., 1984), continuing through to more modern works (Morgan et al., 2011; Zhang et al., 2013). It arises directly in what is called the type-III minimum Generalized Extreme Value (GEV) distribution, though here we consider the type-I minimum GEV with real shift μ and scale $\sigma > 0$ parameters defined by distribution function

$$G(x; \mu, \sigma) = 1 - \exp\left(-\exp\left(\frac{x-\mu}{\sigma}\right)\right). \tag{2.7}$$

Note that if $x \sim W(\lambda, \kappa)$ and we define $y = \log(x)$, then since $\log$ is strictly increasing on the positive half-line (the Weibull domain) we can readily confirm that $y \sim G(\mu, \sigma)$ with $\mu = \log(\lambda)$ and $\sigma = 1/\kappa$. The following result gives us a straightforward way to use the CRPS in the case of a Weibull model.

Proposition 3. Given the type-I minimum GEV denoted $G(x; \mu, \sigma)$ as in (2.7), we have that

$$\mathrm{CRPS}(G(\mu, \sigma), x) = x - \mu + \sigma\left(\gamma - \log(2) - 2\,\mathrm{Ei}\left(-\exp\left(\frac{x-\mu}{\sigma}\right)\right)\right) \tag{2.8}$$

where $\mathrm{Ei}(x) = \int_{-\infty}^{x} e^u/u\,du$.

Proof. It is fairly straightforward to find numerous univariate functions which when integrated over their domain are equivalent to the CRPS, and a particularly nice general form is for distribution P is

$$\begin{aligned}\mathrm{CRPS}(P, y) &= \int_0^1 2\left(\mathbf{1}[P^{-1}(\xi) > y] - \xi\right)\left(P^{-1}(\xi) - y\right) d\xi \\ &= 2\int_0^1 P^{-1}(\xi)\mathbf{1}[P^{-1}(\xi) > y]\, d\xi - 2\int_0^1 P^{-1}(\xi)\xi\, d\xi - 2y(1 - P(y)) + y\end{aligned}$$

where P^{-1} denotes the quantile function of P. A straightforward proof of the first identity is given in A.1. For the minimum GEV case, noting that

$$G^{-1}(\xi; \mu, \sigma) = \sigma \log\left(-\log(1-\xi)\right) + \mu$$

we have that doing a simple change of variables we get

$$2\int_0^1 G^{-1}(\xi)\mathbf{1}[G^{-1}(\xi) > y]\,d\xi = 2\sigma\int_0^{1-G(y)} \log(-\log u)\,du + 2\mu(1-G(y)).$$

To deal with this term, we note that the exponential integral (cf. Abramowitz and Stegun, 1964) denoted $\mathrm{Ei}(x)$ and defined

$$\mathrm{Ei}(x) \coloneqq \int_{-\infty}^{x} \frac{e^u}{u}\,du = \gamma + \log|x| + \sum_{n=0}^{\infty}\frac{x^n}{n!n}$$

where

$$\gamma \coloneqq \lim_{N\to\infty}\left(\sum_{n=1}^{N} n^{-1} - \log N\right) = 0.57721566\ldots$$

is the Euler-Mascheroni constant (cf. Havil, 2003), satisfies $\mathrm{Ei}'(x) = e^x/x$, and then we may readily confirm the following indefinite integrals

$$\int \log(-\log u)\,du = u\log(-\log u) - \mathrm{Ei}(\log u) + C$$
$$\int \log(-\log u)u\,du = \frac{1}{2}\left(u^2\log(-\log u) - \mathrm{Ei}(2\log u)\right) + C$$

using the derivative of Ei and the chain rule. This function is obviously continuous on $(0,1)$. To evaluate the integral we first note

$$\lim_{u\downarrow 0}\left(u\log(-\log u) - \mathrm{Ei}(\log u)\right) = 0$$

as $u\log(-\log u) \to 0$ as u approaches 0 from above, and $\mathrm{Ei}(x)$ vanishes as $x \to -\infty$, noting the limits of integration become arbitrarily close. We thus evaluate the first integral as

$$2\int_0^1 G^{-1}(\xi)\mathbf{1}[G^{-1}(\xi) > y]\,d\xi = 2(1-G(y))y - 2\sigma\,\mathrm{Ei}(\log(1-G(y))) \quad (2.9)$$

where we have used

$$\log(-\log(1-G(y))) = \frac{y-\mu}{\sigma}.$$

For the second integral, we note

$$2\int_0^1 P^{-1}(\xi)\xi\, d\xi = 2\sigma \int_0^1 \log(-\log u)(1-u)\, du + \mu.$$

To evaluate this, while the $u \to 0$ case can be handled just as above, for $u \to 1$ we make use of the series expansion of Ei and thus note

$$\lim_{u\uparrow 1}\left(u^2\log(-\log u) - \frac{\gamma}{2} - \frac{\log(-2\log u)}{2} - \sum_{n=0}^{\infty}\frac{(2\log u)^n}{2(n!n)}\right) = -\frac{\gamma+\log 2}{2}$$

$$\lim_{u\uparrow 1}\left(u\log(-\log u) - \gamma - \log(-\log u) - \sum_{n=0}^{\infty}\frac{(\log u)^n}{n!n}\right) = -\gamma$$

since $(u^2-1)\log(-\log u)$, $(u-1)\log(-\log u)$, and the infinite sum vanish as u approaches 1 from below. We thus have that

$$2\int_0^1 P^{-1}(\xi)\xi\, d\xi = \sigma\log 2 - \sigma\gamma + \mu. \tag{2.10}$$

Replacing the two integrals in the CRPS expression with (2.9) and (2.10) above then yields

$$\begin{aligned}\mathrm{CRPS}(G(\mu,\sigma),y) &= y - \mu + \sigma\left(\gamma - \log 2 - 2\,\mathrm{Ei}(\log(1-G(y)))\right)\\ &= y - \mu + \sigma\left(\gamma - \log 2 - 2\,\mathrm{Ei}\left(-\exp\left(\frac{y-\mu}{\sigma}\right)\right)\right)\end{aligned}$$

which is our desired result. □

In the GEV case then, if we assume that GEV parameters μ and σ are functions of $\boldsymbol{\theta}_\mu$ and $\boldsymbol{\theta}_\sigma$ respectively, then denoting $\boldsymbol{\theta} \coloneqq (\boldsymbol{\theta}_\mu, \boldsymbol{\theta}_\sigma)$, we have that the minimum CRPS estimator is

$$\widehat{\boldsymbol{\theta}}_{CRP} = \arg\min_{\boldsymbol{\theta}} \sum_{t=1}^{T}\left(x_t - 2\sigma\,\mathrm{Ei}\left(-\exp\left(\frac{x_t-\mu(\boldsymbol{\theta}_\mu)}{\sigma(\boldsymbol{\theta}_\sigma)}\right)\right)\right) + T(\sigma(\boldsymbol{\theta}_\sigma)(\gamma-\log 2) - \mu(\boldsymbol{\theta}_\mu)). \tag{2.11}$$

We now note a few points with respect to explicit computation of this value. The exponential integral term may be readily computed using routines written in the C programming language by making use of the GNU Scientific Library (Galassi

et al., 2013). As $(x_t - \mu)/\sigma$ grows large, the term approaches 0 quickly, and C may run into underflow issues, and establishing a threshold here is useful (typically not much larger than $(x_t - \mu)/\sigma > 6$). Ei is undefined at zero, and thus some standard systematic method for removing zeroes will be necessary. Speaking from experience, the difference in output between replacing zero-valued $(x_t - \mu)/\sigma$ with some pre-determined value (on the order of $\pm 10^{-4}$) and replacing with randomly generated Normal values (say from $\mathcal{N}(0, 10^{-4})$) is negligible.

For parameter estimation, the first and second-order derivatives may be useful. Obtaining these is straightforward, albeit tedious, and for reference we include the first-order expressions and the diagonal elements of the Hessian as follows. Let shift μ be a function of scalar θ_μ, and scale σ a function of scalar θ_σ. For clean notation we denote $G(x) \coloneqq G(x; \mu, \sigma)$ and $z \coloneqq (x - \mu)/\sigma$. The first-order derivatives of the CRPS in the case of a type-I minimum GEV distribution are

$$\frac{\partial}{\partial \theta_\mu} \operatorname{CRPS}(G, x) = \left(2 \exp\left(-\exp\left(z\right)\right) - 1\right) \frac{\partial \mu}{\partial \theta_\mu},$$
$$\frac{\partial}{\partial \theta_\sigma} \operatorname{CRPS}(G, x) = \left(\gamma - \log(2) - 2\,\mathrm{Ei}(\mathcal{L}(x)) + 2z \exp\left(\mathcal{L}(x)\right)\right) \frac{\partial \sigma}{\partial \theta_\sigma}$$

where $\mathcal{L}(x) \coloneqq \log(1 - G(x))$ and we note

$$\log(1 - G(x)) = -\exp\left(z\right), \text{ and}$$
$$\frac{\exp\left(z - \exp\left(z\right)\right)}{\log\left(1 - G(x)\right)} = -\exp\left(-\exp\left(z\right)\right)$$

being careful with signs. Next, the diagonal terms of the Hessian. First with respect to the shift parameter,

$$\frac{\partial^2}{\partial \theta_\mu^2} \operatorname{CRPS}(G, x) = \frac{\partial^2 \mu}{\partial \theta_\mu^2} \left(2 \exp\left(\mathcal{L}(x)\right) - 1\right) + \left(\frac{\partial \mu}{\partial \theta_\mu}\right)^2 \frac{2}{\sigma} \exp\left(z + \mathcal{L}(x)\right).$$

Next, defining

$$A \coloneqq \gamma - \log(2) - 2\,\mathrm{Ei}(\mathcal{L}(x)) + 2z \exp\left(\mathcal{L}(x)\right)$$

we have

$$\frac{\partial^2}{\partial \theta_\sigma^2} \operatorname{CRPS}(G, x) = \frac{\partial^2 \sigma}{\partial \theta_\sigma^2} A + \left(\frac{\partial \sigma}{\partial \theta_\sigma}\right)^2 2z \exp\left(\mathcal{L}(x)\right) \left(1 - \frac{1}{\sigma} - \frac{x - \mu}{\sigma^2} \mathcal{L}(x)\right).$$

Finally, we note that the minimum CRPS estimator has been reported to be a successful and robust alternative to NLL in the forecasting literature in recent years (Gneiting et al., 2006; Lerch and Thorarinsdottir, 2013; Holland and Ikeda, 2014), and we thus use it as a competitive reference method to beat.

Next we consider additional standard references in addition to NLL and CRPS. The classical deterministic reference is the random walk estimator, also known as persistence (PER) in the meteorology literature, defined $\widehat{x}_{t+k} = x_t$, and is traditionally extremely competitive for short-term forecasts. Another very strong deterministic reference comes from Nielsen et al. (1998), here denoted NIEL, and is a simple autocorrelation-weighted moving average model. The typical form considered here is $\widehat{x}_{t+k} = \rho_k x_t + (1-\rho)\bar{x}$, where ρ_k is the k-lagged autocorrelation coefficient, and $\bar{x}$ is the arithmetic mean of recent observations.

For a classical probabilistic reference dating back at least as far as (Justus et al., 1976), taking two consecutive logarithms of the Weibull CDF $W(x; \lambda, \kappa)$ yields

$$\log(-\log(1-\pi_t)) = \kappa_t \log x_t + \kappa_t \log \lambda_t(\boldsymbol{\theta}_\lambda),$$

where π_t denotes the empirical cumulative probability assigned to observation x_t. Let us model $\kappa = w_0$ as a scalar and $\lambda = \exp(\mathbf{w}_v^T \boldsymbol{\phi}_t)$, where $\boldsymbol{\phi}_t = (\phi_1(t), \ldots, \phi_v(t))$ is a given vector of relevant features/covariates. This allows us to utilize a linear model $\mathbf{c} = \boldsymbol{\Phi}\mathbf{w}$, where $\mathbf{c} = (\log(-\log(1-\pi_1)), \ldots, \log(-\log(1-\pi_T)))$, $\boldsymbol{\Phi}$ is a $T \times (v+1)$ matrix composed of $\log x_t$ values in the first column, and $\boldsymbol{\phi}_t$ values in the remaining columns, and $\mathbf{w} = (w_0, \mathbf{w}_v)$. If we define

$$\widehat{\mathbf{w}}_m \coloneqq (\widehat{w}_{0,m}, \widehat{\mathbf{w}}_{v,m}) \coloneqq \arg\min_{\mathbf{w}} ||\mathbf{c} - \boldsymbol{\Phi}\mathbf{w}||_m$$

where $||\cdot||_m$ denotes the usual l_m norm here on finite dimensional Euclidean space, then we define the L1CDF and L2CDF estimators by

$$\widehat{\boldsymbol{\theta}}_m = \left(\frac{\widehat{\mathbf{w}}_{v,m}}{\widehat{w}_{0,m}}, \widehat{w}_{0,m} \right)$$

for $m = 1, 2$ respectively. Naturally $m = 2$ is standard OLS estimation, while $m = 1$ is minimum mean absolute deviation estimation, and efficient standard algorithms for this computation are due to Barrodale and Roberts (1973), and we have used the R package `robust`.

2.4 Deriving quasi-divergence minimizing estimators

Here we put forward a methodology for constructing classes of proper loss functions, and thus deriving quasi-divergence minimizing estimators. Given a parametric model $(\mathcal{P}, \Theta, v)$, we construct $\mathcal{L}$ on d-dimensional Euclidean space, show (strict) Θ-propriety, and using v then show (strict) $\mathcal{P}$-propriety. Since we can make liberal use of basic properties of convex functions on $\mathbb{R}^d$, this approach is particularly fruitful, in terms of both theoretical justification and computational tractability.

Let f be a real concave function on $S \subset \mathbb{R}^k$. We call $\mathbf{x}^* \in S$ a *supergradient* of f if $-\mathbf{x}^*$ is a subgradient of convex $-f$ (Rockafellar, 1970). We thus have the useful inequality

$$f(\mathbf{z}) \leq f(\mathbf{x}) + \langle \mathbf{x}^*, \mathbf{z} - \mathbf{x} \rangle, \quad \forall\, \mathbf{z} \in S \tag{2.12}$$

which we shall make use of repeatedly. Letting $\partial f(\mathbf{x})$ denote the set of all supergradients of f at $\mathbf{x}$, if $\partial f(\mathbf{x}) \neq \emptyset$, then from the existence of a supergradient $\mathbf{x}^* \in \partial f(\mathbf{x})$ it follows that there exists a hyperplane H in $\mathbb{R}^{d+1}$ defined

$$H = \{(\mathbf{z}, z_0) : (\mathbf{z}, z_0)^T (\mathbf{x}^*, -1) = \langle \mathbf{x}^*, \mathbf{x} \rangle - f(\mathbf{x})\},$$

as f is real. The epigraph for concave f will be $\text{epi}(f) = \{(\mathbf{x}, \lambda) \in S \times \mathbb{R} : \lambda \leq f(\mathbf{x})\}$, a convex set, and note that for arbitrary $(\mathbf{z}, \lambda) \in \text{epi}(f)$,

$$\begin{aligned}(\mathbf{z}, \lambda)^T (\mathbf{x}^*, -1) &\geq \mathbf{z}^T \mathbf{x}^* + f(\mathbf{z}) \\ &\geq \langle \mathbf{x}^*, \mathbf{x} \rangle - f(\mathbf{x}).\end{aligned}$$

This implies $\text{epi}(f)$ is a subset of the upper half-space corresponding to H. Since $(\mathbf{x}, f(\mathbf{x})) \in \text{epi}(f) \cap H$ is clear, the intersection of $\text{epi}(f)$ and the boundary of the half-space containing it is non-empty. Thus H is a supporting hyperplane to $\text{epi}(f)$, a geometrically appealing fact. By Rockafellar (1970), if we restrict ourselves to $\mathbf{x} \in \text{int}(\text{dom}(f))$ only, then $\partial f(\mathbf{x}) \neq \emptyset$ always holds (Thm. 23.4). As well, the uniqueness of the supergradient, i.e., $\partial f(\mathbf{x}) = \{\mathbf{x}^*\}$ holds if and only if $\mathbf{x}^* = \nabla f(\mathbf{x})$ (Thm. 25.1), the gradient of f at $\mathbf{x}$.

With these basic facts in hand, we shall methodically construct loss functions taking a general form, proving $\mathcal{P}$-propriety explicitly using (2.12) to show Θ-propriety, thus deriving at least quasi-divergences. To motivate the general form proposed here, we consider an example from Grünwald and Dawid (2004) in the case of discrete random variables. Let $\{1, 2, \ldots, M\}$ be our sample space. Thus

probability mass functions are specified by $\mathbf{p} = (p_1, \ldots, p_M) \in \mathcal{S}_M$, the probability simplex on $\mathbb{R}^M$. Let $\mathbf{e}(m) \coloneqq (\mathbf{1}[m = 1], \ldots, \mathbf{1}[m = M])$. If we define

$$G(m, \mathbf{p}) \coloneqq f(\mathbf{p}) + \langle \mathbf{p}^*, \mathbf{e}(m) - \mathbf{q} \rangle$$

where f is any concave function on $\text{int}(\mathcal{S}_M)$, and $\mathbf{p}^* \in \partial f(\mathbf{p})$. Since $\mathbb{E}_{\mathbf{q}}[\mathbf{e}] = \mathbf{q}$, we can observe

$$\mathbb{E}_{\mathbf{q}}[\Delta_G(\mathbf{p}, \mathbf{q})] = f(\mathbf{p}) - f(\mathbf{q}) + \langle \mathbf{p}^*, \mathbf{q} - \mathbf{p} \rangle \geq 0,$$

for all $\mathbf{p}, \mathbf{q} \in \text{int}(\mathcal{S}_M)$, which follows from inequality (2.12). Should some additional regularity conditions on f be applied, including differentiability on $\text{int}(\mathcal{S}_M)$, then

$$\mathbb{E}_{\mathbf{q}}[\Delta_G(\mathbf{p}, \mathbf{q})] = d_B(\mathbf{q}, \mathbf{p})$$

where d_B denotes the Bregman divergence with respect to convex $-f$ on $\mathcal{S}_M$. Note that without loss of generality we may consider $\mathcal{S}_M$ to be a subset of $\mathbb{R}^{M-1}$ denoted $\mathcal{S}_M = \{\mathbf{x} \in \mathbb{R}^{M-1} : \sum_{i=1}^{M-1} x_i \leq 1, x_i \geq 0\}$, and using the usual norm, clearly $\text{int}(\mathcal{S}_M)$ is all $\mathbf{x} \in \mathcal{S}_M$ such that $\sum_{i=1}^{M-1} x_i < 1$ and $x_i > 0, i = 1, \ldots, M$. This says that WLOG the Mth coordinate of any $\mathbf{p} \in \text{int}(\mathcal{S}_M)$ is in $(0, 1)$. Thus, since the map from the set of all discrete probability distributions assigning nonzero probability to all events in a finite sample space of size M is clearly bijective, we have $\text{int}(\mathcal{S}_M)$-propriety of G.

We now return to the general probability space $(\mathcal{X}, \mathcal{A}, P)$. In the following statement, we give a natural form for a loss function L to take, noted to be a straightforward generalization of the example above.

Proposition 4. Let $(\mathcal{P}, \Theta, v)$ be a parametric model where Θ is an open subset of $\mathbb{R}^d$. Define loss function $L : \mathcal{X} \times \Theta \to \mathbb{R}$ to be

$$L(\mathbf{x}, \boldsymbol{\theta}) = f(\boldsymbol{\theta}) + \langle \boldsymbol{\theta}^*, h_v(\mathbf{x}, \boldsymbol{\theta}) - \boldsymbol{\theta} \rangle \tag{2.13}$$

with f concave on Θ and $h_v : \mathcal{X} \times \Theta \to \Theta$. If there exists an operator $\boldsymbol{\theta} \mapsto r(\boldsymbol{\theta})$ returning values in Θ such that

$$\begin{aligned} &\text{(i)} \quad \mathbb{E}_Q[h_v(\mathbf{x}, \boldsymbol{\theta})] = v(Q) + r(\boldsymbol{\theta}) \\ &\text{(ii)} \quad \langle \boldsymbol{\theta}^*, r(\boldsymbol{\theta}) \rangle \geq 0, \quad \forall\, \boldsymbol{\theta} \in \Theta \end{aligned}$$

then $d_{\lambda(v)}$ is a quasi-divergence.

Proof. The existence of supergradient $\boldsymbol{\theta}^*$ at $\boldsymbol{\theta} \in \Theta$ follows as $\text{int}(\text{dom}(f)) = \text{dom}(f)$ is assumed open (cf. A.2). As $\mathbb{E}_Q[L(\mathbf{x}, \boldsymbol{\theta})] = f(\boldsymbol{\theta}) + \langle \boldsymbol{\theta}^*, r(\boldsymbol{\theta}) \rangle$ whenever $Q \in v^{-1}(\boldsymbol{\theta})$, letting

$$\rho \coloneqq \langle \boldsymbol{\theta}'^*, r(\boldsymbol{\theta}') \rangle + \langle \boldsymbol{\theta}^*, r(\boldsymbol{\theta}) \rangle \geq 0$$

for arbitrary $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$, then for any such Q we have

$$\mathbb{E}_Q[\Delta_L(\boldsymbol{\theta}', \boldsymbol{\theta})] = f(\boldsymbol{\theta}') - f(\boldsymbol{\theta}) + \langle \boldsymbol{\theta}'^*, \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle + \rho \geq 0$$

since $v(Q) = \boldsymbol{\theta}$, using (2.12) for the last inequality. □

With $(\mathcal{P}, \Theta, v)$ then, we need to select a $h_v(\mathbf{x}, \boldsymbol{\theta})$ such that L is Θ-proper, with the only constraint being the concavity of f. We note that $\mathbb{S}^d$ is a subset of $\mathbb{R}^{d \times d}$, and is isomorphic to $\mathbb{R}^{d(d+1)/2}$ since every vector $\mathbf{a} \in \mathbb{R}^{d(d+1)/2}$ specifies one and only one $\mathbf{A} \in \mathbb{S}^d$ given a rule mapping long vectors to matrices. The Frobenius norm on $\mathbb{S}^d$ is defined $|\mathbf{A}| \coloneqq \sqrt{\text{tr}\, \mathbf{A}^T \mathbf{A}}$, and for some $\mathbf{A}, \mathbf{B} \in \mathbb{S}^d$ we see

$$|\mathbf{A} - \mathbf{B}|^2 = \sum_{i=1}^{d} (\mathbf{a}_i - \mathbf{b}_i)^T (\mathbf{a}_i - \mathbf{b}_i) = |\mathbf{a} - \mathbf{b}|^2$$

where $\mathbf{a}_j$ denotes the jth column vector of $\mathbf{A}$, and $\mathbf{a}$ is the vector of length $d(d+1)/2$ specifying $\mathbf{A}$. Namely, the Frobenius norm for $d \times d$ matrices coincides with the Euclidean norm on $\mathbb{R}^{d \times d}$. We define reverse parametrizations

$$\begin{aligned} v_1(P) &\coloneqq \mathbb{E}_P[\mathbf{x}] \\ v_2(P) &\coloneqq \mathbb{E}_P[\mathbf{x}\mathbf{x}^T] \\ v_3(P) &\coloneqq (v_1(P), v_2(P) - v_1(P) v_1(P)^T) \end{aligned}$$

and in the following statement, we show that for many parametric models, there exist very natural rudimentary forms for h to take in which the desired propriety may be readily proven.

Proposition 5. Let $\Theta_1 \subset \mathbb{R}^d$ and $\Theta_2 \subset \mathbb{S}^d$ be open subsets of their respective parents. Let $\Theta_3 = \mathbb{R}^d \times \mathbb{S}^d_+$. If real loss functions L_i defined on $\mathcal{X} \times \Theta_i$ for $i = 1, 2, 3$ are of the form

$$\begin{aligned} L_1(\mathbf{x}, \boldsymbol{\mu}) &= f_1(\boldsymbol{\mu}) + \langle \boldsymbol{\mu}^*, \mathbf{x} - \boldsymbol{\mu} \rangle \\ L_2(\mathbf{x}, \mathbf{W}) &= f_2(\mathbf{W}) + \langle \mathbf{W}^*, \mathbf{x}\mathbf{x}^T - \mathbf{W} \rangle \\ L_3(\mathbf{x}, (\boldsymbol{\mu}, \mathbf{W})) &= f_3(\mathbf{W}) + \langle \mathbf{W}^*, (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T - \mathbf{W} \rangle \end{aligned}$$

then we have that

$$d_{\lambda(v_i)} \text{ is a quasi-divergence on } \mathrm{dom}(v_i),$$

respectively on $\mathcal{P}^1(\Theta_1)$, $\mathcal{P}^2(\Theta_2)$, and $\mathcal{P}(\mathbb{R}^d, \mathbb{S}_+^d)$.

Proof. For $\lambda(v_1)$ and $\lambda(v_2)$, the results are immediate by Prop. 4 since $\mathbb{E}_Q[\mathbf{x}] = v_1(Q)$ and $\mathbb{E}_Q[\mathbf{x}\mathbf{x}^T] = v_2(Q)$, which again follows from openness.

For $\lambda(v_3)$, note that $\mathbb{S}_+^d$ is an open subset of $\mathbb{R}^{d(d+1)/2}$ (cf. A.2). Then as $\langle \mathbf{A}, \mathbf{x}\mathbf{x}^T \rangle = \operatorname{tr} \mathbf{A}\mathbf{x}\mathbf{x}^T = \mathbf{x}^T\mathbf{A}\mathbf{x}$ for $\mathbf{A} \in \mathbb{S}^d$ we may note

$$\begin{aligned}\mathbb{E}_Q[\langle \mathbf{W}^*, (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T \rangle] &= \mathbb{E}_Q[(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{W}^* (\mathbf{x}-\boldsymbol{\mu})] \\ &= \langle \mathbf{W}^*, \mathbb{E}_Q[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T] \rangle\end{aligned}$$

and some straightforward algebra results in

$$\mathbb{E}_Q[L_3(\mathbf{x}, (\boldsymbol{\mu}, \mathbf{W}))] = f_3(\mathbf{W}) + \langle \mathbf{W}^*, \mathbf{V}_Q + (\boldsymbol{\mu} - \boldsymbol{\mu}_Q)(\boldsymbol{\mu} - \boldsymbol{\mu}_Q)^T - \mathbf{W} \rangle.$$

Since $\mathbf{W}^* \in \mathbb{S}_+^d$, we see

$$\langle \mathbf{W}^*, (\boldsymbol{\mu} - \boldsymbol{\mu}_Q)(\boldsymbol{\mu} - \boldsymbol{\mu}_Q)^T \rangle = (\boldsymbol{\mu} - \boldsymbol{\mu}_Q)^T \mathbf{W}^* (\boldsymbol{\mu} - \boldsymbol{\mu}_Q) \geq 0.$$

As $v_3(Q) = (\boldsymbol{\mu}_Q, \mathbf{V}_Q)$ and $\mathrm{dom}(f_3) = \mathbb{S}_+^d$, we may apply the proof of Prop. 4 here in a similar way, giving us that$d_{\lambda(v_3)}$ is a quasi-divergence. □

Many important cases of parametric models have parameter space of $\mathbb{R}^d \times \mathbb{S}_+^d$, and thus results specifically for L_3 may be considered particularly pertinent. If we strengthen the conditions imposed on f, a valid divergence may be obtained.

Proposition 6. Let $\mathcal{P} \subset \mathcal{P}(\mathbb{R}^d, \mathbb{S}_+^d)$, for which a bijective $w : \mathcal{P} \to \Theta_3$ exists. Taking L_3 as in Prop. 5, we have that

$$f_3 \text{ is strictly concave } \implies d_{\lambda(w)} \text{ is a divergence.}$$

Proof. Let f_3 be strictly concave. Given the form of $\mathbb{E}_Q[\Delta_{L_3}(\mathbf{W}, \mathbf{W}_0)]$, it remains only to verify that $f_3(\mathbf{W}) - f_3(\mathbf{W}_0) + \langle \mathbf{W}^*, \mathbf{W}_0 - \mathbf{W} \rangle = 0$ if and only if $\mathbf{W} = \mathbf{W}_0$. For a general f strictly concave on $\mathbb{R}^d$ for integer $d \geq 1$, assume

$$f(\mathbf{a}) - f(\mathbf{b}) + \langle \mathbf{a}^*, \mathbf{b} - \mathbf{a} \rangle = 0 \tag{2.14}$$

for $\mathbf{a} \neq \mathbf{b}$. Then by strict concavity of f, for $\lambda \in (0,1)$ we have

$$\begin{aligned} f(\lambda\mathbf{a} + (1-\lambda)\mathbf{b}) &> \lambda f(\mathbf{a}) + (1-\lambda) f(\mathbf{b}) \\ &= \lambda f(\mathbf{a}) + (1-\lambda)(f(\mathbf{a}) + \langle \mathbf{a}^*, \mathbf{b} - \mathbf{a} \rangle) \\ &= f(\mathbf{a}) + \langle \mathbf{a}^*, (\lambda\mathbf{a} + (1-\lambda)\mathbf{b}) - \mathbf{a} \rangle \\ &\geq f(\lambda\mathbf{a} + (1-\lambda)\mathbf{b}) \end{aligned}$$

where (2.12) gives us the final inequality. This is a contradiction, and thus $\mathbf{a} = \mathbf{b}$. That the equality in (2.14) holds then implies $\mathbf{W} = \mathbf{W}_0$. Conversely, assuming $\mathbf{W} = \mathbf{W}_0$, the result is trivial. L_3 is thus strictly Θ_3-proper. Using the bijectivity of w and Prop. 2, we have that $d_{\lambda(w)}$ is a valid divergence. □

We have looked at properties for loss functions constructed using rudimentary h. Next we see that it is not difficult to derive more sophisticated loss functions retaining the desirable properties found in the simpler case.

Proposition 7. Let l be a non-decreasing, concave real function on $(0, \infty)$. Then $f : \mathbb{S}_+^d \to \mathbb{R}$ defined $f(\mathbf{W}) \coloneqq l\left((\det \mathbf{W})^{1/d}\right)$ is concave on its domain.

Proof. Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ for integer $d > 1$, $\lambda \in [0,1]$, and let g be concave on $\mathbb{R}^d$. By our hypotheses then,

$$\begin{aligned} f(\lambda\mathbf{a} + (1-\lambda)\mathbf{b}) &= l(g(\lambda\mathbf{a} + (1-\lambda)\mathbf{b})) \\ &\geq l(\lambda g(\mathbf{a}) + (1-\lambda) g(\mathbf{b})) \\ &\geq \lambda l(g(\mathbf{a})) + (1-\lambda) l(g(\mathbf{b})) \end{aligned}$$

which is to say that $f = l \circ g$ is concave.

We next note that for any $\mathbf{A}, \mathbf{B} \in \mathbb{S}_+^d$,

$$\begin{aligned} (\det(\lambda\mathbf{A} + (1-\lambda)\mathbf{B}))^{1/d} &\geq (\det \lambda\mathbf{A})^{1/d} + (\det(1-\lambda)\mathbf{B})^{1/d} \\ &= \lambda(\det \mathbf{A})^{1/d} + (1-\lambda)(\det \mathbf{B})^{1/d} \end{aligned}$$

holds. The first inequality is Minkowski's determinant theorem, and this result gives us concavity of $(\det \mathbf{W})^{1/d}$ and concludes the proof. □

Given the statement of Prop. 7, $l(x) = \log(x)$ and $l(x) = x^{1/\alpha}$ for real $\alpha \geq 1$ are two very natural selections we might make for compositional loss functions.

Both are clearly concave and non-decreasing on $(0, \infty)$. First some basic facts. On $\mathbb{S}_+^d$, we have that

$$\partial_{i,j} \det \mathbf{W} \coloneqq \partial \det \mathbf{W} / \partial w_{i,j} = (\det \mathbf{W}) \operatorname{tr} \mathbf{W}^{-1} \partial \mathbf{W} / \partial w_{i,j},$$

and the matrix which has the partial derivative $\partial_{i,j} \det \mathbf{W}$ as its (i,j)th element is readily confirmed to be

$$\partial \det \mathbf{W} / \partial \mathbf{W} \coloneqq [\partial_{i,j} \det \mathbf{W}] = \det(\mathbf{W}) \mathbf{W}^{-1}.$$

Then letting L_α and L_{ld} denote L_3 with f_3 set to $f_3 = (\det(\cdot))^{1/\alpha}$ and $f_3 = \log\det$ respectively, we have

$$L_\alpha(\mathbf{x}, (\boldsymbol{\mu}, \mathbf{W})) = (\det \mathbf{W})^{1/\alpha} \left(1 + \frac{(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{W}^{-1} (\mathbf{x} - \boldsymbol{\mu})}{\alpha}\right) \tag{2.15}$$

$$L_{ld}(\mathbf{x}, (\boldsymbol{\mu}, \mathbf{W})) = \log\det \mathbf{W} + (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{W}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \tag{2.16}$$

for $\alpha \geq d$, noting we have added d to the original derivations to cancel out an irrelevant constant term. As noted in 2.4, uniqueness of the supergradient follows from differentiability of the corresponding f_3.

Using the concavity of $\log$ and $x^{1/\alpha}$ on the positive half-line lets us apply Prop. 5 to (2.15) and (2.16), giving us by Prop. 7 that

- Both L_{ld} and L_α are $\mathcal{P}(\mathbb{R}^d, \mathbb{S}_+^d)$-proper.
- $\mathbb{E}_Q[\Delta_{L_{ld}}(v_3(P), v_3(Q))]$ and $\mathbb{E}_Q[\Delta_{L_\alpha}(v_3(P), v_3(Q))]$ are quasi-divergences.

It is also straightforward to verify that these results hold strictly for L_{ld}, as we see below.

Proposition 8. Let $\mathcal{P} \subset \mathcal{P}(\mathbb{R}^d, \mathbb{S}_+^d)$ be any model isomorphic to Θ_3 with mapping w. Then L_{ld} is strictly Θ_3-proper and thus $d_{\lambda(w)}$ is a divergence.

Proof. Given the result of Prop. 6, we need only prove $\log\det$ is strictly concave on $\mathbb{S}_+^d$. Fix a $\mathbf{W} \in \mathbb{S}_+^d$, and let $\mathbf{Y}$ be a symmetric matrix with elements $|y_{i,j}| \in [-1, 1]$. Let $g(u; \mathbf{Y}) \coloneqq \log\det(\mathbf{W} + u\mathbf{Y})$. Since $\mathbb{S}_+^d$ is an open subset of $\mathbb{R}^{d(d+1)/2}$, there exists $\varepsilon > 0$ such that $B_\varepsilon(\mathbf{W}) \subset \mathbb{S}_+^d$. If $u \in (-\varepsilon/d, \varepsilon/d)$, then clearly $\mathbf{W} + u\mathbf{Y} \in B_\varepsilon(\mathbf{W})$ and is thus positive definite. Setting $\operatorname{dom}(g) = (-\varepsilon/d, \varepsilon/d)$, it follows that $g(u; \mathbf{Y})$ is differentiable on $\operatorname{dom}(g)$.

Let $\sqrt{\mathbf{W}} = \mathbf{S}\sqrt{\mathbf{\Lambda}}\mathbf{S}^T$, where $\mathbf{W} = \mathbf{S}\mathbf{\Lambda}\mathbf{S}^T$ is the spectral decomposition of $\mathbf{W}$ with $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$. It is thus immediate that $\sqrt{\mathbf{W}} \in \mathbb{S}_+^d$ as eigenvalues

are the $\sqrt{\lambda_i} > 0$. Defining $\mathbf{\Gamma} \coloneqq \sqrt{\mathbf{W}}^{-1}\mathbf{Y}\sqrt{\mathbf{W}}^{-1} \in \mathbb{S}^d$, with eigenvalues γ_i, we have

$$\begin{aligned} g(u;\mathbf{Y}) &= \log\det\sqrt{\mathbf{W}}(\mathbf{I}+u\mathbf{\Gamma})\sqrt{\mathbf{W}} \\ &= \log\det\mathbf{W} + \sum_{i=1}^{d}\log(1+u\gamma_i). \end{aligned}$$

The second-order derivative is $\partial^2 g(u;\mathbf{Y}) = -\sum\gamma_i^2/(1+u\gamma_i)^2 < 0$, and thus $g(u;\mathbf{Y})$ is strictly concave on its domain (Thm. 4.4) (Rockafellar, 1970). Denoting $B_u \coloneqq \mathbf{W} + u\mathbf{Y}$, by definition of g we see that

$$\log\det(\lambda\mathbf{B}_t + (1-\lambda)\mathbf{B}_u) > \lambda\log\det\mathbf{B}_t + (1-\lambda)\log\det\mathbf{B}_u \tag{2.17}$$

for any $t \neq u$ on dom(g), and any $\mathbf{Y}$ as described. We now let $\zeta \coloneqq \varepsilon/2d$, so $[-\zeta,\zeta] \subset$ dom(g), and clearly for any $\mathbf{B} \in B_\zeta(\mathbf{W})$ there exists $u \in$ dom(g) and $\mathbf{Y}$ such that $\mathbf{B} = \mathbf{W} + u\mathbf{Y}$. The inequality (2.17) implies that $\log\det$ is concave over $B_\zeta(\mathbf{W}) \subset B_\varepsilon(\mathbf{W})$, and as $\mathbf{W}$ was arbitrary, the proof is finished. □

With a parametric model $(\mathcal{P}, \Theta, v)$ of interest, we consider the estimators

$$P_N^\alpha \coloneqq \operatorname*{arg\,min}_{P\in\mathcal{P}} \sum_{i=1}^{N} \Delta_{L_\alpha}(v(P), v(Q)) \tag{2.18}$$

$$P_N^{ld} \coloneqq \operatorname*{arg\,min}_{P\in\mathcal{P}} \sum_{i=1}^{N} \Delta_{L_{ld}}(v(P), v(Q)). \tag{2.19}$$

Without regularity assumptions, we informally have $P_N^\alpha \approx \arg\min_P d_\alpha(P,Q)$ and $P_N^{ld} \approx \arg\min_P d_{ld}(P,Q)$, with quasi-divergence

$$d_\alpha(P,Q) \coloneqq \mathbb{E}_Q[\Delta_{L_\alpha}(v(P), v(Q))]$$

and divergence

$$d_{ld}(P,Q) \coloneqq \mathbb{E}_Q[\Delta_{L_{ld}}(v(P), v(Q))].$$

In Chapter 3, in the context of short-term wind speed forecasting we empirically evaluate the utility of these estimators compared with the references put forward in 2.3, looking at both "accuracy" in the sense of high-quality predictive distributions, as well as robustness to model error that results from changes in seasonal, diurnal, and spatial factors.

2.5 Application to Weibull-based models

The properties of the loss functions minimized in estimators (2.18) and (2.19) as shown in the previous subsection hold for arbitrary dimension $d \geq 1$. As only the first and second moments are required, the proposed metric(s) are valid for a large number of distributions, including of course the Normal and Truncated Normal, Gamma, Rayleigh, log-Normal, generalized extreme value, and Weibull distributions, among many others. As an interesting example which also is particularly relevant to the research problem at hand, in this section we note the basic identities related to L_α and L_{ld} in the univariate Weibull case, and also discuss several parameter initialization methods.

Let W_N^α denote the estimator defined using loss function L_α in (2.15), where the model of interest is

$$\mathcal{P}_W = \{P : dP/d\mu = p(x) = w(x; \lambda, \kappa), \lambda > 0, \kappa > 0\}$$

and the reverse parametrization here is naturally $v(P) \coloneqq (\lambda, \sigma) \in \mathbb{R}^2$, and $\Theta = \mathbb{R}^2_{++}$. In this case we have

$$W_N^\alpha = \arg\min_{\lambda,\kappa} \sum_{i=1}^{N} \left(\lambda^2(\Gamma_2 - \Gamma_1^2)\right)^{1/\alpha} \left(1 + \frac{(x_i - \lambda\Gamma_1)^2}{\alpha\lambda^2(\Gamma_2 - \Gamma_1^2)}\right)$$

where $\Gamma_n \coloneqq \Gamma(1 + n/\kappa)$ and $\Gamma(z)$ denotes the usual Gamma function (Artin, 1964). If we model λ and κ such that they are respectively dependent on scalars u and v, the gradient of the loss function being summed above, here denoted simply $\mathcal{L}$, may be readily acquired as

$$\frac{\partial\mathcal{L}}{\partial u} = \left(\frac{\partial\lambda}{\partial u}\right) \frac{2(\lambda^{2-\alpha}\mathcal{G})^{1/\alpha}}{\alpha} \left(1 - \frac{1}{\alpha} - \frac{1}{\lambda\mathcal{G}} \left((x - \lambda\Gamma_1)\Gamma_1 + \left(1 - \frac{1}{\alpha}\right) \frac{(x - \lambda\Gamma_1)^2}{\lambda}\right)\right)$$

$$\frac{\partial\mathcal{L}}{\partial v} = \left(\frac{\partial\kappa}{\partial v}\right) \frac{2}{\alpha\kappa^2} \left(\frac{\lambda^2}{\mathcal{G}^{\alpha-1}}\right)^{1/\alpha} \left(\left(1 - \frac{1}{\alpha}\right)(\Gamma_2' - \Gamma_1\Gamma_1') \left(\frac{(x - \lambda\Gamma_1)^2}{\lambda^2\mathcal{G}} - 1\right) + \frac{(x - \lambda\Gamma_1)\Gamma_1'}{\lambda}\right)$$

where for notational clarity we denote $\mathcal{G} \coloneqq (\Gamma_2 - \Gamma_1^2)$ and $\Gamma_n' = \Gamma'(1 + n/\kappa)$. Optimization algorithms utilizing gradients, such as the quasi-Newton BFGS method as implemented in the R language and environment (R Core Team, 2014) may thus readily be made use of here, and where required, linear inequality constraints can be implemented via logarithmic barrier functions. Analogous expressions for

W_N^{ld} can be readily obtained in a similar fashion and take a similar form, though are tangential to our discussion here.

Next, we give a very brief overview of the Weibull parameter initialization methods used in the experiments in Chapter 3. The following methods were employed:

- MOM initializer from Newby (1980)
- Maximum likelihood initializer
- Past-value initializer
- Random initializer
- Fixed initializer

The first method is straightforward and moment-based, proposed originally by Newby (1980). Noting that the mean and variance of $x \sim W(\lambda, \kappa)$ are respectively given by

$$\mathbb{E}_W[x] = \lambda\Gamma(1 + 1/\kappa), \qquad \mathbb{V}_W[x] = \lambda^2 \left(\Gamma(1 + 2/\kappa) - \Gamma^2(1 + 1/\kappa)\right),$$

this suggests naturally that a moment-based approach using the "coefficient of variation" $\sqrt{\mathbb{V}_W[x]}/\mathbb{E}_W[x]$ may be effective since dependence is only on κ. If we let m_1 and m_2 denote the first and second sample moments, letting

$$g(\kappa) = \frac{\sqrt{\Gamma(1 + 2/\kappa) - \Gamma^2(1 + 1/\kappa)}}{\Gamma(1 + 1/\kappa} - \frac{\sqrt{m_2 - m_1^2}}{m_1}$$

we have

$$g'(\kappa) = \frac{1}{\kappa^2\Gamma(1 + 1/\kappa)} \left(\psi(1 + 1/\kappa)(\Gamma^2(1 + 1/\kappa) + 1) - \psi(1 + 2/\kappa)\Gamma(1 + 2\kappa)\right),$$

where ψ denotes the digamma function $\psi(z) = \Gamma'(z)/\Gamma(z)$. Given any reasonable small positive initial values for κ, one may use the standard Newton-Raphson method to iteratively seek the root of g, denoted $\widehat{\kappa}$, and then simply set $\widehat{\lambda} = m_1/\Gamma(1 + 1/\widehat{\kappa})$. In the experiments used in Chapter 3, we consider Weibull models where the scale λ is given as a linear combination of relevant features, including a constant term (i.e., the "intercept" term). In such a multi-parameter case, we let the intercept term weight(s) be initialized using the Newby method,

with the remaining terms set to some small positive constant on the order of 10^{-1}. We also considered method of moments approaches proposed by Mihram (1977) and Blischke and Scheuer (1986), which are very similar, and yield results in controlled tests which are so close to results of the Newby approach that the differences are almost negligible. Strictly speaking the Newby method did perform better in these tests, and thus we elect to use it here.

The second parameter initializer is a maximum likelihood method. One simple method is to estimate scalar κ using Newton-Raphson and then the intercept weight in a linear model of λ once $\widehat{\kappa}$ is obtained, and the non-intercept weights to sufficiently small positive values. More sophisticated optimization methods may also naturally be used. The third initializer fixes the intercept terms to 1 for the first time step $t = 1$ at which a forecast is carried out, and then after all parameter estimation is finished (for a given model and estimation method), the resulting parameters are used as initial values for the following time step, and so forth. The fourth initializer simply sets the intercept weights to $1 + |\mathcal{N}(0, 0.5)|$, generated anew for each time step, and for each model. The final initializer simply sets the intercept weights to 1, for all time steps.

Chapter 3

Evaluation of performance

3.1 Experiment details

To carry out a sufficiently in-depth empirical investigation of the utility of the proposed methods, we consider two large data sets which let us consider the task of density estimation and forecasting of wind velocity.

AMeDAS network data The first data set is consists of weather data from across several islands of Japan, collected by a network of observation sites called the *Automated Meteorological Data Acquisition System*, or AMeDAS, operated by the Japanese Meteorological Agency (JMA) (Japan Meteorological Agency, 2013). In addition to hourly, monthly, and seasonal measurements since November 1974, 10-minute measurements from April 1994 to the present are made public over the internet by the JMA. We have obtained a large subset of the full database, with 10-minute measurements at 1,134 sites between Jan. 1 2003 and Dec. 31, 2013. In addition to wind speed (m/s), direction (rad), temperature (C), rainfall (mm) and hours of daylight, we have geographic coordinates, site elevation (above sea level), height above ground of both anemometer and thermometer. We note that the temporal and spatial resolution of AMeDAS data is sufficiently low that most independent power providers could reasonably be expected to be able to acquire the same data used here in their region of interest at a similar resolution (and in all likelihood far higher). For the specific forecasting task considered in this study, we have filtered this network down to 44 sites with exceedingly high-quality data using the following criteria:

- **Min. average annual wind speed:** 2.5m/s

- **Max. missing value rate (per annum):** 0.1%
- **Min. contiguous sequence length (per annum):** 250 days (36,000 raw observations)

The forecasting task will be carried out over calendar year 2012 at all sites satisfying these criteria.

Geographically, the selected sites range from Okinawa and Kyushu through central Honshu and all the way north to Hokkaido, with sites in both urban and rural areas at both high and low elevations, spanning regions of varied topography, climate, and wind volatility (Fig. 3.1). The sampling period is 10 minutes at all sites considered. All implementations which call AMeDAS data are carried out using a library of routines we have written in the C and R programming languages explicitly for efficiently working with large quantities of AMeDAS data. The library is called `amemgr` and has been prepared and tested for public distribution on UNIX-like systems, though most functionality carries over to Windows systems as well.

Heliostat project data The second data set is provided by the Heliostat research and development project funded by Google, and is a high-quality set of very high-frequency observations, with a sampling rate of roughly 7.6Hz (Google, 2011). The observations were made between May 17–June 14, 2011 in central California. The former data set will be used for estimation of predictive densities with a horizon ranging between 1–5 hours, and the latter for densities with horizons spanning 1–5 seconds. There are no missing values, and at the single observation site, an array of five anemometers were arranged as displayed in the schematic in Fig. 3.2. For our purposes, we consider only observations from the anemometer labelled "E" with a height of 2.76m above ground.

Forecasting task The basic task is, at present time t, given a T-length sequence of observations $\mathbf{X}_t = (\mathbf{x}_{t-T+1}, \ldots, \mathbf{x}_t)$, and a parametric model $(\mathcal{P}, \Theta, v)$ where

$$\mathcal{P} = \{P : dP/d\mu = p(x; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$$

and $v(P) \coloneqq \boldsymbol{\theta}$, to estimate $\boldsymbol{\theta}$ and thus specify density function p. More specifically, for a horizon $k > 0$, we will typically have a model for the parameters $\boldsymbol{\theta}(\mathbf{X}_t, \boldsymbol{\gamma})$, indicating a dependence on the observations of interest and controllable parameters $\boldsymbol{\gamma}$. Let us denote by $\boldsymbol{\gamma}_{t+k}$ the parameters that, given $\mathbf{X}_t$, yield the true distribution, that is, $x_{t+k} \sim P(\boldsymbol{\theta}(\mathbf{X}_t, \boldsymbol{\gamma}_{t+k}))$. This is a density estimation task,

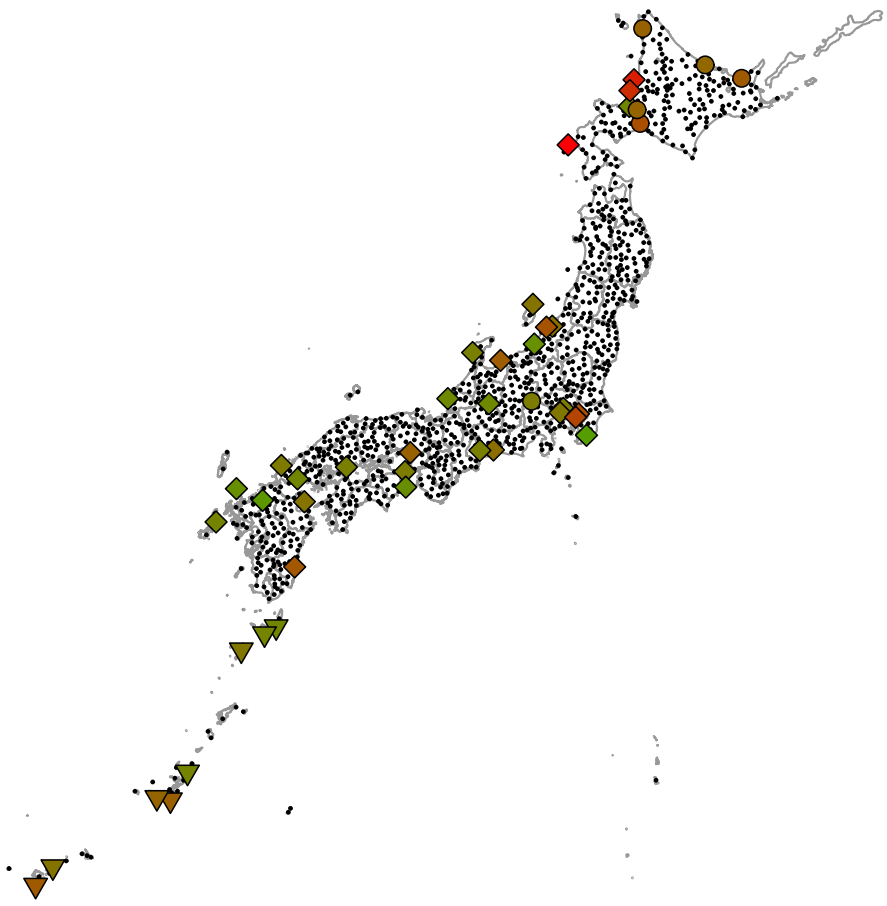

Figure 3.1: AMeDAS network, with forecasting target sites used in this study indicated by large symbols. Shapes denote annual average temperature (C): circles are < 8, diamonds are [8,18), triangles are $\geq$ 18 over 2012–2013. The green-red colour gradient represents normalized standard deviation of wind speed (m/s), with minimum of 1.5 and maximum 3.8.

and we shall call the density estimate made at time t, namely $p(x; \boldsymbol{\theta}(\mathbf{X}_t, \widehat{\boldsymbol{\gamma}}_t))$ the *predictive distribution*. There are numerous ways to generate point forecasts from a density forecast, though here we shall set $\widehat{x}_{t+k} = P^{-1}(0.5; \boldsymbol{\theta}(\mathbf{X}_t, \widehat{\boldsymbol{\gamma}}_t))$, i.e., a median-based estimate. The two task classes are k hour-ahead average hourly wind speed forecasts (using AMeDAS network), and k second-ahead 1s average

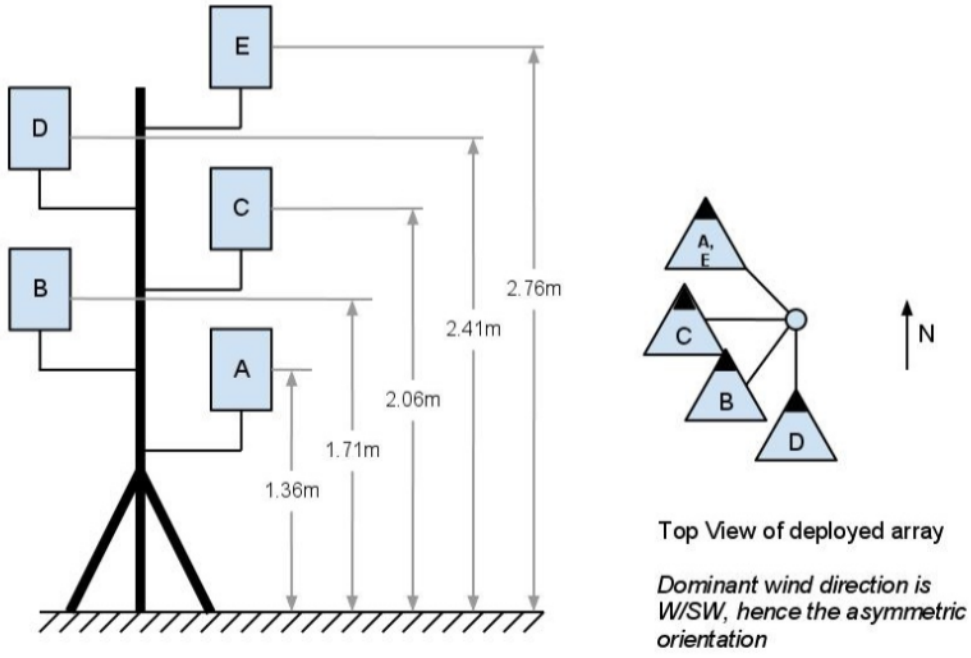

Figure 3.2: Anemometer array schematic from page 3 of the Heliostat project report documentation.

wind speed forecasts (using Heliostat data). For each class, the test is carried out for $k = 1, 2, \ldots, 5$.

With respect to notation, we shall denote by DET_A, DET_B, and DET_C the estimator W_N^α for α values of 1.05, 2, and 3. We denote W_N^{ld} by LOGDET. The tasks considered assume a general model of $\mathcal{P} = \mathcal{P}_W$ as given in section 2.5. Shape parameter κ is estimated as-is, while scale parameter λ is expressed as a simple AR(m) model, where $m = k + 3$. A window length of 15 days was used. Both shorter and longer window lengths, as well as smaller and larger order AR models were also tested, and overall trends in performance metrics remained the same, and thus the results shown here can be considered sufficiently representative. For the Heliostat project data, a similar model has been used with a window length of 3 minutes, AR model order between 1–6 time steps (here, seconds) determined using the calendar year prior to the yearlong test set to compute autocorrelation and setting a minimum threshold at 0.8. While the general para-

metric model is $\mathcal{P}_W$, the specific model depends on the horizon length (i.e., is task-dependent), though naturally will be common across estimators being compared for a given task.

3.2 Accuracy and robustness to horizon

We begin by focusing on general parameter estimation accuracy, and look at how sensitive the competing estimation methods are to horizon length. Note that in this sub-section, performance metrics for the AMeDAS data are averaged over all measurement sites.

Evaluation Four key metrics are used to quantify performance, that is, the quality of density estimation and accuracy of forecasts over the test set. We use the root mean squared error (RMSE) of signal estimates, long-run volatility difference, R^2 value, and probability of gross error. Note that mean absolute error (MAE) was also computed, but relative performance was essentially identical to RMSE, so to reduce redundancy we display only the former. If predictive distribution $\widehat{P}_t := P(x_t; \widehat{\boldsymbol{\theta}}_{t-k})$ yields forecast $\widehat{x}_t$, then defining $\epsilon_t := |x_t - \widehat{x}_t|$, we have for a test set of length $N > 0$ that RMSE is

$$\text{RMSE}(\widehat{P}) = \left(\frac{1}{N} \sum_{t=1}^{N} \epsilon_t^2 \right)^{1/2}.$$

Long-run volatility difference is simply the absolute value of the difference in standard deviation between the observed x_{t+k} and the forecast $\widehat{x}_{t+k} = P^{-1}(0.5; \widehat{\boldsymbol{\theta}}_t)$. If we let $P(x_t; \widehat{\boldsymbol{\theta}}_{t-k})$ be the cumulative probability of observation x_t assigned by the predictive distribution estimated at time $t-k$, and $\overline{P} = \sum_{t=1}^{N} P(x_t; \widehat{\boldsymbol{\theta}}_{t-k})$, then R^2 is defined

$$R^2(\widehat{P}) = \frac{\sum_{t=1}^{N} (P(x_t; \widehat{\boldsymbol{\theta}}_{t-k}) - \overline{P})^2}{\sum_{t=1}^{N} ((P(x_t; \widehat{\boldsymbol{\theta}}_{t-k}) - \overline{P})^2 + (\pi_t - \overline{P})^2)},$$

where π_t is the empirical cumulative probability of observation x_t. The probability of gross errors (PGE) is

$$\text{PGE}(\widehat{P}) = \frac{\sum_{t=1}^{N} \mathbf{1}[\epsilon_t \geq \bar{x}]}{N},$$

the relative frequency of poor estimates, where $\bar{x}$ is the arithmetic mean for a given site (in AMeDAS case) over the test period. This is a strict threshold for the 1h+ horizons, and we would thus be satisfied with < 0.10 as no site-specific model tuning is done here (Hering and Genton, 2010). Key results are given in Fig. 3.3.

Discussion of results Considering the eight panels of Fig. 3.3, we may readily confirm that all of the proposed methods showed dominant superiority over all probabilistic rivals in terms of RMSE, volatility difference, and PGE, over all time scales 1–5s and 1–5h. As well, using the most naive possible model, RMSE (and similarly MAE, not pictured) performance for the proposed estimators has already matched or outperformed the deterministic references, which tend to be exceedingly difficult to beat in very short-term estimation tasks. We note the model fit as gauged by R^2 is better for L1/L2CDF, which should be expected as its parameter optimization effectively maximizes this value. In any case, the model fit found using the proposed methods is far better than when using NLL or CRPS, is maintained on both hour and second time scales, and in addition the volatility difference is by far the smallest, suggesting a more balanced fit. We note that PGE at the 1–5h scale is within acceptable limits, and gross errors are virtually non-existent at the 1–5s time scale. Probabilistic forecasts at the order of seconds are important for many turbine control applications and yet the literature remains very sparse (Pinson, 2012; Jiang et al., 2013). Our initial results here may suggest a promising solution to such tasks.

3.3 Robustness to changes in spatio-temporal conditions

In this sub-section we focus exclusively on the AMeDAS network data, which includes observations taken at locations with different surrounding topographies, presence of nearby man-made obstructions, seasonal changes in climate, and anemometer elevation, among other factors known to impact short-term wind velocity forecasting models (Lynch, 2008; Ishihara and Yamaguchi, 2014). We shall consider how various performance metrics change with respect to signal measurement site. The reason is as follows: while in this paper we have not included a formal discussion regarding the robustness of proposed estimators, e.g. in the sense of Hampel et al. (1986), against standard references, we shall take sensitivity of estimation quality as a function of location to be an empirical proxy for robustness of the

proposed estimators to the above noted spatial and temporal conditions.

Evaluation As discussed in sub-section 3.2, we compute the PGE for each method at each site. If site index is given by s, we denote this explicitly as $\text{PGE}(\widehat{P}; s)$. For each method, fixing the site at which the smallest such probability was recorded (in the event of a tie, uniformly select one from tied candidates), denoted index s^*, we compute $\Delta(s) = \text{PGE}(\widehat{P}; s) - \text{PGE}(\widehat{P}; s^*)$, and Fig. 3.4 is the histogram of the $\Delta(s)$ values for all $s \neq s^*$, done separately for each competing method. Intuitively, this expresses the likelihood that, as the result of a random (uniform) change in measurement location away from the site at which a given method performs best, an increase in gross error probability will occur.

In addition, we include a number of illustrative visualizations of several of the four metrics considered above in 3.2; namely, we highlight the change in performance for each method and at each horizon as a function of site and average annual wind speed. These are given in Figs. 3.5–3.7. We note that similar results have been obtained (figures not included here) with the independent variable being average annual temperature, standard deviation of wind speed, and anemometer height. Finally, we quantify the sensitivity of competing estimators to site-specific conditions as well as temporal factors in Table 3.1 by looking at the standard deviation of RMSE and PGE for each method taken over different sites and seasons.

Discussion of results Both LOGDET and all DET methods performed similarly across all time horizons, and as such due to space constraints we show only the former. It is clear that the reference method showed little to no sensitivity to location at the 1h time scale, and even at the longest time horizon maintained a sensitivity on par with L1/L2CDF. The key references are shown to have higher sensitivity to location in both shown time scales, and indeed all horizons from 1–5h. These findings can be taken as empirical evidence of higher robustness to model error in the proposed approach.

In Figs. 3.5–**??**, we have plotted the performance of RMSE, R^2, and PGE as functions of site annual wind speed. The blue-red gradient spans 1–5 hours, and each separate plot corresponds to a particular method (same methods as Fig. 3.4). An OLS-fit line is also plotted for each time scale. As a general trend we see that for high velocity sites, forecast error increases, R^2 decreases, and GEP values are largely independent. Most importantly, one may visually confirm that the rate of performance deterioration of the proposed method is slower than the main probabilistic references, and overall site-to-site volatility is far less than NLL and

CRPS. We note that the linear trends observed in the metrics as a function of wind speed are essentially the same as that observed for standard deviation.

A natural way to visually confirm this sensitivity or lack thereof to site-dependent conditions is to use geographical coordinates to visualize this performance spatially on a map (Figs. 3.6–3.7). In these figures we have normalized RMSE and PGE for the 1h and 5h horizons for all representative methods of interest. The relative superiority of LOGDET (and indeed the other DET methods) across horizons is clear, and while CRPS performs relatively well for shorter horizons, it does not do as well when forecasts grow longer. It is also apparent that while NLL (and to a lesser extent CRPS and L1/2CDF) performs very well at some sites, it also performs very poorly at others, a stark contrast to the proposed methods which maintain steady performance across locations, in addition to superior output in an absolute sense.

To more clearly quantify sensitivity to spatio-temporal conditions, and thus empirically shed light on estimator robustness to such condition changes as well, we refer to Table 3.1. In the first five columns, we look at the standard deviation of RMSE taken over all the test sites, for each horizon. It is evident that the proposed methods show lower volatility across sites at all horizons, numerical values which corroborate visuals in Figs. 3.5–**??**. We may also confirm a similar trend in the volatility of PGE values across the entire tested network in the last five table columns. With respect to temporal non-stationarity, we consider explicitly model error which may periodically arise over seasons of the year. Dividing the year into four equal quarters, the standard deviation of metrics taken over these periods is given in the middle five table columns. While the common model used inherently adapts to seasonal changes, via a sliding window, this naive model in all likelihood does not capture all seasonal non-stationarity, and thus the results here can be taken to imply the proposed methods have superior robustness to the model error that arises. We note as well that very similar trends were found in metrics taken as a function of average annual site temperature, standard deviation of site wind speed, and anemometer height.

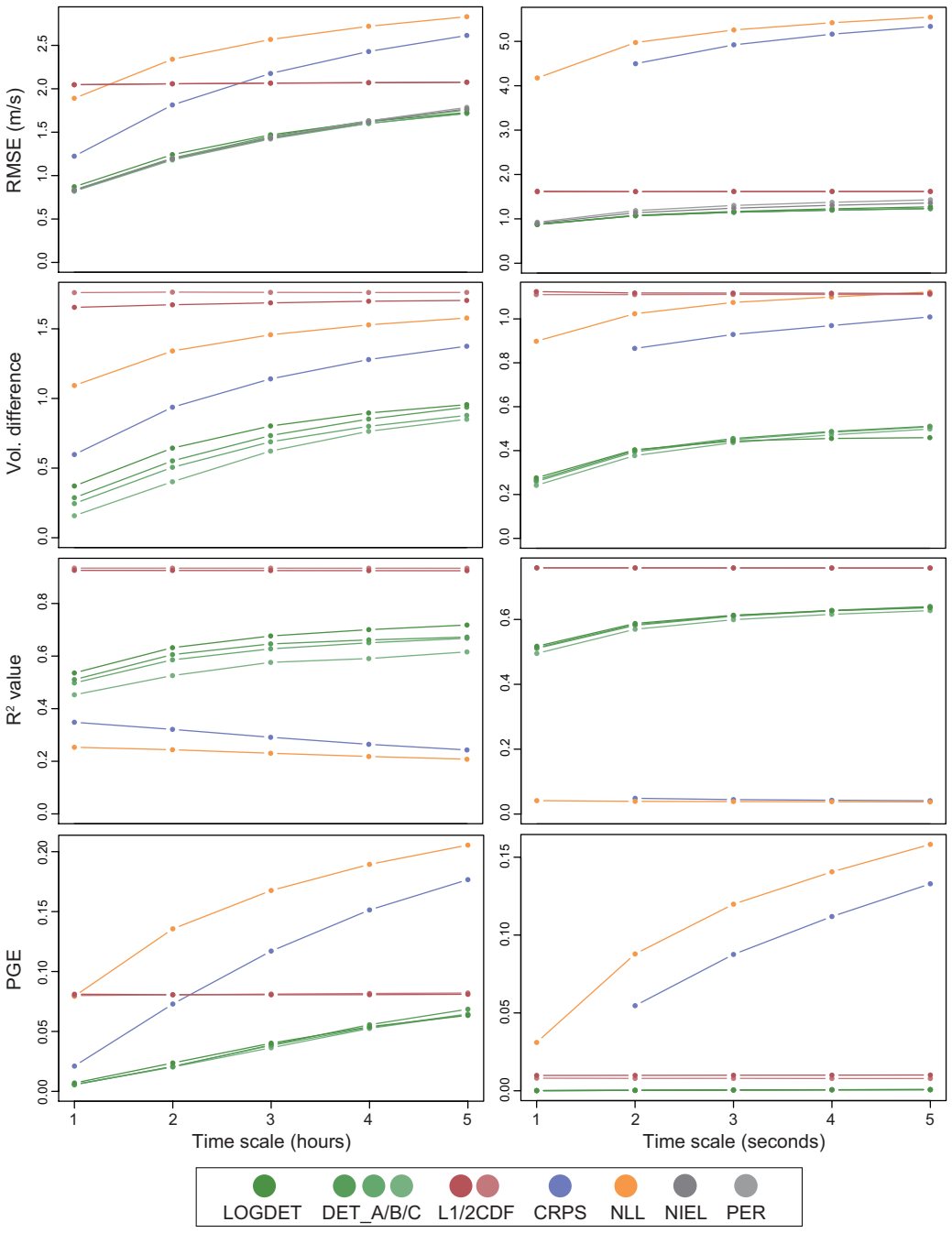

Figure 3.3: Overall results for all time scales. Left is 1–5h (AMeDAS), right is 1–5s (Heliostat).

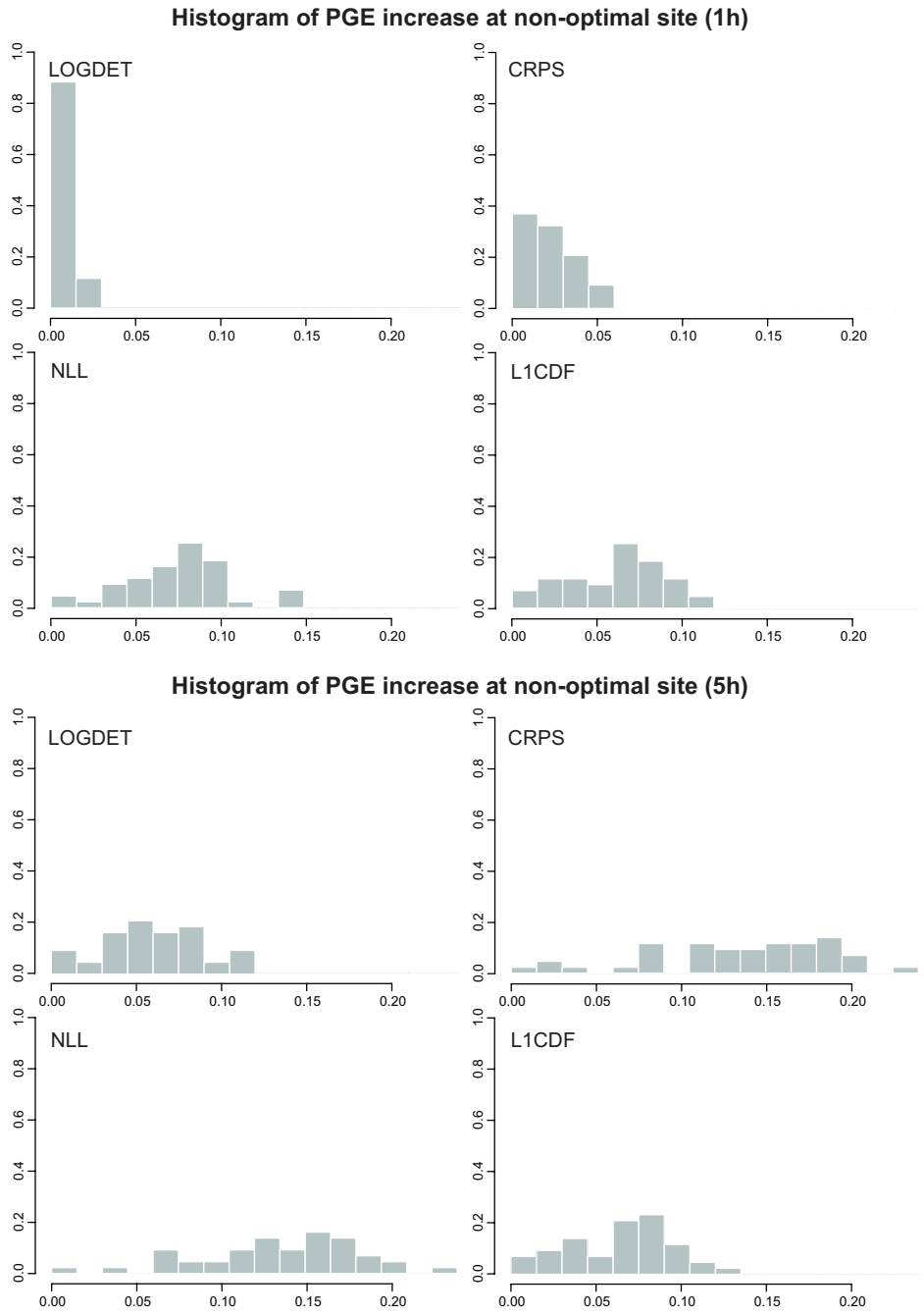

Figure 3.4: Increased likelihood of gross error by deviation from best fit site. First two rows are for 1h horizon, latter two rows for 5h horizon.

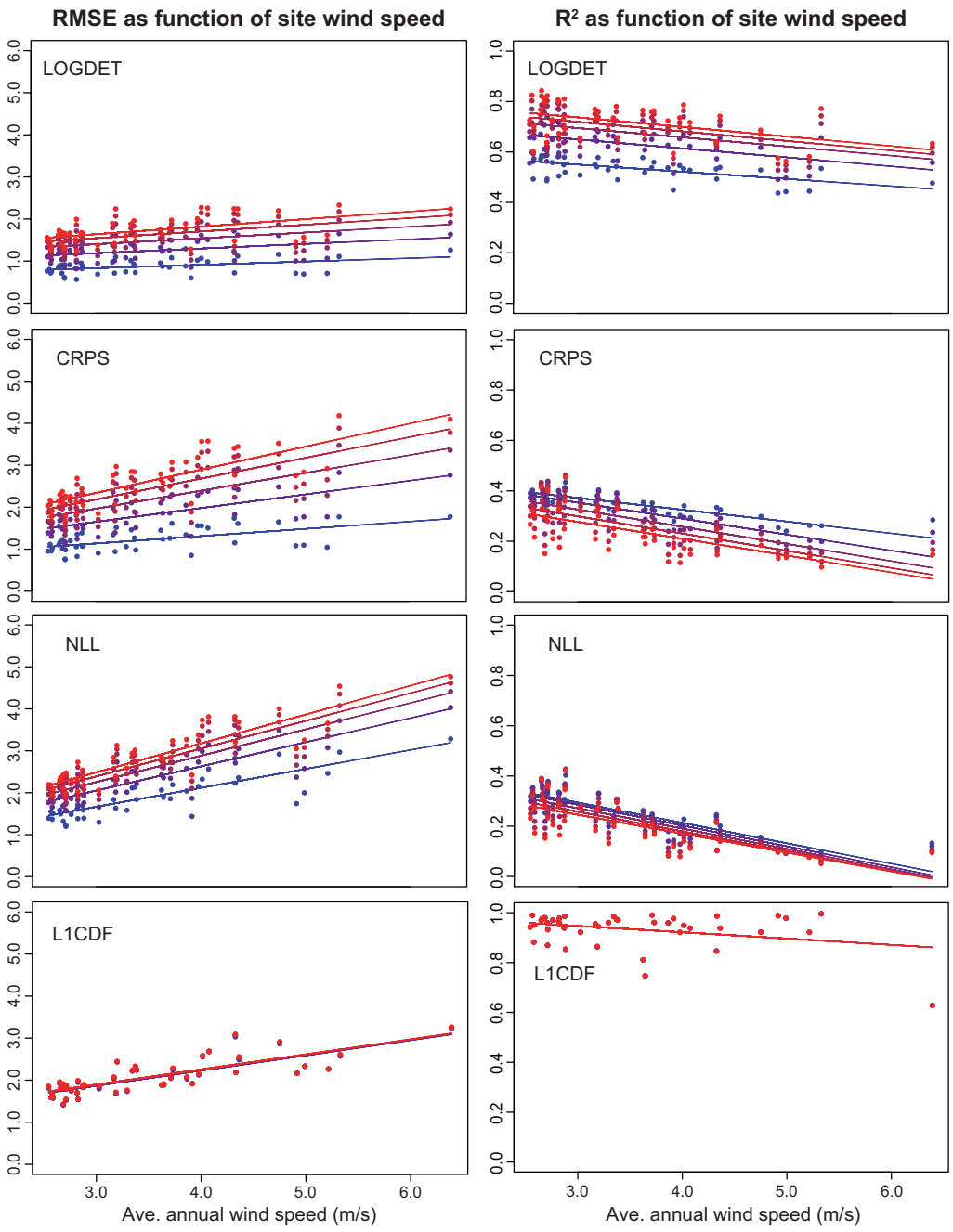

Figure 3.5: RMSE and R^2 value as function of wind speed over all horizons.

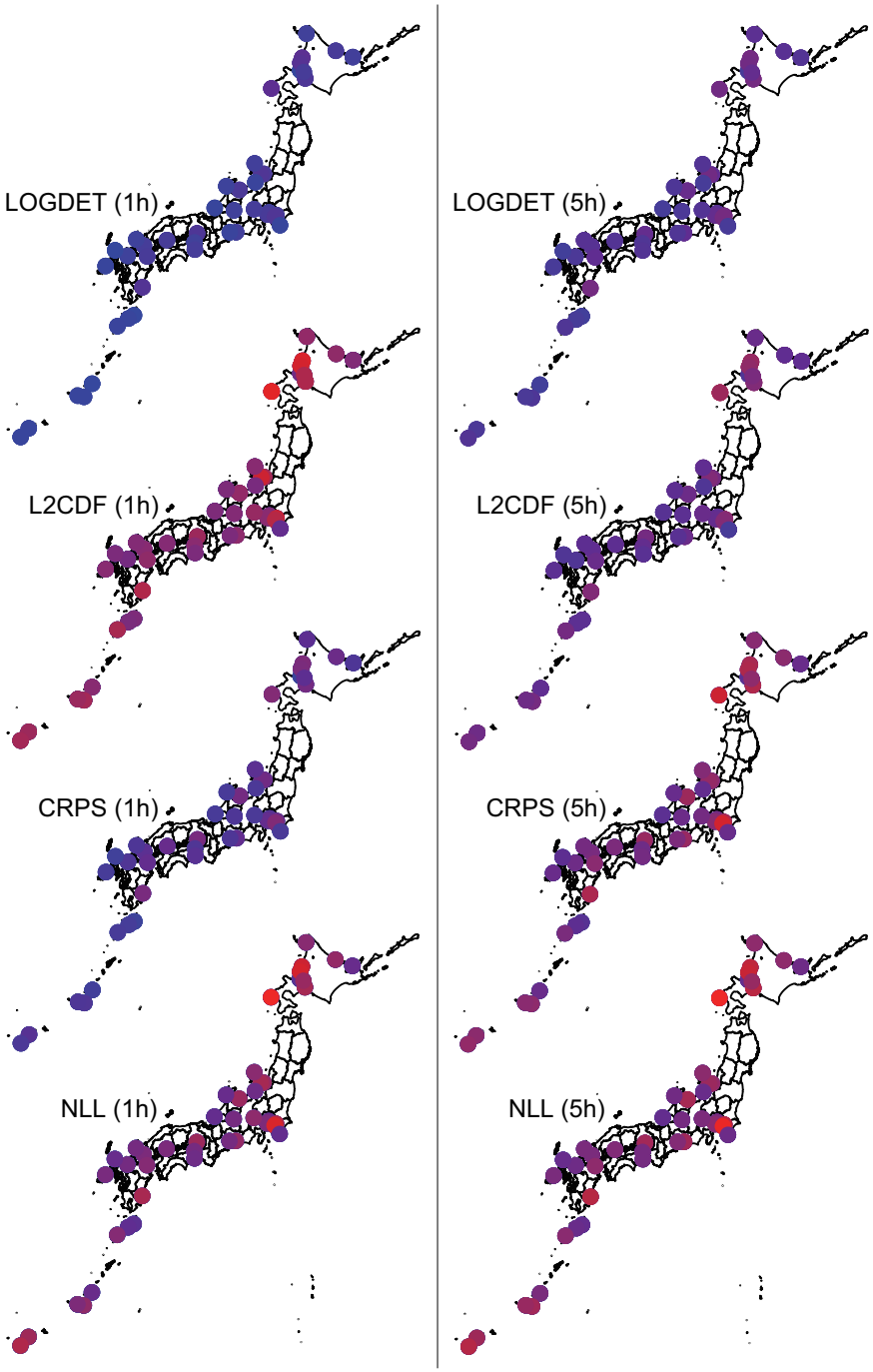

Figure 3.6: RMSE at all sites, for four methods and shortest/longest horizons, normalized over all methods.

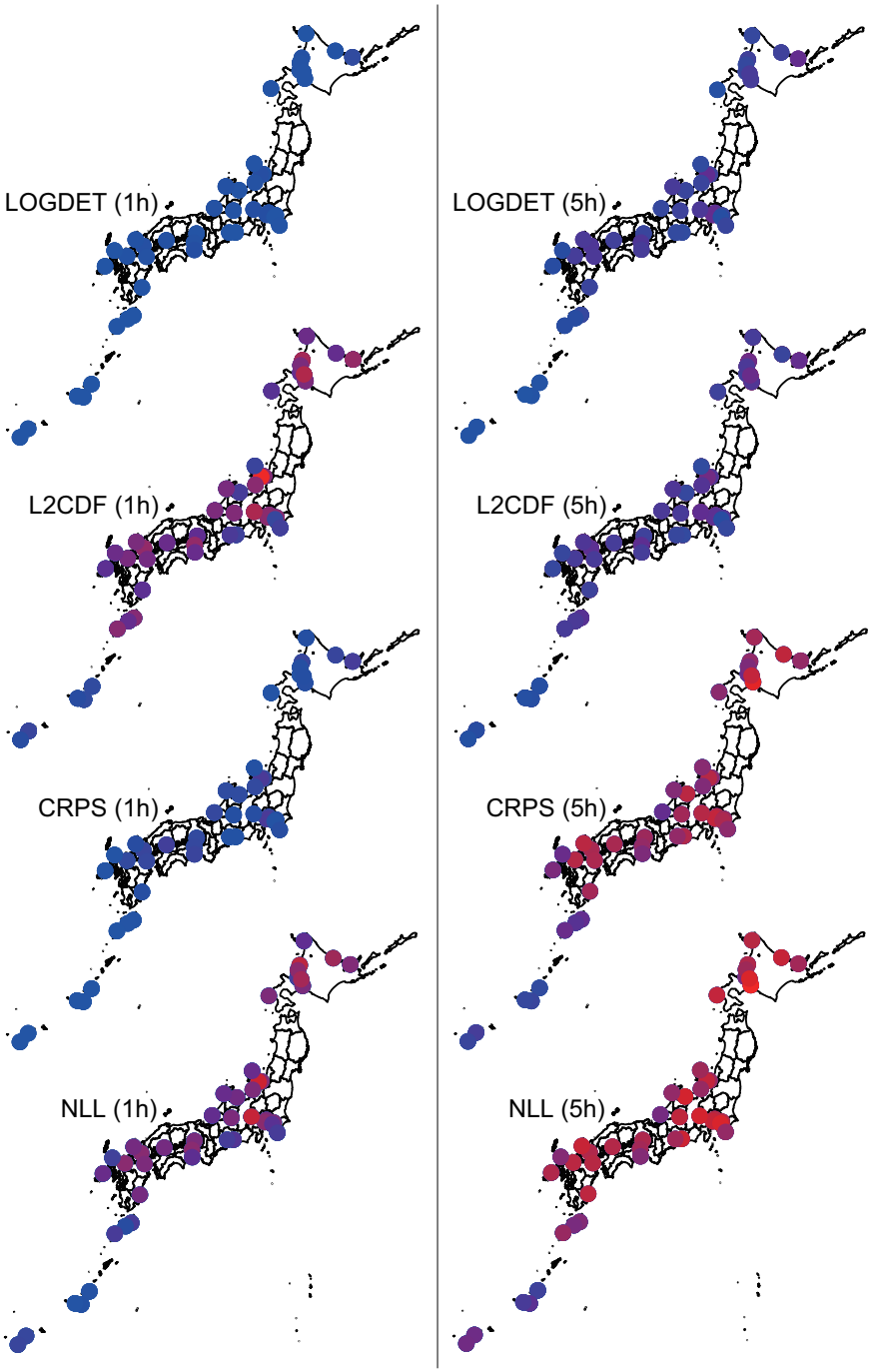

Figure 3.7: PGE at all sites, for four methods and shortest/longest horizons, normalized over all methods.

	SD of RMSE by site					**SD of RMSE by season**					**SD of PGE by site**				
	1h	*2h*	*3h*	*4h*	*5h*	*1h*	*2h*	*3h*	*4h*	*5h*	*1h*	*2h*	*3h*	*4h*	*5h*
DET_A	**0.155**	**0.235**	0.289	0.315	0.324	**0.060**	**0.086**	0.105	0.119	0.126	**0.005**	**0.015**	**0.022**	**0.030**	0.033
DET_B	0.157	0.237	**0.281**	**0.302**	0.320	0.063	0.089	**0.100**	**0.110**	**0.120**	**0.005**	**0.015**	0.025	0.031	0.034
DET_C	0.159	0.238	0.284	0.317	0.338	0.064	0.094	0.108	0.120	0.132	**0.005**	**0.015**	0.023	0.030	0.034
LOGDET	0.175	0.247	0.283	0.303	**0.318**	0.074	0.099	0.113	0.124	0.134	0.006	0.017	0.024	0.028	**0.031**
CRPS	0.280	0.430	0.520	0.580	0.619	0.100	0.155	0.189	0.214	0.234	0.016	0.037	0.049	0.055	0.057
NLL	0.502	0.606	0.652	0.681	0.702	0.218	0.264	0.279	0.287	0.291	0.035	0.044	0.048	0.050	0.050
L1CDF	0.408	0.408	0.409	0.411	0.412	0.163	0.163	0.164	0.166	0.166	0.030	0.030	0.031	0.031	**0.031**
L2CDF	0.406	0.402	0.401	0.400	0.400	0.166	0.169	0.170	0.169	0.170	0.033	0.031	0.031	0.031	**0.031**
PER	0.159	0.237	0.288	0.330	0.363	**0.060**	0.110	0.288	0.124	0.133	–	–	–	–	–
NIEL	0.159	0.237	0.288	0.328	0.361	0.061	0.089	0.109	0.121	0.130	–	–	–	–	–

Table 3.1: Sensitivity to spatio-temporal condition changes, discussion in main text.

Chapter 4

Concluding remarks

We begin by summarizing the work described in this thesis. First, we proved a straightforward result and proposed a minimum CRPS estimator in the Weibull case to be used as a strong benchmark in addition to likelihood-based estimators. Next, we presented a systematic new approach to constructing divergence-minimizing estimators, using elementary properties of convex functions that let us readily show the propriety of the loss functions used in said construction. In addition, we empirically validated the utility of several estimators derived following the proposed approach. Superior performance of all the proposed methods against standard reference estimators was confirmed across all forecast horizons considered was. Required data inputs (local observations only) were modest, and the proposed methods were derived with no explicit domain dependence.

Now considering the research goals described at the end of Chapter 1, we believe that the results of this study are strongly suggestive of a promising new parameter estimation method which can be used as-is for both all the forecast horizon lengths considered (short and very short). With respect to very short-term forecasts, as the literature on methods for horizons on the order of several seconds is sparse, these results may be considered promising for high-frequency applications. With respect to the short-term prediction, we were able to confirm that even using a simple, reliable, parsimonious predictive distribution model which is readily adapted to an arbitrary site, there were dramatic differences in performance across estimators. This observation is in line with our initial hypothesis regarding the role of parameter estimation methods in the forecasting problem. The lack of robustness of NLL has of course been reported before (Gneiting et al., 2006), but it is interesting to note that forecasters using the proposed estimators saw decidedly better accuracy and were empirically more robust than even the CRPS-based fore-

caster. These observations suggest that an application of the proposed estimation methods to a more general class of site-adaptive models may realize a significant step towards realizing a non-trivial benchmark forecaster usable under arbitrary conditions.

The task of carrying out such a general test of utility, namely on a wider class of distributions (non-Weibull, multi-modal) and models (spatio-temporal) is a very natural next step for this research to take in the wind-related forecasting context. In a more general context of inference in parametric models, the examples of estimators proposed here were rudimentary and domain-free, and still achieved very strong estimation accuracy as well as empirical robustness to model error arising from known temporal non-stationarity. The design of more sophisticated estimators with specific problem domains explicitly reflected is another very natural line of future work. From a different perspective, a more formal theoretical evaluation of the robustness of the estimators considered may be of interest. While stricter assumptions on the classes of models permitted will likely be required, this may yield valuable new insights into the empirical results of this study.

Appendix A

Supplementary materials

A.1 Equivalent expressions of CRPS

There are a number of equivalent expressions for the CRPS as defined above. We begin with a form that can be useful in the pursuit of closed-form CRPS expressions. One of the standard references here is Laio and Tamea (2007), though we note the second equality in equation (6) in their paper is incorrect, however the expression on the final line of that equation is the one we want. The argument of importance is as follows. Let y be a random variable $y \sim F$, where F here denotes a distribution function, with density function f. Now, fix some y, and note that

$$\int_0^1 \{|F^{-1}(1-\gamma) - y| + 2(\gamma - 0.5)(F^{-1}(1-\gamma) - y)\}\, d\gamma$$

can, after a change of variables $y^* := F^{-1}(1-\gamma)$ be written

$$\begin{aligned}
C(y^*) &:= \int_{\infty}^{-\infty} \{|y^* - y| + 2(0.5 - F(y^*))(y^* - y)\}(-1) f(y^*)\, dy^* \\
&= \int_{-\infty}^{\infty} \{|y^* - y| + 2(0.5 - F(y^*))(y^* - y)\} f(y^*)\, dy^* \\
&= \int_{-\infty}^{\infty} \{|y^* - y| + (y^* - y) - 2F(y^*)(y^* - y)\} f(y^*)\, dy^* \\
&= \int_{-\infty}^{\infty} \{\mathbf{1}[y^* > y] 2(y^* - y) - 2F(y^*)(y^* - y)\} f(y^*)\, dy^* \\
&= \int_{-\infty}^{\infty} 2(\mathbf{1}[y^* > y] - F(y^*))(y^* - y) f(y^*)\, dy^*.
\end{aligned}$$

We now proceed to use integration by parts. Note that as a function of y^* the indicator function is not continuous at y so we break it up, noting that for $y \neq y^*$ we have

$$\frac{d}{dy^*}(-1)(\mathbf{1}[y^* > y] - F(y^*))^2 = 2(\mathbf{1}[y^* > y] - F(y^*)) f(y^*)$$

and thus using integration by parts,

$$\int_y^{\infty} 2(\mathbf{1}[y^* > y] - F(y^*)) f(y^*)(y^* - y)\, dy^* =$$
$$\left[(-1)(\mathbf{1}[y^* > y] - F(y^*))^2 (y^* - y)\right]_y^{\infty} - \int_y^{\infty} (-1)(\mathbf{1}[y^* > y] - F(y^*))^2 (1)\, dy^*.$$

Since the squared factor in the term without the integral sends the entire term to 0 as $y^* \to \pm\infty$, and the resulting term as $y^* \to y$ in both cases will cancel out, we have clearly that

$$\left[(-1)(\mathbf{1}[y^* > y] - F(y^*))^2 (y^* - y)\right]_y^{\infty} + \left[(-1)(\mathbf{1}[y^* > y] - F(y^*))^2 (y^* - y)\right]_{-\infty}^{y} = 0$$

and thus

$$\begin{aligned}
&\int_{-\infty}^{\infty} 2(\mathbf{1}[y^* > y] - F(y^*))(y^* - y) f(y^*)\, dy^* = \\
&\qquad \int_{-\infty}^{y} 2(\mathbf{1}[y^* > y] - F(y^*))(y^* - y) f(y^*)\, dy^* + \int_{y}^{\infty} 2(\mathbf{1}[y^* > y] - F(y^*))(y^* - y) f(y^*)\, dy^* \\
&\qquad = -\int_{-\infty}^{y} (-1)(\mathbf{1}[y^* > y] - F(y^*))^2\, dy^* - \int_{y}^{\infty} (-1)(\mathbf{1}[y^* > y] - F(y^*))^2\, dy^* \\
&\qquad = \int_{-\infty}^{\infty} (\mathbf{1}[y^* > y] - F(y^*))^2\, dy^* \\
&\qquad = \mathrm{CRPS}(F, y).
\end{aligned}$$

There are thus many possible integrals we can equate to the CRPS from this result; the key form is the one from which we are able to do the integration by parts. As long as we can get a particular expression into that form, we have equivalence to the CRPS.

An equivalent and quite nice and clean form is

$$\mathrm{CRPS}(F, y) = \int_0^1 2\left(\mathbf{1}[F^{-1}(\gamma) > y] - \gamma\right)\left(F^{-1}(\gamma) - y\right) d\gamma,$$

used explicitly by Friederichs and Thorarinsdottir (2012) as well as Gneiting and Ranjan (2011). Proof of this form is essentially immediate, as setting $y^* \coloneqq F^{-1}(\gamma)$ yields

$$\int_0^1 2\left(\mathbf{1}[F^{-1}(\gamma) > y] - \gamma\right)\left(F^{-1}(\gamma) - y\right) d\gamma = \int_{-\infty}^{\infty} 2\left(\mathbf{1}[y^* > y] - F(y^*)\right)(y^* - y) f(y^*)\, dy^*$$

which is precisely the form we used in the preceding argument for the key integration step.

A.2 Auxiliary results

Proposition. $\mathbb{S}_+^d$ is an open, convex subset of $\mathbb{R}^{d(d+1)/2}$.

Proof. Using the bilinearity of inner product on real vector spaces, it clearly follows for arbitrary $\mathbf{W}, \mathbf{U} \in \mathbb{S}_+^d$ that $\lambda\mathbf{W} + (1-\lambda)\mathbf{U} \in \mathbb{S}_+^d$, $\lambda \in (0, 1)$, so $\mathbb{S}_+^d$ is convex.

With the Frobenius norm on $d \times d$ symmetric positive definite matrices, having fixed some $\mathbf{W}$ and $\delta > 0$, we have immediately that $\mathbf{U} \in B_\delta(\mathbf{W}) \implies |w_{i,j} - u_{i,j}| < \delta$. If symmetric $\mathbf{Y} \in \mathbb{S}^d$ satisfies $|y_{i,j}| \in [0, 1]$, then we note that

$$\det(\mathbf{W} + \varepsilon\mathbf{Y}) = \det \mathbf{W} + p(\varepsilon; \mathbf{Y})$$

where p is a d-degree polynomial in ε. By the given assumptions, for any ε we have that p is a bounded function of $\mathbf{Y}$. Also, if we note $\det \mathbf{W} > 0$, we may take $\varepsilon > 0$ such that $\det(\mathbf{W} + \varepsilon\mathbf{Y}) > 0$ for all valid $\mathbf{Y}$.

An identical argument can be applied to the principal minors, which gives us that for any principal minor of diagonal length $k = 1, 2 \ldots, d$ we may take a $\varepsilon > 0$ such that the $\det[\mathbf{W} + \varepsilon\mathbf{Y}]_k > 0$, and from this it follows that $\mathbf{W} + \varepsilon\mathbf{Y} \in \mathbb{S}^d_+$. Fixing ε as the smallest such perturbance taken over finitely many $k = 1, \ldots, d$ we note that for all $\mathbf{U} \in B_\varepsilon(\mathbf{W})$, clearly there exists a $\mathbf{Y}$ as required above which satisfies $\mathbf{U} = \mathbf{W} + \varepsilon\mathbf{Y}$. Thus we have $B_\varepsilon(\mathbf{W}) \subset \mathbb{S}^d_+$, implying that $\mathbb{S}^d_+$ is an open subset of $\mathbb{R}^{d(d+1)/2}$. □

Bibliography

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*, volume 55. National Bureau of Standards, Applied Mathematics Series.

Alexiadis, M., Dokopoulos, P., Sahsamanoglou, H., and Manousaridis, I. (1998). Short-term forecasting of wind speed and related electrical power. *Solar Energy*, 63(1):61–68.

Amari, S.-I. (2009). α-divergence is unique, belonging to both f-divergence and Bregman divergence classes. *IEEE Transactions on Information Theory*, 55(11):4925–4931.

Artin, E. (1964). *The Gamma Function*. Holt, Rinehart and Winston.

Baïle, R., Muzy, J.-F., and Poggi, P. (2011). Short-term forecasting of surface layer wind speed using a continuous random cascade model. *Wind Energy*, 14(6):719–734.

Banerjee, A., Guo, X., and Wang, H. (2005). On the optimality of conditional expectation as a Bregman predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669.

Barnard, G. (1951). The theory of information. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(1):46–64.

Barrodale, I. and Roberts, F. D. (1973). An improved algorithm for discrete l_1 linear approximation. *SIAM Journal on Numerical Analysis*, 10(5):839–848.

Bathurst, G. N., Weatherill, J., and Strbac, G. (2002). Trading wind generation in short term energy markets. *IEEE Transactions on Power Systems*, 17(3):782–789.

BIBLIOGRAPHY

Bechrakis, D. and Sparis, P. (1998). Wind speed prediction using artificial neural networks. *Wind Engineering*, 22(6):287–296.

Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore.

Bjerknes, V. (1904). Das problem der wettervorhersage, betrachtet vom standpunkte der mechanik und der physik. *Meteorolgische Zeitschrift*, pages 1–7.

Blischke, W. R. and Scheuer, E. (1986). Tabular aids for fitting Weibull moment estimates. *Naval research logistics quarterly*, 33(1):145–153.

Bossanyi, E. (1985). Short-term wind prediction using Lalman filters. *Wind Engineering*, 9(1):1–8.

Boukhezzar, B. and Siguerdidjane, H. (2011). Nonlinear control of a variable-speed wind turbine using a two-mass model. *IEEE Transactions on Energy Conversion*, 26(1):149–162.

Box, G. E., Jenkins, G. M., and Reinsel, G. C. (2008). *Time Series Analysis: Forecasting and Control*. John Wiley & Sons.

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217.

Brown, B. G., Katz, R. W., and Murphy, A. H. (1984). Time series models to simulate and forecast wind speed and wind power. *Journal of Climate and Applied Meteorology*, 23:1184–1195.

Brown, T. (1974). *Admissible scoring systems for continuous distributions (Manuscript P-5235)*. The Rand Corporation, Santa Monica, CA.

Burton, T., Jenkins, N., Sharpe, D., and Bossanyi, E. (2011). *Wind Energy Handbook*. John Wiley & Sons.

Charney, J. G., Fjörtoft, R., and Neumann, J. v. (1950). Numerical integration of the barotropic vorticity equation. *Tellus*, 2(4):237–254.

BIBLIOGRAPHY

Cichocki, A. and Amari, S.-I. (2010). Families of alpha-beta- and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568.

Collins, M., Schapire, R. E., and Singer, Y. (2002). Logistic regression, AdaBoost and Bregman distances. *Machine Learning*, 48(1-3):253–285.

Conradsen, K., Nielsen, L., and Prahm, L. (1984). Review of Weibull statistics for estimation of wind speed distributions. *Journal of Climate and Applied Meteorology*, 23(8):1173–1183.

Csiszár, I. (1972). A class of measures of informativity of observation channels. *Periodica Mathematica Hungarica*, 2(1):191–213.

Csiszár, I. (1991). Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *The Annals of Statistics*, pages 2032–2066.

Csiszár, I. (2008). Axiomatic characterizations of information measures. *Entropy*, 10(3):261–273.

Fabbri, A., Roman, T. G. S., Abbad, J. R., and Quezada, V. M. (2005). Assessment of the cost associated with wind generation prediction errors in a liberalized electricity market. *IEEE Transactions on Power Systems*, 20(3):1440–1446.

Friederichs, P. and Thorarinsdottir, T. L. (2012). Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics*, 23(7):579–594.

Fujisawa, H. and Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9):2053–2081.

Galassi, M., Gough, B., Jungman, G., Theiler, J., Davies, J., Booth, M., and Rossi, F. (2013). *The GNU Scientific Library Reference Manual (3rd Ed.)*. ISBN 0954612078.

Giebel, G., Brownsword, R., Kariniotakis, G., Denhard, M., and Draxl, C. (2011). The state-of-the-art in short-term prediction of wind power: A literature overview. Technical report, ANEMOS.plus.

BIBLIOGRAPHY

Gneiting, T., Larson, K., Westrick, K., Genton, M. G., and Aldrich, E. (2006). Calibrated probabilistic forecasting at the stateline wind energy center: The regime-switching space-time method. *Journal of the American Statistical Association*, 101(475):968–979.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gneiting, T., Raftery, A. E., Westveld III, A. H., and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133(5):1098–1118.

Gneiting, T. and Ranjan, R. (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, 29(3).

Google (2011). REC Initiative. `https://www.google.org/rec.html`. Online; accessed 21-January-2015.

Grünwald, P. D. and Dawid, A. P. (2004). Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *Annals of Statistics*, pages 1367–1433.

Halmos, P. R. (1974). *Measure Theory, Graduate Texts in Mathematics (18)*. Springer-Verlag New York.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust statistics: the approach based on influence functions*. John Wiley & Sons.

Havil, J. (2003). *Gamma: Exploring Euler's Constant*. Princeton University Press.

Hein, M. and Bousquet, O. (2005). Hilbertian metrics and positive definite kernels on probability measures. *AISTATS2005*.

Hennessey, J. P. (1977). Some aspects of wind power statistics. *Journal of Applied Meteorology*, 16(2):119–128.

Hering, A. S. and Genton, M. G. (2010). Powering up with space-time wind forecasting. *Journal of the American Statistical Association*, 105(489):92–104.

BIBLIOGRAPHY

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570.

Holland, M. J. and Ikeda, K. (2014). Forecasting in wind energy applications with site-adaptive Weibull estimation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014)*, Florence, Italy.

Huber, P. J. (1981). *Robust Statistics*. John Wiley & Sons.

Hutting, H. and Cleijne, J. (1999). The price of large scale offshore wind energy in a free electricity market. In *European Wind Energy Conference*, Nice, France.

Ishihara, T. and Yamaguchi, A. (2014). Prediction of the extreme wind speed in the mixed climate region by using Monte Carlo simulation and measure-correlate-predict method. *Wind Energy*.

Japan Meteorological Agency (2013). About AMeDAS. `http://www.jma.go.jp/jma/kishou/know/amedas/kaisetsu.html`. Online (Japanese); accessed 21-January-2015.

Jiang, Y., Song, Z., and Kusiak, A. (2013). Very short-term wind speed forecasting with Bayesian structural break model. *Renewable Energy*, 50:637–647.

Justus, C., Hargraves, W., and Yalcin, A. (1976). Nationwide assessment of potential output from wind-powered generators. *Journal of Applied Meteorology*, 15(7):673–678.

Kalnay, E., Lord, S. J., and McPherson, R. D. (1998). Maturity of operational numerical weather prediction: Medium range. *Bulletin of the American Meteorological Society*, 79(12):2753–2769.

Kamal, L. and Jafri, Y. Z. (1997). Time series models to simulate and forecast hourly averaged wind speed in Quetta, Pakistan. *Solar Energy*, 61(1):23–32.

Kariniotakis, G., Matos, M., and Miranda, V. (1999). Assessment of the benefits from advanced load and wind power forecasting in autonomous power systems. In *European Wind Energy Conference*, pages 391–394, Nice, France.

Kavasseri, R. G. and Seetharaman, K. (2009). Day-ahead wind speed forecasting using f-ARIMA models. *Renewable Energy*, 34(5):1388–1393.

Kullback, S. (1968). *Information theory and statistics*. Dover Publications.

BIBLIOGRAPHY

Kusiak, A., Li, W., and Song, Z. (2010). Dynamic control of wind turbines. *Renewable Energy*, 35(2):456–463.

Kusiak, A. and Zhang, Z. (2012). Control of wind turbine power and vibration with a data-driven approach. *Renewable Energy*, 43:73–82.

Lafferty, J., Della Pietra, S., and Della Pietra, V. (1997). Statistical learning algorithms based on Bregman distances.

Laio, F. and Tamea, S. (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences*, 11(4):1267–1277.

Lerch, S. and Thorarinsdottir, T. L. (2013). Comparison of nonhomogeneous regression models for probabilistic wind speed forecasting. *Tellus A*, 65.

Lorenz, E. N. (1969). The predictability of a flow which possesses many scales of motion. *Tellus*, 21(3):289–307.

Lynch, P. (2006). *The emergence of numerical weather prediction: Richardson's dream*. Cambridge University Press.

Lynch, P. (2008). The origins of computer weather prediction and climate modeling. *Journal of Computational Physics*, 227(7):3431–3444.

Makarov, Y. V., Loutan, C., Ma, J., and de Mello, P. (2009). Operational impacts of wind generation on California power systems. *IEEE Transactions on Power Systems*, 24(2):1039–1050.

Mihram, G. (1977). Weibull shape parameter: estimation by moments. In *Proceedings of 31st Annual Technical Conference of American Society for Quality Control*, pages 315–322.

Minami, M. and Eguchi, S. (2002). Robust blind source separation by beta divergence. *Neural Computation*, 14(8):1859–1886.

Morgan, E. C., Lackner, M., Vogel, R. M., and Baise, L. G. (2011). Probability distributions for offshore wind speeds. *Energy Conversion and Management*, 52(1):15–26.

Newby, M. (1980). The properties of moment estimators for the Weibull distribution based on the sample coeffkient of variation. *Technometrics*, 22(2):187–194.

BIBLIOGRAPHY

Nielsen, L., Morthorst, P., Skytte, K., Jensen, P. H., Jørgensen, P., Eriksen, P., Sørensen, A., Nissen, F., Godske, B., Ravn, H., Søndergreen, C., Stærkind, K., and Havsager, J. (1999). Wind power and a liberalised North European electricity exchange. In *European Wind Energy Conference*, pages 379–382, Nice, France.

Nielsen, T. S., Joensen, A., Madsen, H., Landberg, L., and Giebel, G. (1998). A new reference for wind power forecasting. *Wind Energy*, 1(1):29–34.

Pinson, P. (2012). Very-short-term probabilistic forecasting of wind power with generalized logit–normal distributions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(4):555–576.

Pinson, P., Chevallier, C., and Kariniotakis, G. N. (2007). Trading wind generation from short-term probabilistic forecasts of wind power. *IEEE Transactions on Power Systems*, 22(3):1148–1156.

Potter, C. W. and Negnevitsky, M. (2006). Very short-term wind forecasting for Tasmanian power generation. *IEEE Transactions on Power Systems*, 21(2):965–972.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rinne, H. (2010). *The Weibull distribution: A handbook*. CRC Press.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press.

Seguro, J. and Lambert, T. (2000). Modern estimation of the parameters of the Weibull wind speed distribution for wind energy analysis. *Journal of Wind Engineering and Industrial Aerodynamics*, 85(1):75–84.

Slootweg, J., De Haan, S., Polinder, H., and Kling, W. (2003). General model for representing variable speed wind turbines in power system dynamics simulations. *IEEE Transactions on Power Systems*, 18(1):144–151.

Soman, S. S., Zareipour, H., Malik, O., and Mandal, P. (2010). A review of wind power and wind speed forecasting methods with different time horizons. In *North American Power Symposium (NAPS), 2010*, pages 1–8.

BIBLIOGRAPHY

Sørensen, B. and Meibom, P. (1999). Can wind power be sold in a deregulated electricity market? In *European Wind Energy Conference*, pages 375–378, Nice, France.

Taylor, J. W., McSharry, P. E., and Buizza, R. (2009). Wind power density forecasting using ensemble predictions and time series models. *IEEE Transactions on Energy Conversion*, 24(3):775–782.

Walford, C. A. (2006). *Wind turbine reliability: understanding and minimizing wind turbine operation and maintenance costs*. United States Department of Energy.

Wallace, J. M. and Hobbs, P. V. (2006). *Atmospheric Science: An Introductory Survey, Second Edition*. Academic Press.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT Press.

Zhang, J. (2004). Divergence function, duality, and convex analysis. *Neural Computation*, 16(1):159–195.

Zhang, J., Chowdhury, S., Messac, A., and Castillo, L. (2013). A multivariate and multimodal wind distribution model. *Renewable Energy*, 51:436–447.

Zugno, M., Morales, J., Pinson, P., and Madsen, H. (2013). Pool strategy of a price-maker wind power producer. *IEEE Transactions on Power Systems*, 28(3):3440–3450.