Regions of Reliability in the Evaluation of Multivariate Probabilistic Forecasts

Étienne Marcotte [†] **Valentina Zantedeschi** [†] **Alexandre Drouin** [†] **Nicolas Chapados** [†]

Abstract

Multivariate probabilistic time series forecasts are commonly evaluated via proper scoring rules, i.e., functions that are minimal in expectation for the ground-truth distribution. However, this property is not sufficient to guarantee good discrimination in the non-asymptotic regime. In this paper, we provide the first systematic finite-sample study of proper scoring rules for time-series forecasting evaluation. Through a power analysis, we identify the "region of reliability" of a scoring rule, i.e., the set of practical conditions where it can be relied on to identify forecasting errors. We carry out our analysis on a comprehensive synthetic benchmark, specifically designed to test several key discrepancies between ground-truth and forecast distributions, and we gauge the generalizability of our findings to real-world tasks with an application to an electricity production problem. Our results reveal critical shortcomings in the evaluation of multivariate probabilistic forecasts as commonly performed in the literature.

1. Introduction

The forecasting of time-varying quantities is a fundamental component of decision-making in fields like economics, operation management, and healthcare (Peterson, 2017; Heizer et al., 2023). In this context, a proper characterization of uncertainty is key to reasoning about potential futures and their respective likelihood. This has motivated the problem of *multivariate probabilistic forecasting*, which consists in estimating the joint distribution of the future values of quantities of interest (Gneiting & Katzfuss, 2014).

In this setting, all estimates are not equal. Depending on the application, certain kinds of errors, e.g., failures to properly model statistical dependencies between variables, can have a catastrophic impact on downstream decisions. It is thus critical to develop methodological tools to assess the quality of distributional forecasts produced by statistical models.

For this, the literature has primarily focused on developing *proper scoring rules* (Gneiting & Raftery, 2007), which are designed to reach a minimum when the forecast and the ground-truth distributions match. Among them, the *Negative Log-Likelihood* has been shown to be an optimal discriminator of erroneous distributions (Neyman & Pearson, 1933). However, the use of the likelihood is not always practical since many models do not allow for its efficient calculation. Hence, time-series practitioners have turned to other proper scoring rules, such as the *Continuous Ranked Probability Score* (CRPS; Matheson & Winkler 1976) to evaluate forecasts. While proper in theory, the discriminative performance of such scoring rules in practical conditions, where the dimensionality of the problem is large and the sample size is small in comparison, is poorly understood.

This work aims to study the reliability of proper scoring rules for the evaluation of multivariate probabilistic forecasts in realistic finite-sample settings. We quantify reliability as the statistical power of a rule at discriminating between data sampled respectively from the ground truth and an erroneous forecast. We introduce a comprehensive benchmark to systematically measure the ability of scoring rules to detect failures in forecasts of practical interest. Our results emphasize sets of conditions (in terms of problem dimensionality and Monte Carlo approximation quality) under which each scoring rule is reliable, dubbed *regions of reliability*,[1] and, most interestingly, reveal significant shortcomings such as the general inability of the studied scoring rules at detecting some basic forecasting errors.

Contributions:

- We propose a methodology, based on power analysis, to assess the reliability of proper scoring rules in the evaluation of multivariate probabilistic forecasts (§4);
- We propose an extensive benchmark that reveals *regions of reliability* for five common proper scoring rules and 19 types of forecasting errors (§5);
- We show that our findings generalize to three real-world

[†]ServiceNow Research, Montréal, Canada. Correspondence to: Étienne Marcotte <etienne.marcotte@servicenow.com>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1]We note that our regions of reliability are distinct from *reliability diagrams*, which are graphical tools for checking forecast calibration.

settings beyond this benchmark (§6);

- We present a critical review of experimental practices in the recent literature in light of these results and make recommendations for future developments in the field (§7).

2. Background

2.1. Multivariate Probabilistic Forecasting

We consider the problem of discrete-time probabilistic forecasting of multivariate time series, where we seek to accurately estimate the joint distribution of some numerical quantities of interest over a future time horizon, given historical data. Formally, let $X_t \in \mathbb{R}^v$ be a random vector containing the values of v variables at time t. The problem consists in accurately estimating the joint conditional distribution

$$P(X_{t+1},...,X_T \mid X_1,...,X_t), \tag{1}$$

where $X_1,...,X_t$ are past observations of the variables and $X_{t+1},...,X_T$ are forecasted values up to time T. This is in contrast with point and quantile forecasting, which typically aim at estimating the conditional mean or quantiles of this distribution, respectively (West & Harrison, 1997; Cai, 2002).

2.2. Motivating Example in Decision-Making

Many problems can only be effectively solved by accurately estimating Eq. (1). Consider for instance the following real-world decision-making task, inspired by the `solar` dataset (Godahewa et al., 2021). Suppose that one manages a network of solar power plants and that, one day in advance, needs to decide (i) which plants to shut down for maintenance, and (ii) how much electricity to commit to selling at every hour. Any shortage in production would result in a hefty fine, which implies that just predicting the expectation of Eq. (1) is not sufficient for the task. Consider the following formalization of the problem:

$$\begin{aligned} \max_{a_i,s_t} \quad & \mathop{\mathbb{E}}_{p\sim P}\left[\sum_t \left(s_t - \lambda \max\left(0, s_t - \sum_i a_i p_{it}\right)\right)\right], \\ \text{s.t.} \quad & \sum_i a_i \leq M, \;\; a_i \in \{0,1\}, \;\; s_t \geq 0, \end{aligned} \tag{2}$$

where $a_i = 1$ indicates that plant $i \in \{1,...,N\}$ is active (with at most M active at once), s_t is our production commitment for period $t \in \{1,...,T\}$, P_{it} is the distribution of electricity produced by plant i during period t (to be forecasted), P represents the multivariate distribution for all plants and periods, and $\lambda > 1$ is the penalty factor for not delivering the promised electricity.

Given a perfect estimate of Eq. (1), i.e., the future power production for each station, one could optimally solve this problem. However, in practice, this distribution is unknown and must be estimated using imperfect forecasting models; some imperfections could lead to critically bad solutions. For instance, if the forecast does not capture the statistical associations between stations, one could decide to only activate stations in one geographical area, making the total production vulnerable to local weather events.

It is thus essential to develop methodological tools to holistically assess the quality of probabilistic forecasts, not limited to their marginals or expectations. However, it is not straightforward to carry out such an evaluation as it requires knowledge of both the forecast and ground-truth distributions. As the latter is, in practice, unknown, the community has turned to *proper scoring rules*, i.e., evaluation criteria that only require samples/observations from these distributions.

2.3. Proper Scoring Rules

Denote $\mathcal{D}^{gt}$ the ground-truth process with Eq. (1) as probability density function and $\mathcal{D}^f$ an arbitrary forecast distribution over the same domain. A *scoring rule* is a function $S(y,\mathcal{D}^f)$ that measures the loss incurred if event $y \sim \mathcal{D}^{gt}$ realizes under forecast $\mathcal{D}^f$, assessing, for instance, how unlikely event y is according to $\mathcal{D}^f$. An example of a scoring rule is the *Energy Score* (Gneiting & Raftery, 2007), defined as

$$\text{ES}(y,\mathcal{D}^f) = \mathop{\mathbb{E}}_{x\sim\mathcal{D}^f} \|y-x\|_2^p - \frac{1}{2} \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}^f \\ x'\sim\mathcal{D}^f}} \|x-x'\|_2^p, \tag{3}$$

with $\|\cdot\|_2$ the Euclidean norm and $p \in (0,2)$ a parameter.

In general a scoring rule S must obey some basic regularity conditions, such as being *proper*, i.e., given a ground-truth distribution $\mathcal{D}^{gt}$ for any forecast $\mathcal{D}^f$, we must have

$$\mathop{\mathbb{E}}_{y\sim\mathcal{D}^{gt}}[S(y,\mathcal{D}^{gt})] \leq \mathop{\mathbb{E}}_{y\sim\mathcal{D}^{gt}}[S(y,\mathcal{D}^f)]. \tag{4}$$

In other words, the scoring rule must achieve its minimum, in expectation over all realizations of $\mathcal{D}^{gt}$, for a forecast that perfectly matches the ground truth. Further, a scoring rule is said to be *strictly proper* if the minimum is unique.

While it is a necessary condition for a scoring rule to be minimal in the ground truth, it is important to note that being proper does not imply that a scoring rule will be able to detect any prediction flaw. Several failure modes have been pointed out in the literature (e.g., Hamill (2001); Gneiting et al. (2007)), but an understudied issue is the behavior of proper scoring rules in the finite-sample regime. Indeed, properness only holds in expectation, while in practice evaluation is conducted based on a few samples from $\mathcal{D}^{gt}$ (e.g., in a rolling window evaluation). In what follows, we raise concerns about the practical reliability of common proper scoring rules, showing that, in realistic experimental conditions, they fail to distinguish between the ground-truth distribution and forecasts with significant imperfections.

3. Related Works

Early concern with forecast evaluation measures was driven in large part by the first forecasting competitions, such as the influential M-competitions (Makridakis & Hibon, 1979; Makridakis et al., 1982; Makridakis & Hibon, 2000), which prior to M4 (Makridakis et al., 2018) only focused on univariate point accuracy measures. Such measures include the mean absolute, squared, and absolute percentage errors (Mahmoud, 1984). The mean absolute scaled error was also proposed by Hyndman & Koehler (2006) as an improved scale-invariant measure. Interval forecasts (Chatfield, 1993; 2001) are a mid-point between point and full probabilistic forecasts, and were part of the M4 and M5 competitions (Hewamalage et al., 2021; Makridakis et al., 2022).

Development of Scoring Rules for Stochastic Forecasts Given that point and interval forecasting accuracy metrics are incomplete for probabilistic assessments, the literature considered alternatives such as the Continuous Ranked Probability Score (CRPS) (Matheson & Winkler, 1976; Winkler, 1996), and introduced the Energy Score (ES) (Gneiting & Raftery, 2007; Gneiting et al., 2008) as a multivariate generalization of the CRPS. Gneiting & Ranjan (2011) studied weighting schemes for the CRPS aimed at improving its tail behavior. As limitations of these rules became better understood, new ones, such as the Variogram (Scheuerer & Hamill, 2015), which is popular in the weather forecasting literature, have emerged. Salinas et al. (2019) introduced the CRPS-Sum as a simple scoring rule for multivariate time series. In another direction, Ziel & Berk (2019) proposed to split the forecast distribution into its marginals and a copula, allowing the use of specialized scoring rules on each component.

Assessment of Proper Scoring Rules Although a scoring rule cannot give a total ranking of forecasts for all possible applications of the target quantities (Diebold et al., 1998), much interest has been given to the specific discrimination abilities of particular scoring rules. In a univariate setting, Hamill (2001) illustrated a case where the rank histogram is uniform, yet every probabilistic forecast is biased; Gneiting et al. (2007) studied several rules in light of a proposed tuning approach to multivariate forecasts. Bao et al. (2007) used the Kullback-Leibler information criterion to compare univariate probabilistic forecasting models. Pinson & Tastu (2013) explored how well the Energy Score can distinguish between two bivariate Gaussian distributions, and gave a theoretical bound on how it fares as the number of variables increases. Alexander et al. (2022) compared the Energy Score and the Variogram on multiple distributions built from real-world data. However, their work does not explore the impact of dimensionality and sampling size on robustness. The effectiveness of the CRPS-Sum has been studied by Koochali et al. (2022), who showed how it can fail to distinguish between naive and state-of-the-art forecasts on real-world data. In these studies, a systematic assessment of forecasting rules against known deviations in the finite-sample regime is lacking, preventing a clear understanding of their limitations. This is what the present work seeks to address.

4. Regions of Reliability

As exemplified in our motivating example (§ 2.2), the detection of certain discrepancies between forecasts and ground truth, e.g., adequately capturing the correlation structure, is of high practical importance. Nonetheless, the conditions in which proper scoring rules are used in practice do not always allow for the detection of such discrepancies. In what follows, we devise a methodology based on power analysis (e.g., Cohen (1992)) to assess the reliability of a scoring rule at evaluating forecasts. Aside from being a well-established framework for carrying out significance tests, power analysis has the advantage of providing power laws for the minimal size of the ground-truth sample sufficient for obtaining a certain discriminatory power, which would have to be assessed empirically otherwise.

4.1. Measuring Reliability via Power Analysis

To quantify the reliability of a scoring rule, we conduct a power analysis that tests whether the rule can discriminate an incorrect forecast distribution from the ground-truth distribution, given samples from both. That is, we measure the statistical power of the scoring rule on the task of rejecting the null hypothesis that the forecast and the ground truth are statistically indistinguishable at a chosen significance level. Note that, contrary to common applications of power analysis, we seek to assess the ability of a scoring rule to detect a known effect, rather than detecting a purported effect with a scoring rule that is known to be reliable.

Formally, consider a pair of ground-truth $\mathcal{D}^{gt}$ and forecast $\mathcal{D}^{f}$ distributions over d variables. We examine the following random variable, which denotes the gap between ground truth and forecast according to the scoring function S,

$$\Delta_m = S(y, X_m^f) - S(y, X_m^{gt}), \tag{5}$$

with $y \sim \mathcal{D}^{gt}$ a realization of the ground truth, and $X_m^{gt} = \{x_i \sim \mathcal{D}^{gt}\}_{i=1}^m$ and $X_m^f = \{x_i \sim \mathcal{D}^f\}_{i=1}^m$ random variables corresponding to samples of size m of the ground-truth and forecast distributions, respectively. We can empirically estimate the mean and variance of this random variable as

$$\mu_m = \mathbb{E}_{y, X_m^{gt}, X_m^f}[\Delta_m] \approx \frac{1}{K}\sum_{k=1}^{K} \delta_m^{(k)}, \tag{6}$$

$$\sigma_m^2 = \mathrm{Var}_{y, X_m^{gt}, X_m^f}[\Delta_m] \approx \frac{1}{K-1}\sum_{k=1}^{K} \left(\delta_m^{(k)} - \mu_m\right)^2, \tag{7}$$

where $\delta_m^{(k)} \sim \Delta_m$, which we measure through K independent

trials. [2] Our null hypothesis corresponds to assuming that the difference Δ_m has mean $\mu_m = 0$, i.e., the forecast is indistinguishable from the ground truth.

Instead of making assumptions on the distribution of Δ_m, we leverage the central limit theorem[3] and consider that the average of n independent replications of Δ_m (i.e., $\frac{1}{n}\sum_{j=1}^{n}\Delta_m^{(j)}$) approximately follows the normal distribution $H_S(n,m) = N\left(\mu_m, \sigma_m^2/n\right)$, which we take to be the *distribution under the alternative hypothesis*. Similarly, we take $H_0(n,m) = N\left(0, \sigma_m^2/n\right)$ to be the corresponding *distribution under the null hypothesis*. As illustrated in Fig. 1, our power analysis boils down to studying how $H_0(n,m)$ and $H_S(n,m)$ overlap, i.e., the power of the scoring rule at setting them apart.

Concretely, we fix the level of significance to $\alpha = 5\%$, i.e., the false positive rate (shaded blue area in Fig. 1), and determine the corresponding critical value $t_\alpha \in \mathbb{R}^+$,

$$\mathbb{P}[H_0(n,m) \geq t_\alpha] = \alpha. \tag{8}$$

We then quantify the reliability of the scoring rule via its statistical power.

Definition 4.1 (Statistical power)**.** The statistical power of a scoring rule S is defined as its true positive rate given a significance level of α and is given by

$$\mathbb{P}[H_S(n,m) \geq t_\alpha] = 1 - \beta, \tag{9}$$

where β is the false negative rate (shaded red area in Fig. 1).

Remarks: Increasing the number of replications (n) leads to a reduction in the variance of the distributions of H_0 and H_S, resulting in decreased overlap and thus, increased power (up to the perfect power of 1 in the limit). In practice, this can be achieved to some extent by increasing the number of rolling windows used for evaluation in a time series setting (up to the limits of data availability). As for the sample size (m), small values might inflate the variance estimated in Eq. (7), resulting in increased overlap between H_0 and H_S and reduced power. Finally, note that the dimensionality of the distributions (d) may degrade the ability of some scoring rules to detect discrepancies (e.g., due to the curse of dimensionality).

4.2. Identifying Regions of Reliability

By performing the power analysis described above for a variety of conditions: number of variables (d), ground-truth sample size (n), and forecast sample size (m), we can isolate the set of conditions (region) under which a scoring rule S achieves a power of at least $1-\beta$ for some β of interest:

Definition 4.2 (Region of reliability)**.** We define the region of reliability of level $1-\beta$ for a scoring rule S, denoted by

[2] In practice, we set $K = 1000$.

[3] We investigate how accurate this approximation is in § C.4.

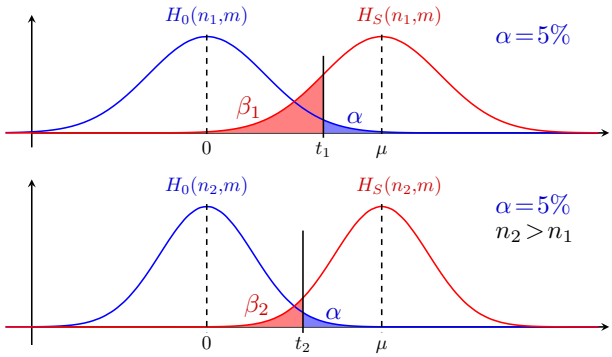

Figure 1. Schematic of our power analysis. It can be thought of as a classification problem, where the task is to distinguish between effects that are due to statistical noise (i.e., samples from the null hypothesis distribution H_0, in blue, or the negative class) and effects due to the discrepancy between ground truth and forecast as measured by the scoring rule S (i.e., samples from the alternative hypothesis H_S, in red, or the positive class). The critical value t_α for the classification task is set so that the false positive rate α equals 0.05. The statistical power of a test corresponds then to the true positive rate $1-\beta$ and varies depending on three factors: (1) The number of trials n; (2) The sample size m; (3) The studied scoring rule S. For instance, by increasing the number of trials from the top (n_1) to the bottom (n_2) plots, the same scoring rule S has better power, as the false negative rate decreases (from β_1 to β_2). On the same problem and for the same n and m, different scoring rules can have different power, depending on their ability to capture the mismatch between ground truth and forecast.

$\mathsf{RoR}_{(1-\beta)}$, as the subset of $dom(d) \times dom(n) \times dom(m)$, with $dom(\cdot)$ the domain of each factor, where S achieves at least $1-\beta$ statistical power as defined in Definition 4.1.

Note that the above methodology cannot directly be applied to real-world datasets since their ground-truth distributions are unknown, making it infeasible to sample n ground-truth trajectories and to produce forecasts that are qualitatively different from them. Hence, in what follows, we construct a synthetic benchmark that enables the controlled evaluation of scoring rules in a comprehensive set of experimental conditions where the ground-truth distribution is known, and we test the generalizability of our findings beyond these simulated cases on three temporal real datasets.

5. Benchmark experiments

We propose a benchmark consisting of a comprehensive array of test cases, each made up of a ground-truth and forecast distributions that differ in a chosen feature (e.g., a statistical moment) by a controlled amount (ε). These test cases are carefully selected to enable a systematic evaluation of scoring rules on specific discrepancies that may occur when building probabilistic forecasts for real-world tasks.

Table 1. Nomenclature of test cases where distributions vary in their marginal distributions. Each test case is defined by distributions with a specific kind of marginals, a subset of dimensions that change, a parameter to modify them (denoted ε for generality), and a shift direction. For instance, Exponential(All,$\mu\uparrow$) designates a test case where the ground-truth distribution consists of multiple independent Exponential distributions with mean 1, while the forecast distribution is the same but with mean $\mu=\varepsilon>1$ for all dimensions.

Marginal distribution	Dimension subset
Normal $N(\mu,\ \sigma^2)$, the Normal distribution with mean $\mu = 0$ and standard deviation $\sigma=1$.	**Single** Only the first dimension is modified in the forecast.
Exponential $\mathrm{Exp}(\lambda = 1/\mu)$, the Exponential distribution with mean $\mu=1$.	**All** All dimensions are modified identically in the forecast.
Skew Normal $\mathrm{Skew}(\mu=0,\sigma^2=1,\alpha)$, the Skew Normal distribution with mean equal to 0, variance equal to 1, and shape parameter α.	**Direction** $\uparrow$ The parameter is increased in the forecast. $\downarrow$ The parameter is decreased in the forecast.

Table 2. Nomenclature of test cases based on multivariate Gaussian distributions, with means set to 0 and variances to 1, but varying correlations. For each test case, we decide whether it is the ground truth or the forecast whose covariance matrix is not the identity, and which entries of said covariance matrix are set to ε. For instance, Full Cov(Extra) describes a test where the ground-truth distribution is a multivariate Gaussian distribution with 0-mean and identity covariance matrix, while the forecast distribution is a multivariate Gaussian distribution of 0-mean and covariance matrix equal to 1 on its diagonal and ε elsewhere.

Covariance matrix	Non-trivial covariance
Full Cov $\Sigma_{ab}=\varepsilon \quad \forall a\neq b$. All correlations are equal to some positive constant.	**Missing** The ground-truth covariance matrix is not the identity.
Checker Cov $\Sigma_{ab}=(-1)^{a+b}\varepsilon$ $\forall a\neq b$. All correlations alternate between $+\varepsilon$ and $-\varepsilon$.	**Extra** The Forecast distribution covariance matrix is not the identity.
Block Cov $\Sigma_{2h,2h+1}=\Sigma_{2h+1,2h}=\varepsilon$ $\forall h\in\{1,...,d/2\}$. The correlation matrix is block-diagonal with blocks of size 2-by-2.	

5.1. Proper Scoring Rules

We analyze five scoring rules that are common in the multivariate probabilistic forecasting literature: the Negative Log-Likelihood (NLL), the Continuous Ranked Probability Score (CRPS), the Energy Score (ES), the Variogram (VG), and the Dawid-Sebastiani score (DS). For the CRPS, we use its average over the dimensions[4] and we consider two variations, which we denote by CRPS-Q (quantiles) and CRPS-E (expectations), where CRPS-Q is the most commonly used. Similarly, for the ES, we consider the complete numerical approximation (ES-Full) and a faster approximation (ES-Partial). Detailed definitions of these rules and their respective parameters can be found in § A.

5.2. Test Cases and Tuning

We consider 19 test cases, or discrepancy types, categorized into (i) distributions that differ in their marginal distributions (detailed in Tab. 1), (ii) distributions that differ in their covariance structure (detailed in Tab. 2), and (iii) multivariate Gaussian mixture distributions where the forecast has one more or one fewer mixture component than the ground truth, denoted by *Mixture (Missing)* and *Mixture (Extra)*, respectively. An explicit definition of each test case can be found in § B.

Tuning In all cases, we tune the magnitude of the discrepancy between distributions (ε), based on the performance of the NLL scoring rule. More precisely, we fix ε such that the NLL has a statistical power of $1-\beta=0.8$ for a significance level of $\alpha=0.05$. This ensures that each test case is meaningful in that the magnitude of the effect is sufficient to be detected (most of the time) by the NLL. The scoring rules are thus evaluated w.r.t. their ability to serve as surrogates for the NLL in forecast evaluation. In § C.1 we report additional results not related to the NLL power.

[4]Taking the average CRPS is often used in practice, although a multivariate extension of CRPS has been proposed (Gneiting & Raftery, 2007).

5.3. Results

We start by emphasizing results of key interest by illustrating regions of reliability for specific scoring rules and test cases in Figs. 2 to 5. We then report a summary of all results in Fig. 6 and detailed results in § C.

Visualizing RoRs As regions of reliability are defined over three axes (number of variables d, ground-truth sample size n, and forecast sample size m), we plot the cross section corresponding to setting $n=30$ and show how the statistical power of a scoring rule varies with d and m using a heatmap. We choose $n=30$ since (i) it corresponds to a rolling window evaluation setting of realistic length, and (ii) the central limit theorem commonly holds around this sample size. Nonetheless, we note that insights on reliability for larger n can be inferred from such visualizations since the values of the heatmap can be interchangeably read as the statistical power at $n=30$ (higher is better) or as the minimal n required to achieve 80% power (lower is better). For ease of readability, we also plot the contour lines for $\mathsf{RoR}_{0.8}$ (solid), $\mathsf{RoR}_{0.5}$ (dashed), and $\mathsf{RoR}_{0.2}$ (dotted) computed based on a kernel-smoothed estimation of the measured statistical power.

Detection of Incorrect Mean Fig. 2 shows the statistical power of CRPS-Q, ES-Partial, and VG when the ground truth and the forecast means differ in a single dimension. When

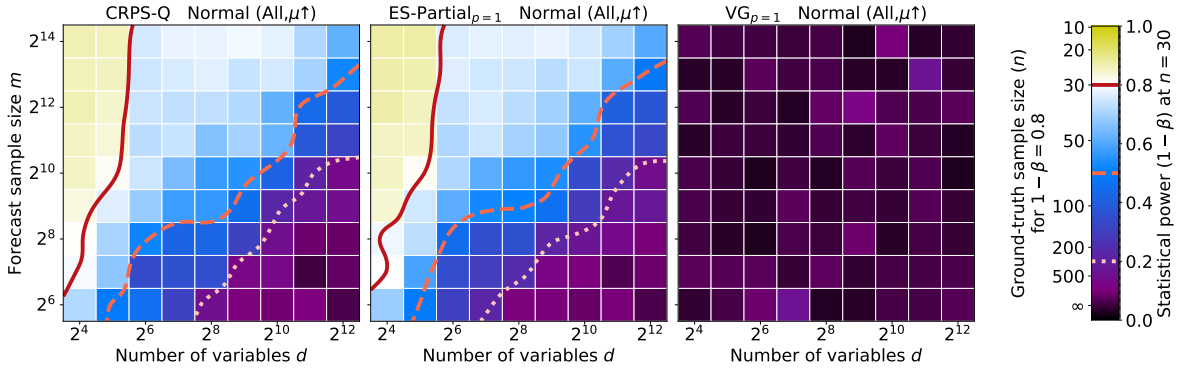

Figure 2. Statistical power of *CRPS-Q* (left), *ES-Partial* (middle) and *VG* (right) for detecting a mismatch in mean on all dimensions, depending on the problem dimensionality (d) and the forecast sample size (m).

$m > d$, the CRPS-Q and ES-Partial are able to detect such discrepancies: their $\mathrm{RoR}_{0.8}$, which indicates a power greater or equal to that of the NLL, spans $d \leq 2^5 \cap m \geq 2^{10}$. However, when $m < d$, a setting that is ubiquitous in real-world benchmark datasets (Godahewa et al., 2021), their power drops below 20% and we find that $n = 300$ ground-truth samples (e.g., rolling windows) would be needed for these scoring rules to perform on par with the NLL. In contrast, VG never succeeds at detecting the discrepancy, with empty $\mathrm{RoR}_{0.8}$ and $\mathrm{RoR}_{0.5}$, and a power below 20% even when $m \gg d$. We obtain similar results on the test cases with incorrect means for all dimensions (cf. § C).

Detection of Missing Correlations Fig. 3 shows the statistical power of ES-Partial, DS, and VG when the ground truth and forecast differ only in their correlation structure. DS is the only scoring rule with a non-empty, albeit very small, $\mathrm{RoR}_{0.8}$ ($d = 2^5, m = 2^{14}$). However, this scoring rule makes parametric assumptions[5] that match the underlying distributions, giving it a significant advantage. As for ES-Partial, its performance is far from that of the NLL, with a small $\mathrm{RoR}_{0.5}$ ($d = 2^4 \cap m \geq 2^{13}$) and a larger $\mathrm{RoR}_{0.2}$ ($d \leq 2^6 \cap m \geq 2^9$). VG performs comparably but, interestingly, it seems less affected by the forecast sample size (m), with a $\mathrm{RoR}_{0.2}$ spanning all values of m ($d \leq 2^7$). Nevertheless, we find that the ground-truth sample size (n) needed for ES-Partial and VG to perform on par with the NLL is of the order of hundreds or thousands, a setting that is highly impractical.

Detection of Incorrect Higher Moments Fig. 4 shows the power of CRPS-Q and ES-Partial when the distributions differ in higher moments, i.e., their mean and covariance are identical. We illustrate two cases: (i) differences in skewness (Fig. 4, left-middle) and (ii) the inclusion of an extra distribution mode in the forecast (Fig. 4, right). In all cases, the scoring rules completely underperform in comparison to the NLL. For differences in skewness, CRPS-Q is the only scoring rule that shows any statistical power, reaching a power of around 20% for very high sample sizes ($m \geq 2^{13}$) or very low numbers of dimensions ($d \leq 2^5$). As for ES-Partial, it reaches power values close to the false positive rate ($\alpha = 5\%$), indicating that its scores are essentially random. Finally, for the case where the forecast contains an extra mode, ES-Partial is the only scoring rule that was found to achieve non-negligible power. The results indicate a small $\mathrm{RoR}_{0.2}$ spanning $d \leq 2^5 \cap n \geq 2^{10}$ and that an impractical n of the order of hundreds would be required to match NLL performance.

Extrapolating these Results Can these results be extrapolated beyond the ranges considered for d, m, n? Statistical power generally increases with m and decreases with d and this is clearly reflected in our analysis. However, this does not necessarily happen monotonically, as illustrated in Fig. 5. We provide an explanation for this in § C.1.

Summary of Results Finally, we summarize the results of our analysis in Fig. 6 by listing, for each scoring rule and test case, the maximal power (over m) averaged over all values of d. This measure gives insights into the performance of a scoring rule in our benchmark, regardless of the dimensionality. For instance, a value of 0.8 indicates that a scoring rule typically performs on par with the NLL, whereas values < 0.8 indicate that it underperforms. Here are a few notable observations:

- Most scoring rules are only capable of detecting errors in the marginal distributions, as revealed by their small power in the test cases that induce errors in covariance.
- For covariance-related test cases, DS shows the greatest power, followed by VG. Yet, they both significantly underperform the NLL. DS's stronger performance is likely due to its Gaussian parametric assumptions (see § A.5), a thesis supported by its failure in the mixture test cases.
- VG is complementary to the CRPS and ES in our settings. When VG achieves lower power, CRPS/ES generally achieves higher power and vice versa. The only exceptions are the mixture test cases, where all scoring rules fail.

[5]The Dawid-Sebastiani scoring rule approximates the multivariate Gaussian NLL (see § A.5), which matches the data.

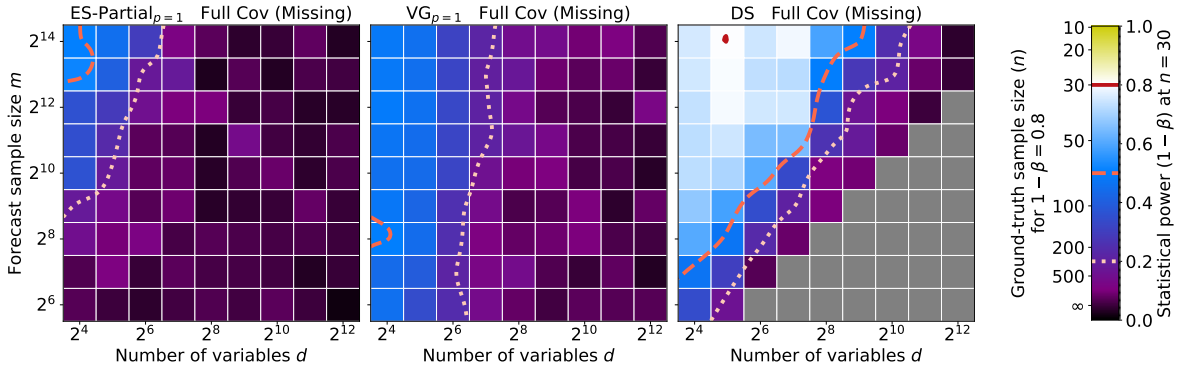

Figure 3. Statistical power of *ES-Partial* (left), *VG* (middle) and *DS* (right) at detecting that the forecast is missing the positive correlations between variables, depending on the problem dimensionality (d) and the forecast sample size (m). *DS* cannot be computed when $d \geq m$, so the corresponding area is greyed out.

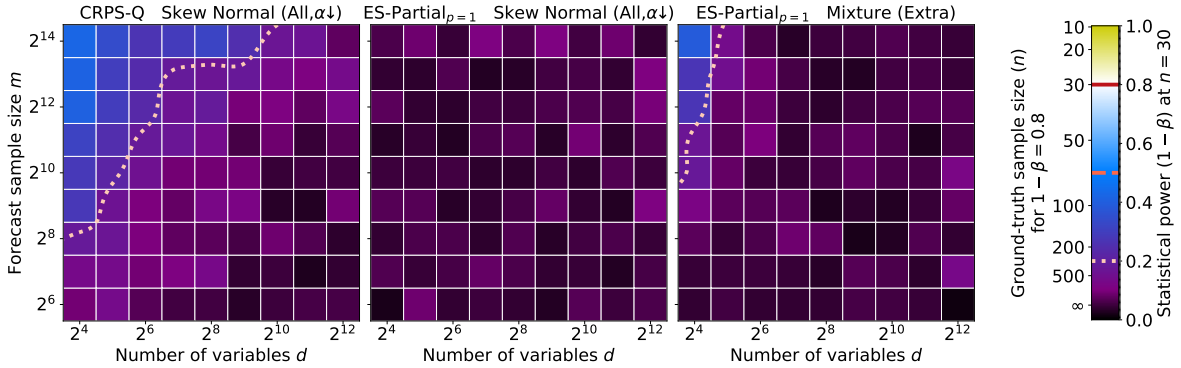

Figure 4. Statistical power of *CRPS-Q* (left) and *ES-Partial* (middle) at detecting that the forecast is not capturing the skewness of the ground-truth distribution, and of *ES-Partial* (right) at detecting that the forecast is predicting an additional mode, all depending on the problem dimensionality (d) and the forecast sample size (m).

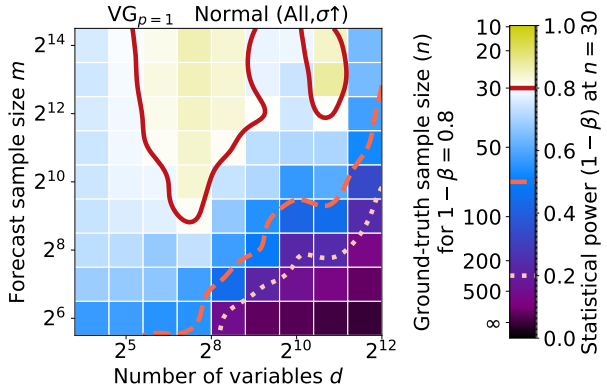

Figure 5. Non-Monotonicity of statistical power of *VG* at correctly detecting that a forecast marginal variance is lower than the ground-truth one on a single dimension.

- The various implementations of CRPS and ES perform comparably. This indicates that the computational gain of CRPS-Q and ES-Partial does not negatively affect their reliability.

Generalizability to Real Data The regions of reliability that we identify in the simulated benchmark provide valuable insights for practitioners. In fact, even if the distributions used in the benchmark do not match their data, observing a low-power result in one setting is indicative of a potential problem. This means that there exists at least one simple setting where the scoring rule fails to detect an error, indicating that one should take additional measures to ensure the validity of one's findings (see §7). As such, the regions of reliability calculated on simulated data serve as a necessary sanity check, i.e., the scoring rules that do not pass the tests proposed herein for a given setting should not be relied upon in practical settings. In the next section, we evaluate whether our findings also hold for real-world temporal data.

6. Application to Real Temporal Data

We now explore the generalizability of our findings to more realistic temporal data distributions in the context of our motivating example (see § 2). Since our methodology can only be applied when the ground-truth distribution

is known, we fit the TACTiS multivariate probabilistic forecasting model (Drouin et al., 2022) to the `solar-10min` dataset ($d = 9864$, $n = 30$, $m = 100$; details in § D) and consider the learned distribution to be the ground truth. We then produce forecast distributions with three kinds of errors, which are likely to affect the resulting profit, Eq. (2), namely: (i) breaking all correlations between the variables, which prevents one from strategically selecting power stations to put into maintenance (e.g., based on location), (ii) multiplying all variables by 1.05, which causes an overestimation of power production, resulting in penalties, and (iii) adding 0.05 to all variables, which has a similar effect on profit.[6] The TACTiS model was chosen because of the ease of removing all dependencies between variables when computing its NLL, as it is a copula-based model.

Given that this is a real-world setting, we cannot tune the difficulty of the task consisting of detecting broken correlations. For the multiplicative and additive errors, we picked their parameters such that their respective decreases in profit are in the same order of magnitude as the one for broken correlations. These profit decreases are computed with the model presented in Eq. (2), with $M = 50$, and $\lambda = 10$. This results in test cases that are somewhat easier than those studied in the benchmark, as reflected by the NLL achieving a power of 1.0 (instead of 0.8) in each case. Nonetheless, all the other scoring rules fail in at least one case:

- **Breaking Correlations:** Profit decrease: 2.3%. As expected, the CRPS fails to capture this, achieving a power of 0.05 (see Fig. 21). ES and VG achieve a perfect power of 1.0, which is surprising given that $d \gg m$, but may be explained by the simplicity of the problem (see Fig. 9 for a study of statistical power as a function of the difficulty of the problem).
- **Multiplying by a Constant:** Profit decrease: 3.0%. The CRPS, ES, and VG all detect the discrepancy, with powers of 0.85, 0.75 and 0.75, respectively. This result is not surprising given that these rules each showed significant power at detecting increases in mean in exponential distributions in Fig. 19, which is the closest analog in our benchmark (more details in § D).
- **Adding a Constant:** Profit decrease: 5.1%. The CRPS succeeds at detecting the error, with a power of 1.0. However, the ES performs poorly, with a power of 0.19 and VG completely fails, with a power of 0.05. The results for the CRPS and VG are in line with those in Fig. 11 for the test case where the marginal means were modified. As for the poor performance of ES, it may be explained by a large d, which places this rule outside of a region of reliability in our benchmark results.

While limited in scale, this experiment suggests that our benchmark results are informative, as most observations transfer to this more realistic setting. In § D.2, we expand the analysis to two additional real-world datasets and arrive at similar conclusions.

[6] For scale comparison, the average mean value is equal to 3.04 in the forecast used for this study.

	CRPS-Q	CRPS-E	ES-Full	ES-Partial	VG	DS
Normal (Single,$\mu\uparrow$)	0.74	0.75	0.76	0.73	0.43	0.37
Normal (All,$\mu\uparrow$)	0.77	0.84	0.83	0.76	0.10	0.38
Normal (Single,$\sigma\downarrow$)	0.40	0.40	0.11	0.10	0.70	0.30
Normal (Single,$\sigma\uparrow$)	0.67	0.72	0.28	0.25	0.66	0.53
Normal (All,$\sigma\downarrow$)	0.45	0.57	0.61	0.50	0.78	0.31
Normal (All,$\sigma\uparrow$)	0.52	0.52	0.62	0.56	0.80	0.64
Exponential (Single,$\mu\downarrow$)	0.69	0.72	0.75	0.68	0.34	0.29
Exponential (Single,$\mu\uparrow$)	0.71	0.73	0.63	0.62	0.47	0.43
Exponential (All,$\mu\downarrow$)	0.74	0.77	0.78	0.73	0.69	0.35
Exponential (All,$\mu\uparrow$)	0.70	0.70	0.78	0.78	0.69	0.81
Skew Normal (All,$\alpha\downarrow$)	0.26	0.33	0.09	0.09	0.08	0.07
Full Cov (Missing)	0.08	0.09	0.25	0.21	0.24	0.51
Full Cov (Extra)	0.08	0.09	0.18	0.15	0.11	0.29
Checker Cov (Missing)	0.08	0.09	0.27	0.20	0.28	0.52
Checker Cov (Extra)	0.09	0.09	0.18	0.16	0.32	0.31
Block Cov (Missing)	0.08	0.08	0.11	0.11	0.21	0.38
Block Cov (Extra)	0.08	0.08	0.12	0.11	0.22	0.42
Mixture (Missing)	0.09	0.09	0.10	0.09	0.08	0.07
Mixture (Extra)	0.09	0.09	0.13	0.13	0.09	0.08

Figure 6. Summary of results: maximal statistical power $(1 - \beta)$ over the sample size m, averaged over the number of variables d.

7. Discussion

This work studied the reliability of common proper scoring rules at detecting a variety of errors of practical significance in multivariate probabilistic time-series forecasting. For this purpose, we introduced the notion of regions of reliability, i.e., the experimental conditions under which a proper scoring rule can reliably be used for evaluation, and devised a methodology to identify them. Our proposed methodology consists in assessing the statistical power of a scoring rule (§ 4) on a comprehensive benchmark (§ 5) of test cases that were carefully selected to be of practical relevance. All test cases were tuned with respect to the Negative Log-Likelihood in order to assess whether existing scoring rules could serve as reliable substitutes for it. Our results paint a clear picture: although they are theoretically grounded in the asymptotic regime, none of the considered proper scoring rules is a good surrogate for the NLL, across all test cases, in the small-sample regimes that are common in practice. What is more, none of these scoring rules is able to reliably detect errors in modeling the statistical dependencies across variables, which is a fundamental goal of *multivariate* forecasting.

What are the implications of these findings for the time-series community? A review of the recent literature (e.g., Salinas

et al. (2019); Rasul et al. (2021b); de Bézenac et al. (2020); Rasul et al. (2021a); Tashiro et al. (2021); Nguyen & Quanz (2021); Tang & Matteson (2021); Drouin et al. (2022)) reveals that most contributions are benchmarked by performing rolling-window evaluation on the following datasets: `electricity` ($d = 8880$, $n = 7$), `exchange` ($d = 240$, $n = 5$), `solar` ($d = 3288$, $n = 7$), `taxi` ($d = 29136$, $n = 57$), `traffic` ($d = 23112$, $n = 7$), `wikipedia` ($d = 60000$, $n = 5$), with $m \in [100, 1000]$ depending on the experimental choices of each study. Notice that, in all cases, the sample size (m) and the number of evaluation windows (n) are small in comparison to the dimensionality of the data (d). Our results suggest that, in these regimes, the assessed scoring rules are generally unreliable, i.e., they detect fewer or none of the errors that the NLL would be able to capture. This observation is alarming, as it puts into question the reliability of how progress is currently measured in the field.

Practical Recommendations We emphasize the need to evaluate models in settings where the number of ground truth samples (n) and the number of forecast samples (m) are significantly larger than current standard practice in the literature. As described in § 4.1, both of these quantities have a direct effect on the statistical power of scoring rules. On the one hand, increasing n always improves the statistical power (see Eq. (7)), but is not always a practical solution as it requires collecting more data or reducing the training set in favor of the test set. Nonetheless, the large size of the aforementioned datasets should be amenable to such a change. For instance, the `electricity` dataset contains observations at 5790 time points, a meager 7 of which are typically used for evaluation. On the other hand, increasing m is only guaranteed to increase power up to an asymptote that is not necessarily $\beta = 0$, but it is a viable option that comes at a reasonable computational cost. Therefore, we recommend taking action on both quantities. Furthermore, when planning an experiment on real-world data, we recommend that practitioners query our results, with the appropriate values for d, m, and n, and look at the power for the kinds of errors that would be detrimental to their downstream task. If a scoring rule has low power in our simulated benchmark, they should proceed with caution, as there exists at least one data distribution where the rule is unreliable. Finally, we align with the common recommendation of complementing evaluation with other diagnostic tools, such as inspecting calibration and sharpness (Gneiting et al., 2007) and plotting correlations (Drouin et al., 2022).

Limitations and Future Work Our results are of empirical nature. As such, we cannot guarantee their validity beyond the considered domains (especially for d and m), which are constrained for computational reasons. This is particularly true in light of the non-monotonic behavior of the statistical power on certain test cases, as reported for instance in Fig. 5. Deriving finite-sample theoretical guarantees would be key to characterizing the domains not studied in the current work and is a promising direction for future work. Overall, our analysis highlights the need for developing new scoring rules, with better finite-sample behavior. By providing a simulated benchmark of diverse test cases with known statistical properties, our work enables researchers to assess and compare the performance of different scoring rules in a concrete and standardized way. Our hope is that our work will stimulate the development of new scoring rules, and more extensive benchmarks, to better capture different aspects of prediction quality, and ultimately enable objective assessment of progress in the field.[7] As some scoring rules appear complementary (e.g., CRPS + VG, see § 5), a promising direction for future work would be to use our benchmark to learn combinations of scoring rules that achieve high power across a variety of settings.

Broader Impact Our research examines the reliability of scoring rules currently used in the evaluation of multivariate time series forecasting models. Our findings highlight the need for the scientific community to reassess how progress is measured in this field. By raising awareness of this issue, our research should encourage scientists to supplement their model evaluation with additional measures of correctness. This will stimulate the development of better evaluation protocols, leading to more accurate detection and tracking of innovation in the field. Ultimately, our research should lead to better screening of the shortcomings of forecasting models used in practical applications of machine learning, which have become increasingly ubiquitous in our daily lives. While there is a potential for this work to delay the deployment of innovative solutions due to concerns over forecast accuracy, we believe that the positive benefits to science will ultimately outweigh potential short-term negatives.

References

Alexander, C., Coulon, M., Han, Y., and Meng, X. Evaluating the discrimination ability of proper multi-variate scoring rules. *Annals of Operations Research*, 2022. doi: 10.1007/s10479-022-04611-9.

Bao, Y., Lee, T.-H., and Saltoğlu, B. Comparing density forecast models. *Journal of Forecasting*, 26(3):203–225, 2007.

Cai, Z. Regression quantiles for time series. *Econometric Theory*, 18(1):169–192, 2002. ISSN 02664666, 14694360. URL `http://www.jstor.org/stable/3533031`.

Chatfield, C. Calculating interval forecasts. *Journal of Business & Economic Statistics*, 11(2):121–135, 1993.

Chatfield, C. Prediction intervals for time-series forecasting. In Armstrong, J. S. (ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*, pp.

[7] Benchmark data and supporting code are available at `https://github.com/ServiceNow/regions-of-reliability`.

475–494, Boston, MA, 2001. Springer US. doi: 10.1007/978-0-306-47630-3_21.

Cohen, J. Statistical power analysis. *Current directions in psychological science*, 1992.

de Bézenac, E., Rangapuram, S. S., Benidis, K., Bohlke-Schneider, M., Kurle, R., Stella, L., Hasson, H., Gallinari, P., and Januschowski, T. Normalizing Kalman filters for multivariate time series analysis. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 2995–3007. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1f47cef5e38c952f94c5d61726027439-Paper.pdf.

Diebold, F. X., Gunther, T. A., and Tay, A. S. Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39(4):863, 1998.

Drouin, A., Marcotte, E., and Chapados, N. TACTiS: Transformer-attentional copulas for time series. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5447–5493. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/drouin22a.html.

Gneiting, T. and Katzfuss, M. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1: 125–151, 2014.

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

Gneiting, T. and Ranjan, R. Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, 29(3): 411–422, 2011.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69 (2):243–268, 2007.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., and Johnson, N. A. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test*, 17:211–235, 2008.

Godahewa, R., Bergmeir, C., Webb, G. I., Hyndman, R. J., and Montero-Manso, P. Monash time series forecasting archive. In *Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. forthcoming.

Hamill, T. M. Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 2001.

Heizer, J., Render, B., and Munson, C. *Operations Management: Sustainability and Supply Chain Management*. Pearson, 14th edition, 2023.

Hewamalage, H., Montero-Manso, P., Bergmeir, C., and Hyndman, R. J. A look at the evaluation setup of the M5 forecasting competition, 2021. URL https://arxiv.org/abs/2108.03588.

Hyndman, R. J. and Koehler, A. B. Another look at measures of forecast accuracy. *International Journal of forecasting*, 22(4):679–688, 2006.

Koochali, A., Schichtel, P., Dengel, A., and Ahmed, S. Random noise vs. state-of-the-art probabilistic forecasting methods: A case study on CRPS-Sum discrimination ability. *Applied Sciences*, 12(10):5104, 2022.

Mahmoud, E. Accuracy in forecasting: A survey. *Journal of Forecasting*, 3(2):139–159, 1984.

Makridakis, S. and Hibon, M. Accuracy of forecasting: An empirical investigation. *Journal of the Royal Statistical Society. Series A (General)*, 142(2):97–145, 1979.

Makridakis, S. and Hibon, M. The M3-competition: results, conclusions and implications. *International Journal of forecasting*, 16(4):451–476, 2000.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., and Winkler, R. The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, 1(2):111–153, 1982.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. The M4 competition: Results, findings, conclusion and way forward. *International Journal of forecasting*, 34(4): 802–808, 2018.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., Chen, Z., Gaba, A., Tsetlin, I., and Winkler, R. L. The M5 uncertainty competition: Results, findings and conclusions. *International Journal of Forecasting*, 38 (4):1365–1385, 2022. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2021.10.009. URL https://www.sciencedirect.com/science/article/pii/S0169207021001722. Special Issue: M5 competition.

Matheson, J. E. and Winkler, R. L. Scoring rules for continuous probability distributions. *Management science*, 22(10):1087–1096, 1976.

Neyman, J. and Pearson, E. S. On the problem of the most efficient tests of statistical hypotheses. *Philosophical*

Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 231(694-706):289–337, 1933.

Nguyen, N. and Quanz, B. Temporal latent auto-encoder: A method for probabilistic multivariate time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9117–9125, 2021.

Peterson, M. *An introduction to decision theory*. Cambridge University Press, 2017.

Pinson, P. and Tastu, J. Discrimination ability of the energy score. Technical report, Technical University of Denmark, DTU Compute Technical Report-2013 No. 15, 2013.

Rasul, K., Seward, C., Schuster, I., and Vollgraf, R. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8857–8868. PMLR, 18–24 Jul 2021a. URL https://proceedings.mlr.press/v139/rasul21a.html.

Rasul, K., Sheikh, A.-S., Schuster, I., Bergmann, U. M., and Vollgraf, R. Multivariate probabilistic time series forecasting via conditioned normalizing flows. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=WiGQBFuVRv.

Salinas, D., Bohlke-Schneider, M., Callot, L., Medico, R., and Gasthaus, J. High-dimensional multivariate forecasting with low-rank Gaussian copula processes. *Advances in neural information processing systems*, 32, 2019.

Scheuerer, M. and Hamill, T. M. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4):1321–1334, 2015.

Tang, B. and Matteson, D. S. Probabilistic transformer for time series analysis. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 23592–23608. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/c68bd9055776bf38d8fc43c0ed283678-Paper.pdf.

Tashiro, Y., Song, J., Song, Y., and Ermon, S. CSDI: Conditional score-based diffusion models for probabilistic time series imputation. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 24804–24816. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/cfe8504bda37b575c70ee1a8276f3486-Paper.pdf.

West, M. and Harrison, J. *Bayesian Forecasting and Dynamic Models*. Springer, second edition, 1997.

Winkler, R. L. Scoring rules and the evaluation of probabilities. *Test*, 5(1):1–60, 1996.

Ziel, F. and Berk, K. Multivariate forecasting evaluation: On sensitive and strictly proper scoring rules. *arXiv*, pp. 1910.07325v1, 2019.

A. Proper Scoring Rules

A.1. Continuous Ranked Probability Score

The Continuous Ranked Probability Score (CRPS) (Matheson & Winkler, 1976) can be written in multiple ways. In particular, it can be written as a comparison between the forecast distribution cumulative density function $\Phi_{\mathcal{D}^f}(x)$ and the realization $y \sim \mathcal{D}^{gt}$:

$$\text{CRPS}(y, \mathcal{D}^f) = \int_{-\infty}^{\infty} \left(\Phi_{\mathcal{D}^f}(x) - \mathbf{1}[x-y]\right)^2 dx, \tag{10}$$

where $\mathbf{1}[x]$ is the Heaviside function. It has been shown (Gneiting & Raftery, 2007) that CRPS can be rewritten in a form akin to the Energy Score:

$$\text{CRPS}(y, \mathcal{D}^f) = \mathop{\mathbb{E}}_{x \sim \mathcal{D}^f} |y - x| - \frac{1}{2} \mathop{\mathbb{E}}_{\substack{x \sim \mathcal{D}^f \\ x' \sim \mathcal{D}^f}} |x - x'|, \tag{11}$$

or using quantiles:

$$\text{CRPS}(y, \mathcal{D}^f) = 2 \int_{q \in [0,1]} \left(\mathbf{1}[\Phi^{-1}_{\mathcal{D}^f}(q) - y] - q\right)\left(\Phi^{-1}_{\mathcal{D}^f}(q) - y\right) dq. \tag{12}$$

Since the CRPS is only defined for univariate distribution, the CRPS of a multivariate distribution is taken as the average CRPS over all dimensions. While the univariate CRPS is a strictly proper scoring rule, the multivariate version is only proper, as it does not capture the correlations between dimensions.

In our numerical experiments, we use numerical approximations of Eqs. (11) and (12) to compute the CRPS. The CRPS-Q (quantiles) uses the quantiles from 0.05 to 0.95 in steps of 0.05:

$$\text{CRPS-Q}(y, \mathcal{D}^f) \approx \frac{1}{|Q|} \sum_{q \in Q} \left(\mathbf{1}[\Phi^{-1}_{\mathcal{D}^f}(q) - y] - q\right)\left(\Phi^{-1}_{\mathcal{D}^f}(q) - y\right), \tag{13}$$

while the CRPS-E (expectations) uses the m samples from $\mathcal{D}^f$ directly:

$$\text{CRPS-E}(y, \mathcal{D}^f) \approx \frac{1}{m} \sum_i |y - x_i| - \frac{1}{m(m-1)} \sum_{i<i'} |x_i - x_{i'}|. \tag{14}$$

Due to the double sum, CRPS-E has complexity $O(dm^2)$, while CRPS-Q has $O(dm\log m)$ since quantiles can be efficiently computed after sorting the samples according to their values.

A.2. Energy Score

The Energy Score (ES) (Gneiting & Raftery, 2007) can be considered a multivariate generalization of the CRPS. Given a parameter $0 < p < 2$ (commonly called β in the literature, which we changed to avoid confusion with the false negative rate of § 4.1), the ES is defined as:

(15)

where $\|z\|_2$ denotes the Euclidean norm. The ES is a strictly proper scoring rule.

We use two numerical approximations for the Energy Score:

$$\text{ES-Full}_p(y, \mathcal{D}^f) \approx \frac{1}{m} \sum_i \|y - x_i\|_2^p - \frac{1}{m(m-1)} \sum_{i<i'} \|x_i - x_{i'}\|_2^p, \tag{16}$$

which uses the full amount of data available in the sample, but has $O(dm^2)$ complexity; and

$$\text{ES-Partial}_p(y, \mathcal{D}^f) \approx \frac{1}{m} \sum_i \|y - x_i\|_2^p - \frac{1}{m} \sum_{i=1}^{m/2} \|x_i - x_{i+m/2}\|_2^p, \tag{17}$$

which uses each data point from the sample only once in the second sum, thus reducing the computing time to $O(dm)$.

A.3. Variogram

The Variogram (VG) (Scheuerer & Hamill, 2015), for a given parameter p, is computed as follows:

$$\mathrm{VG}_p(y,\mathcal{D}^f) = \sum_{a,b} \left(|y_a - y_b|^p - \mathop{\mathbb{E}}_{x\sim\mathcal{D}^f}[|x_a - x_b|^p] \right)^2, \tag{18}$$

where the sum is over all pairs of dimensions of the problem, indexed by a and b, resulting in a $O(d^2m)$ complexity. VG is a proper scoring rule but is not strictly proper since it is invariant to translations in the forecast.

A.4. Negative Log-Likelihood

Given the Probability Density Function (PDF) $p^f_{\mathcal{D}}(x)$ of the forecast $\mathcal{D}^f$, the negative log-likelihood (NLL) is defined as:

$$\mathrm{NLL}(y,\mathcal{D}^f) = -\log p^f_{\mathcal{D}}(y). \tag{19}$$

While the negative log-likelihood is strictly proper and has many other theoretical properties, it cannot be straight-forwardly estimated from a finite sample of $\mathcal{D}^f$, so it often requires access to the PDF to be computed.

A.5. Dawid-Sebastiani

The Dawid-Sebastiani score is computed from the first two moments of $\mathcal{D}^f$, its mean $\mu^f_{\mathcal{D}}$ and its covariance matrix $\Sigma^f_{\mathcal{D}}$, as follows

$$\mathrm{DS}(y,\mathcal{D}^f) = \log\left|\det\Sigma^f_{\mathcal{D}}\right| + (y-\mu^f_{\mathcal{D}})^T\Sigma^{-1}_{\mathcal{D}^f}(y-\mu^f_{\mathcal{D}}). \tag{20}$$

The Dawid-Sebastiani score is very close to the negative log-likelihood for multivariate Gaussian distributions. Unlike the log-likelihood, it can be computed from a finite sample of $\mathcal{D}^f$ by using said sample to estimate $\mu^f_{\mathcal{D}}$ and $\Sigma^f_{\mathcal{D}}$. However, if the rank of $\Sigma_{\mathcal{D}^f}$ is upper bounded by m ($m \leq d$), $\Sigma_{\mathcal{D}^f}$ is guaranteed to not be full rank, hence its inverse (the concentration matrix $\Sigma^{-1}_{\mathcal{D}^f}$) does not exist. Thus, the Dawid-Sebastiani score is only defined for $m > d$.

B. Perturbations Used in the Benchmark

In this section, we detail the test cases of our benchmark. Each test case consists of a couple of ground-truth $\mathcal{D}^{gt}$ and forecast $\mathcal{D}^f$ distributions, specifically conceived to test a particular quality of the scoring rules under scrutiny, for instance, the ability to detect errors in a statistical moment of a significant magnitude. All tests are parameterized by the dimensionality d (number of variables), and a scale parameter ε.

B.1. Incorrect Marginals

For all these distributions, each dimension of either $\mathcal{D}^f$ or $\mathcal{D}^{gt}$ is an independent variable. $\mathcal{D}^f$ or $\mathcal{D}^{gt}$ are thus completely characterized by their respective marginal distributions $\{\mathcal{D}^f_a\}^d_{a=1}$ and $\{\mathcal{D}^{gt}_a\}^d_{a=1}$.

Normal (Single, $\mu\uparrow$)

$$\begin{aligned} \mathcal{D}^{gt}_1 &\sim N(\varepsilon,1) \\ \mathcal{D}^{gt}_a &\sim N(0,1) \quad \forall a \neq 1 \\ \mathcal{D}^f_a &\sim N(0,1) \quad \forall a \end{aligned} \tag{21}$$

Normal (All, $\mu\uparrow$)

$$\begin{aligned} \mathcal{D}^{gt}_a &\sim N(\varepsilon,1) \quad \forall a \\ \mathcal{D}^f_a &\sim N(0,1) \quad \forall a \end{aligned} \tag{22}$$

Normal (Single, $\sigma\uparrow$) and Normal (Single, $\sigma\downarrow$)

$$\begin{aligned} \mathcal{D}^{gt}_1 &\sim N(0,\varepsilon^2) \\ \mathcal{D}^{gt}_a &\sim N(0,1) \quad \forall a \neq 1 \\ \mathcal{D}^f_a &\sim N(0,1) \quad \forall a \end{aligned} \tag{23}$$

Normal (All, $\sigma\uparrow$) and Normal (All, $\sigma\downarrow$)

$$\begin{aligned} \mathcal{D}_a^{gt} &\sim N(0,\varepsilon^2) \quad \forall a \\ \mathcal{D}_a^{f} &\sim N(0,1) \quad \forall a \end{aligned} \tag{24}$$

Exponential (Single, $\mu\uparrow$) and Exponential (Single, $\mu\downarrow$)

$$\begin{aligned} \mathcal{D}_1^{gt} &\sim \mathrm{Exp}(1/\varepsilon) \\ \mathcal{D}_a^{gt} &\sim \mathrm{Exp}(1) \quad \forall a \neq 1 \\ \mathcal{D}_a^{f} &\sim \mathrm{Exp}(1) \quad \forall a \end{aligned} \tag{25}$$

where $\mathrm{Exp}(1)(\lambda)$ is an exponential distribution with rate of change λ, and mean $1/\lambda$.

Exponential (All, $\mu\uparrow$) and Exponential (All, $\mu\downarrow$)

$$\begin{aligned} \mathcal{D}_a^{gt} &\sim \mathrm{Exp}(1/\varepsilon) \quad \forall a \\ \mathcal{D}_a^{f} &\sim \mathrm{Exp}(1) \quad \forall a \end{aligned} \tag{26}$$

Skew Normal (All, $\alpha\downarrow$)

$$\begin{aligned} \mathcal{D}_a^{gt} &\sim \mathrm{Skew}(\xi(\varepsilon),\omega(\varepsilon),\varepsilon) \quad \forall a \\ \mathcal{D}_a^{f} &\sim N(0,1) \quad \forall a \end{aligned} \tag{27}$$

where $\mathrm{Skew}(\xi,\omega,\alpha)$ is a skew normal distribution with location ξ, scale ω, and shape α as parameters. ξ and ω are chosen such that the resulting distribution has a mean of 0 and a variance of 1. Note that we recover the $N(0,1)$ distribution when $\varepsilon=0$.

B.2. Incorrect Correlations

For all these distributions, only the copulas of $\mathcal{D}^f$ and $\mathcal{D}^{gt}$ differ, so their marginals are always identical. For simplicity, we only selected multivariate normal distributions, due to the ease by which their marginals can be selected to always be $N(0,1)$. Therefore, all ground-truth distributions in this section are of the form $\mathcal{D}^{gt} \sim N(0,\Sigma_{\mathcal{D}^{gt}})$ and all forecast distributions are of the form $\mathcal{D}^f \sim N(0,\Sigma_{\mathcal{D}^f})$, for various covariance matrices $\Sigma_{\mathcal{D}^{gt}}$ and $\Sigma_{\mathcal{D}^f}$.

Full Cov (Missing)

$$\begin{aligned} \Sigma_{\mathcal{D}^{gt},aa} &= 1 \quad \forall a \\ \Sigma_{\mathcal{D}^{gt},ab} &= \varepsilon \quad \forall a \neq b \\ \Sigma_{\mathcal{D}^{f},aa} &= 1 \quad \forall a \\ \Sigma_{\mathcal{D}^{f},ab} &= 0 \quad \forall a \neq b \end{aligned} \tag{28}$$

Full Cov (Extra)

$$\begin{aligned} \Sigma_{\mathcal{D}^{gt},aa} &= 1 \quad \forall a \\ \Sigma_{\mathcal{D}^{gt},ab} &= 0 \quad \forall a \neq b \\ \Sigma_{\mathcal{D}^{f},aa} &= 1 \quad \forall a \\ \Sigma_{\mathcal{D}^{f},ab} &= \varepsilon \quad \forall a \neq b \end{aligned} \tag{29}$$

Checker Cov (Missing)

$$\begin{aligned} \Sigma_{\mathcal{D}^{gt},aa} &= 1 \quad \forall a \\ \Sigma_{\mathcal{D}^{gt},ab} &= (-1)^{a+b}\varepsilon \quad \forall a \neq b \\ \Sigma_{\mathcal{D}^{f},aa} &= 1 \quad \forall a \\ \Sigma_{\mathcal{D}^{f},ab} &= 0 \quad \forall a \neq b \end{aligned} \tag{30}$$

Checker Cov (Extra)

$$\begin{aligned} \Sigma_{\mathcal{D}^f,aa} &= 1 && \forall a \\ \Sigma_{\mathcal{D}^f,ab} &= 0 && \forall a \neq b \\ \Sigma_{\mathcal{D}^{gt},aa} &= 1 && \forall a \\ \Sigma_{\mathcal{D}^{gt},ab} &= (-1)^{a+b}\varepsilon && \forall a \neq b \end{aligned} \tag{31}$$

Block Cov (Missing)

$$\begin{aligned} \Sigma_{\mathcal{D}^{gt}} &= \begin{pmatrix} \bar{\Sigma}^{gt}_{\mathcal{D}} & & \\ & \ddots & \\ & & \bar{\Sigma}^{gt}_{\mathcal{D}} \end{pmatrix} \\ \bar{\Sigma}^{gt}_{\mathcal{D}} &= \begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{pmatrix} \\ \Sigma_{\mathcal{D}^f} &= I \end{aligned} \tag{32}$$

Block Cov (Extra)

$$\begin{aligned} \Sigma_{\mathcal{D}^{gt}} &= I \\ \Sigma_{\mathcal{D}^f} &= \begin{pmatrix} \bar{\Sigma}^{f}_{\mathcal{D}} & & \\ & \ddots & \\ & & \bar{\Sigma}^{f}_{\mathcal{D}} \end{pmatrix} \\ \bar{\Sigma}^{f}_{\mathcal{D}} &= \begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{pmatrix} \end{aligned} \tag{33}$$

B.3. Mixture of Distributions

Mixture (Missing)

$$\begin{aligned} \mathcal{D}^{gt} &\sim \text{Mixture}\left(\frac{N(\varepsilon,I)}{2} + \frac{N(-\varepsilon,I)}{2}\right) \\ \mathcal{D}^{f} &\sim N(0,I+\varepsilon^2) \end{aligned} \tag{34}$$

The mean and covariance matrix of $\mathcal{D}^f$ has been chosen to have the same mean and covariance as the mixture used for $\mathcal{D}^{gt}$.

Mixture (Extra)

$$\begin{aligned} \mathcal{D}^{gt} &\sim N(0,I+\varepsilon^2) \\ \mathcal{D}^{f} &\sim \text{Mixture}\left(\frac{N(\varepsilon,I)}{2} + \frac{N(-\varepsilon,I)}{2}\right) \end{aligned} \tag{35}$$

The mean and covariance matrix of $\mathcal{D}^{gt}$ has been chosen to have the same mean and covariance as the mixture used for $\mathcal{D}^f$.

B.4. Tuning Results

The tuned values of ε are listed in Tab. 3 for all test cases and dimensions under consideration in this paper. We computed $\mathbb{E}_{y,X^{gt}_m,X^f_m}[\Delta_m]$ and $\mathrm{Var}_{y,X^{gt}_m,X^f_m}[\Delta_m]$ analytically for the Gaussian test cases, which gave us a numerically precise function of the NLL statistical power in term of ε. Since this function is strictly monotonic in the ranges we are interested in, we used a bisection algorithm to quickly find the correct tuned values.

For the non-Gaussian test cases, $\mathbb{E}_{y,X^{gt}_m,X^f_m}[\Delta_m]$ and $\mathrm{Var}_{y,X^{gt}_m,X^f_m}[\Delta_m]$ were estimated numerically with a sample size of 10'000. The sampling was done with a fixed random number generator seed, to avoid numerical jitter which breaks the monotonicity condition which we relied upon in the bisection algorithm. The randomness inherent to this procedure explains why ε is not constant for Exponential (Single,$\mu\downarrow$) and Exponential (Single,$\mu\uparrow$) even though they should be if the tuning was done exactly.

Table 3. Values of the ε parameter from the tuning procedure which requires the NLL to have a statistical power $1-\beta$ equals to 80%, as described in § 5.2. The values of ε which make the forecast distribution identical to the ground-truth distribution are 0 for Normal μ, Skew Normal α, all covariance test cases, and the mixture test cases; and 1 for Normal σ and Exponential μ.

d	**16**	**32**	**64**	**128**	**256**	**512**	**1024**	**2048**	**4096**
Normal (Single,$\mu\uparrow$)	0.9079	0.9079	0.9079	0.9079	0.9079	0.9079	0.9079	0.9079	0.9079
Normal (All,$\mu\uparrow$)	0.2270	0.1605	0.1135	0.0802	0.0567	0.0401	0.0284	0.0201	0.0142
Normal (Single,$\sigma\downarrow$)	0.5799	0.5799	0.5799	0.5799	0.5799	0.5799	0.5799	0.5799	0.5799
Normal (Single,$\sigma\uparrow$)	2.4514	2.4514	2.4514	2.4514	2.4514	2.4514	2.4514	2.4514	2.4514
Normal (All,$\sigma\downarrow$)	0.8584	0.8963	0.9248	0.9458	0.9612	0.9723	0.9803	0.9860	0.9901
Normal (All,$\sigma\uparrow$)	1.1855	1.1254	1.0860	1.0596	1.0415	1.0291	1.0204	1.0144	1.0101
Exponential (Single,$\mu\downarrow$)	0.4481	0.4487	0.4447	0.4481	0.4538	0.4528	0.4463	0.4463	0.4493
Exponential (Single,$\mu\uparrow$)	3.0032	3.0395	2.9980	3.0000	3.0316	3.0303	3.0327	3.0497	3.0514
Exponential (All,$\mu\downarrow$)	0.8028	0.8539	0.8932	0.9233	0.9451	0.9609	0.9721	0.9800	0.9859
Exponential (All,$\mu\uparrow$)	1.2666	1.1778	1.1209	1.0838	1.0584	1.0411	1.0289	1.0202	1.0142
Skew Normal (All,$\alpha\downarrow$)	2.3987	1.8090	1.4738	1.2036	1.0149	0.8744	0.7532	0.6555	0.5748
Full Cov (Missing)	0.2055	0.1218	0.0680	0.0363	0.0188	0.0096	0.0048	0.0024	0.0012
Full Cov (Extra)	0.1268	0.0629	0.0312	0.0155	0.0077	0.0039	0.0019	0.0010	0.0005
Checker Cov (Missing)	0.2055	0.1218	0.0680	0.0363	0.0188	0.0096	0.0048	0.0024	0.0012
Checker Cov (Extra)	0.1268	0.0629	0.0312	0.0155	0.0077	0.0039	0.0019	0.0010	0.0005
Block Cov (Missing)	0.3058	0.2214	0.1585	0.1128	0.0800	0.0567	0.0401	0.0284	0.0201
Block Cov (Extra)	0.3201	0.2268	0.1605	0.1135	0.0802	0.0567	0.0401	0.0284	0.0201
Mixture (Missing)	0.5906	0.4151	0.2974	0.2083	0.1480	0.1053	0.0739	0.0519	0.0367
Mixture (Extra)	0.8020	0.5749	0.4052	0.2909	0.2040	0.1456	0.1032	0.0727	0.0516

C. Extensive Benchmark Results

Our results for all of our test cases and scoring rules are presented in Figs. 7 and 10 to 28. To help distinguish $\mathsf{RoR}_{0.8}$, $\mathsf{RoR}_{0.5}$, and $\mathsf{RoR}_{0.2}$, we added contour lines for $1-\beta=0.8$, $1-\beta=0.5$, and $1-\beta=0.2$. Since directly using the raw values would result in quite rough contour lines, we first smoothed our data using a Radial Basis Function interpolation (with the Thin Plate Spline function), with $\log_2 d$ and $\log_2 m$ as the independent variables.

C.1. Additional Remarks

Non-Monotonic Statistical Power While statistical power generally improves when increasing m, we notice that it is not necessarily monotonic over the number of dimensions of our tests. Consider the results of Fig. 5 as an example. VG's precision is highest around $d=2^7$, and is lower for both higher and lower dimensionality. We explain this unexpected behavior by the way we tuned the ground-truth and forecast distributions of our test cases, such that the negative log-likelihood has a statistical power of 80%. This non-monotonicity can be the result of the negative log-likelihood and the variogram having a different dependency on the dimensionality, making problems with a higher number of variables not necessarily harder than problems with a lower number of variables.

Asymmetry of Test Cases In most of our tests, a score is not in general equally powerful at detecting whether forecasts underestimate a feature than at detecting forecasts that overestimate the same feature. For instance, Fig. 7 reports the statistical power of VG for a forecast that either underestimates or overestimates the standard deviation on a single dimension. While this scoring rule has a decent statistical power almost everywhere (true positive rate above 50% and minimal n around 50) in the former case, it requires many more samples ($m \geq 2^4 d$) to reach the same power that it has in the latter case. However, the scoring rule continues to improve as the number of samples increases when the forecast is sharp, eventually achieving greater power than the negative log-likelihood ($1-\beta > 0.8$) for $m \geq 2^{13} \cap d \leq 2^7$.

C.2. Study for Varying Problem Complexity

The results presented so far are obtained on test cases tuned w.r.t. the NLL, i.e., on ground-truth and forecast distributions generated so that the NLL can distinguish them with 80% statistical power. The rationale was to ensure that (i) the synthetic discrep-

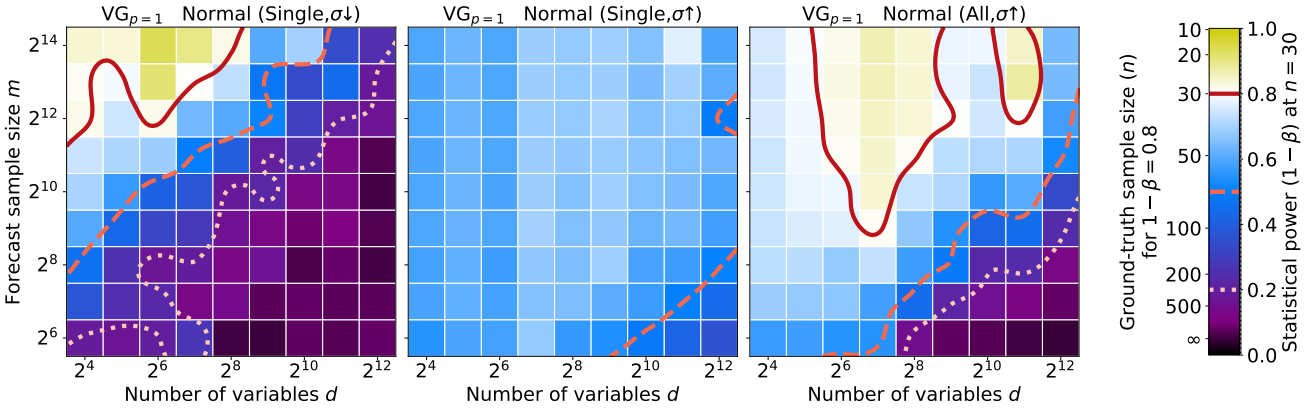

Figure 7. Statistical power of *VG* at correctly detecting that forecast's marginal variances are different from the ground-truth's ones: higher (left) or lower (middle) on a single dimension, or lower on all dimensions (right).

ancies could be captured at least by one scoring rule and (ii) the difficulty of the different test cases was comparable. Doing so we however limited our analysis to a single (albeit compelling) level of difficulty. In Fig. 9 we study the impact of this additional factor on the power of the scoring rules, for the test case where the ground-truth correlations are dropped (**Full Cov (Missing)**). We control the difficulty of the problem by varying the correlation parameter ε: the higher ε the bigger the discrepancy, hence the easier it is to distinguish the two distributions. We observe that, apart from the CRPS which is known to be insensitive to this particular type of discrepancy, all scoring rules converge to perfect statistical power. However, they show different convergence rates, with the DS and VG being significantly more reliable than the ES on the most difficult settings ($\varepsilon \leq 0.4$).

C.3. Sizes of the Regions of Reliability

As an alternative point of view to assess the quality of the various scoring rules, Fig. 8 shows how large each $\mathrm{RoR}_{0.5}$ are in our experiments. Very similar conclusions can be taken from it than from Fig. 6, since the complementarity between scoring rules, and which test cases cause issues for them, are still quite apparent. However, it reveals test cases where some scoring rules never reach the reasonable $1-\beta=0.5$ threshold.

C.4. Validity of normality assumption

In the statistical analysis of our results detailed in §4.1, we have opted to assume that taking an ensemble of $n=30$ independent replications of the gap between the ground truth and forecasting scores Δ_m is sufficient to use the Normal distribution in our computations. To test this hypothesis for a single experiment, we can use a Quantile-Quantile plot, where we compare the distribution of the ensemble average of n independent Δ_m, and theoretical Gaussian distribution of equal mean and standard deviation. Instead of randomly picking from our experiments, we select a few of them according to the excess kurtosis of Δ_m. Fig. 29 shows the Q-Q plot for 4 experiments, selected such that their excess kurtosis prior to ensembling are at quantiles 0.02, 0.05, 0.95, and 0.98. From these, we can see that only the experiments with the highest 2% of kurtosis suffer from a visible deviation between both distributions. This means that our normality assumption is reasonable for our experiments.

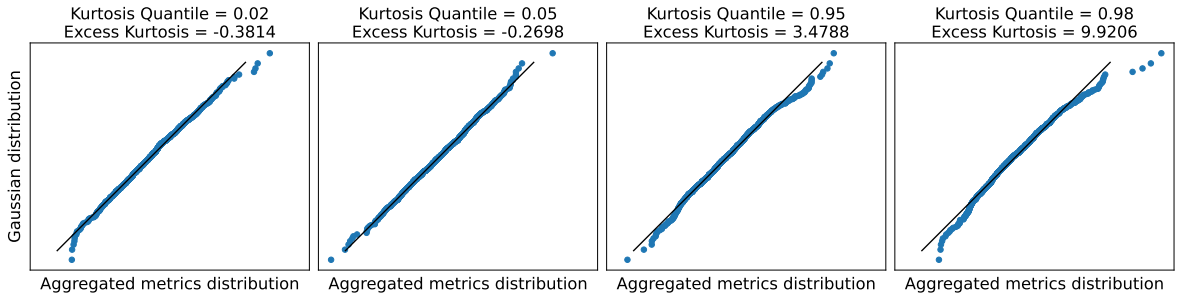

Figure 29. Q-Q plot comparing the Gaussian distribution and distribution we obtain by averaging $n=30$ independent Δ_m values from our experiments. The experiments shown are selected based on their rank when sorted by Δ_m excess kurtosis.

	CRPS-Q	CRPS-E	ES-Full	ES-Partial	VG	DS
Normal (Single,$\mu\uparrow$)	0.89	0.89	0.94	0.85	0.04	0.23
Normal (All,$\mu\uparrow$)	0.92	0.92	0.94	0.89	0.00	0.25
Normal (Single,$\sigma\downarrow$)	0.13	0.13	0.00	0.00	0.45	0.15
Normal (Single,$\sigma\uparrow$)	0.68	0.72	0.08	0.08	1.00	0.60
Normal (All,$\sigma\downarrow$)	0.21	0.30	0.30	0.25	1.00	0.17
Normal (All,$\sigma\uparrow$)	0.26	0.30	0.43	0.30	1.00	0.66
Exponential (Single,$\mu\downarrow$)	0.55	0.55	0.53	0.45	0.09	0.08
Exponential (Single,$\mu\uparrow$)	1.00	1.00	1.00	1.00	0.23	0.15
Exponential (All,$\mu\downarrow$)	0.91	0.89	0.98	0.94	0.98	0.28
Exponential (All,$\mu\uparrow$)	0.85	0.85	1.00	1.00	0.92	1.00
Skew Normal (All,$\alpha\downarrow$)	0.00	0.02	0.00	0.00	0.00	0.00
Full Cov (Missing)	0.00	0.00	0.06	0.04	0.02	0.45
Full Cov (Extra)	0.00	0.00	0.02	0.02	0.00	0.15
Checker Cov (Missing)	0.00	0.00	0.06	0.04	0.02	0.49
Checker Cov (Extra)	0.00	0.00	0.02	0.02	0.17	0.15
Block Cov (Missing)	0.00	0.00	0.00	0.00	0.08	0.25
Block Cov (Extra)	0.00	0.00	0.00	0.00	0.08	0.32
Mixture (Missing)	0.00	0.00	0.00	0.00	0.00	0.00
Mixture (Extra)	0.00	0.00	0.00	0.00	0.00	0.00

Figure 8. Summary of results: proportion of our experiments that are in $\mathtt{RoR}_{0.5}$ amongst those where the sample size is higher than the number of variables $m > d$.

C.5. Effects of Modifying the Experiments Parameters

In our main results, all of our statistical analysis has been done with a constant significance level $\alpha = 0.05$, and our experiment tuning with a constant statistical power of $1-\beta_{NLL} = 0.8$ for the NLL. Since these two choices are arbitrary, we show in this section that our conclusions still hold even if we repeat our experiments with other values for α and β.

Figs. 30 to 33 show the impact of varying α on the ability of CRPS-Q, ES-Partial$_p = 1$, and VG$_p = 1$ to distinguish a forecast distribution from a ground-truth distribution when they differ only by their covariance. From these, we can observe that the regions of high statistical power ($1-\beta \geq 0.5$) stay untouched by the increased in α. While increasing α does increase statistical power, this effect is perfectly balanced by the tuning procedure in those regions. On the flip side, regions of low statistical power saw an increase in said power. This is expected since for purely random scoring rules, the statistical power is equal to α.

Figs. 34 to 37 show the impact of varying $1-\beta_{NLL}$ on the ability of CRPS-Q, ES-Partial$_p = 1$, and VG$_p = 1$ to distinguish a forecast distribution from a ground-truth distribution when they differ only by their covariance. To reach an increased targeted statistical power for the NLL, the tuning requires the forecast and ground-truth distributions to be more dissimilar. As expected, these pairs of distributions are also easier to distinguish by the scoring rules under tests, leading to better reliability. However, the shapes of the various regions of reliability remain unchanged, so these results do not contradict our earlier conclusions.

D. Details on Real-Data Experiments

For our real-data experiments, we started with TACTiS (Drouin et al., 2022), a state-of-the-art timeseries forecasting model, trained on the 10-minute increment version of the `solar` dataset. We chose this model since it allows us to compute the NLL of multiple perturbations of the forecast, including making all variables independent or adding multiplication or additive biases. Using this model, we generated a multivariate stochastic forecast over a 12 hours period (thus 72 time steps), with a sample of size 100.

The impact of the perturbations on the revenues is computed with the model presented in Eq. (2), with $M = 50$, and $\lambda = 10$. This model can readily be converted to a Mixed Integer Programming one, and be solved exactly by a wide range of solvers.

The NLL is computed directly on each element from this sample, which we take as the ground truth sample. However, since it is numerically prohibitive to get independent samples for the other scoring rules, we have to reuse the ground truth sample

Table 4. Statistical power $1-\beta$ for the ability of various scoring rules to distinguish a ground-truth distribution based on the `solar` dataset and a perturbation of said distribution.

Scoring rule	Breaking correlations	Multiplying by 1.05	Adding 0.05
NLL	$1-5.6\times10^{-7728}$	$1-1.6\times10^{-43}$	$1-1.1\times10^{-200}$
CRPS-Q	0.05	0.37	$1-2.2\times10^{-28}$
ES-Partial$_{p=1}$	$1-1.7\times10^{-3}$	0.30	0.10
VG$_{p=1}$	0.99	0.30	0.05

Table 5. Statistical power $1-\beta$ for the ability of various scoring rules to distinguish a ground-truth distribution based on the `kdd-cup` dataset and a perturbation of said distribution.

Scoring rule	Breaking correlations	Multiplying by 1.05	Adding 0.5
NLL	$1-5.8\times10^{-301}$	$1-7.0\times10^{-40010}$	$1-3.0\times10^{-19042}$
CRPS-Q	0.05	0.61	0.96
ES-Partial$_{p=1}$	0.34	0.24	0.08
VG$_{p=1}$	0.30	0.24	0.05

to generate the forecasting sample. Thus, for each element from the ground truth sample, we use all other elements to generate the forecasting sample, after applying the perturbation. To break the correlations in a sample, we shuffle the values of each variable across the sample, which keeps the marginals intact. It should be noted that due to reusing the same sample to generate each forecasting sample, we ignore the extra variance in the scoring rules due to the sampling process, thus our estimates of $1-\beta$ for them are biased toward higher values. The statistical power for these experiments is shown in Tab. 4, but rewritten to put the emphasis on how close the NLL statistical power is to 1, which would only be possible for a perfect scoring rule with distributions with no overlap.

D.1. Additional Remarks

Multiplying by a Constant The closest analog to this test case in our benchmark is **Exponential (All, $\mu\uparrow$)** since multiplying an exponential variable by some factor is the same as multiplying its mean by the same factor. Alternatively we could have **considered** our test cases **Normal (All, $\mu\uparrow$)** and **Normal (All, $\sigma\uparrow$)** jointly.

D.2. Further Real-Data Experiments

To check whether the previous results are valid for other data than the `solar` dataset, we did the experiment with the `kdd-cup` (an air pollution dataset with hourly increments, 48 hours forecast, and $d=12960$) and `electricity` (an electricity dataset with hourly increments, 24 hours forecast, and $d=8880$) datasets. We kept the multiplication perturbation factor equal to the one used for the `solar` dataset (1.05). However, since the datasets do not share the same scale, we adjusted the additive perturbation term to roughly keep a constant ratio to the average value of the data. This gave an additive term of 0.5 for `kdd-cup` and 40 for `electricity`.

The statistical power for these two experiments is shown in Tabs. 5 and 6. The dominance of the NLL over the other scoring rules remains, although the three other scoring rules had very high statistical power for the `electricity` dataset when perturbing the data with either a multiplicative factor or an additive term. On the flip side, the energy score and the variogram lost much of the statistical power they showed for the correlation-breaking perturbation on the `solar` dataset.

D.3. Comparison with Previous Work

Our approach to testing scoring rules using real-world data is similar to the one in Alexander et al. (2022) in that we both use forecasting models trained on real-world data as the ground truth. Their differs from ours in how they create the erroneous forecasts in that they use alternative forecasting models to do so, while we apply explicit transformations. Both approaches are valid, in that theirs is closer to how forecast inaccuracies will behave while ours is easier to interpret.

Table 6. Statistical power $1-\beta$ for the ability of various scoring rules to distinguish a ground-truth distribution based on the `electricity` dataset and a perturbation of said distribution.

Scoring rule	Breaking correlations	Multiplying by 1.05	Adding 40
NLL	$1-3.3\times10^{-3274}$	$1-3.7\times10^{-11560}$	$1-1.7\times10^{-222}$
CRPS-Q	0.05	$1-2.5\times10^{-30}$	$1-7.3\times10^{-233}$
ES-Partial$_{p=1}$	0.35	$1-5.1\times10^{-3}$	$1-7.8\times10^{-3}$
VG$_{p=1}$	0.71	0.98	0.05

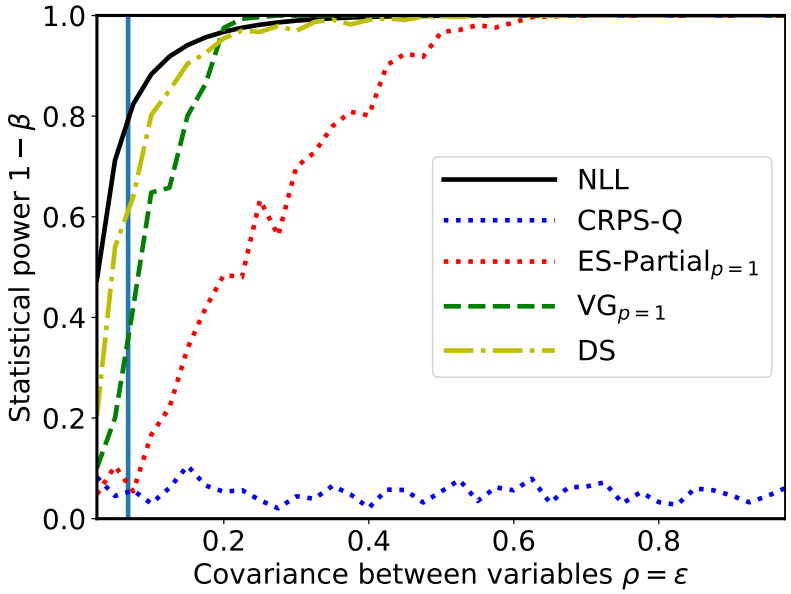

Figure 9. Statistical power of all scoring rules as a function of the simplicity of the problem (the higher ε the simpler the task) for $n=30$, $m=2^{12}$ and $d=16$. The test case corresponds to a forecast with all independent variables, while they are all positively correlated in the ground truth, for Normal distribution marginals. The vertical line marks the value of ε at which NLL has 80% power, corresponding to the main setting of our other studies.

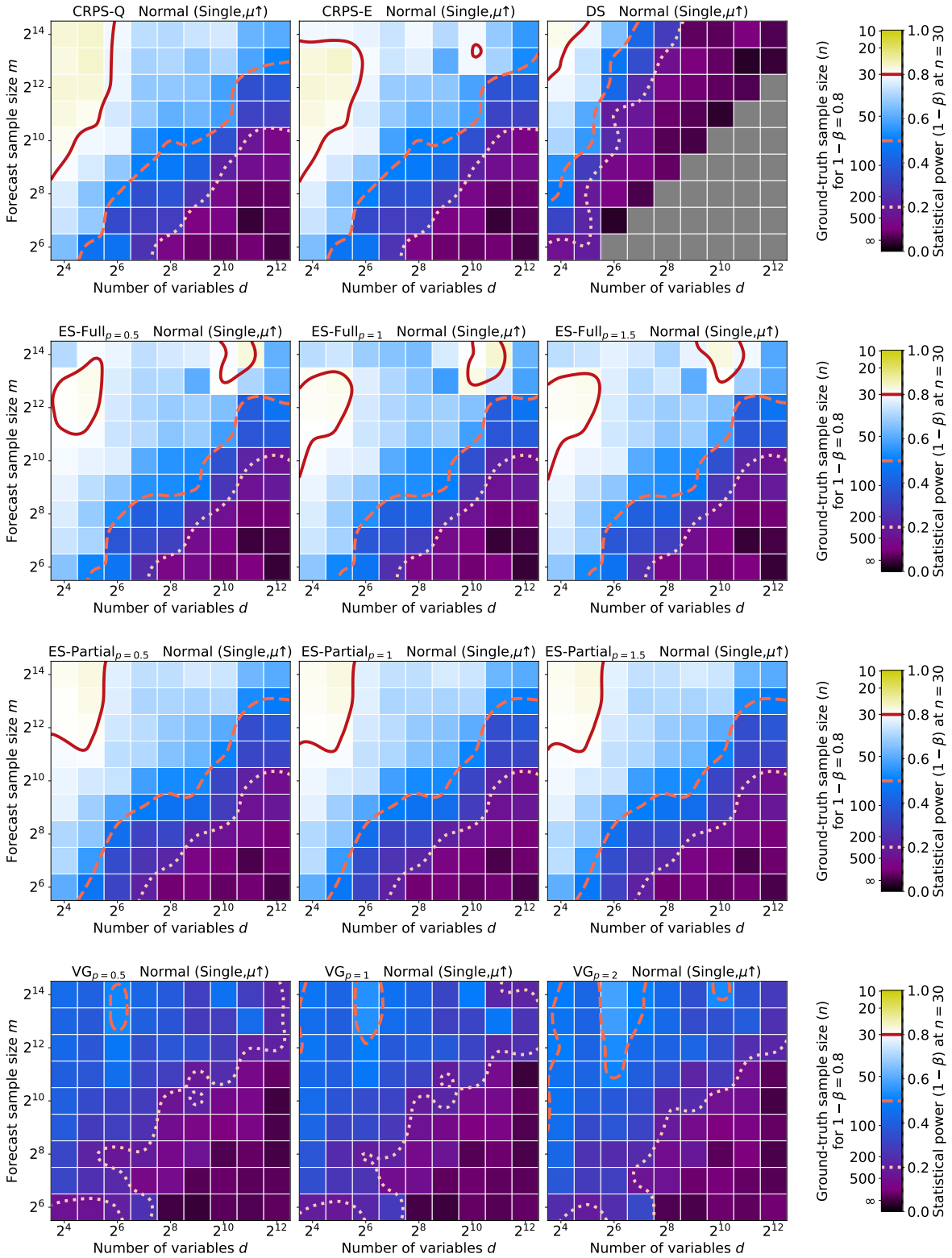

Figure 10. Statistical power of all scoring rules at correctly detecting that forecast's marginal mean is different than the ground-truth's one for a single dimension, for Normal distribution marginals. DS cannot be computed when $d \geq m$, so the corresponding area is greyed out.

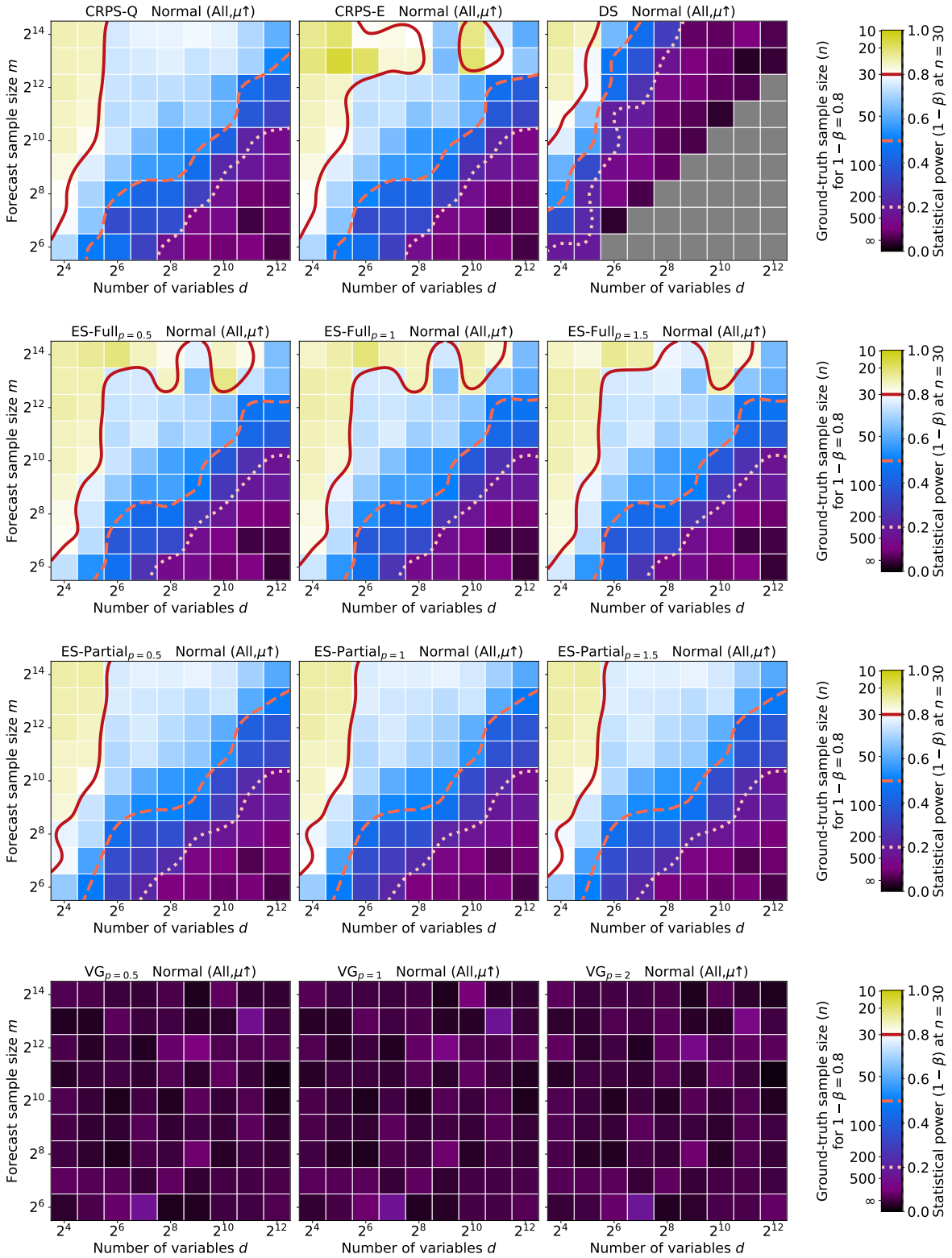

Figure 11. Statistical power of all scoring rules at correctly detecting that forecast's marginal means are different than the ground-truth's ones for all single dimensions, for Normal distribution marginals. DS cannot be computed when $d \geq m$, so the corresponding area is greyed out.

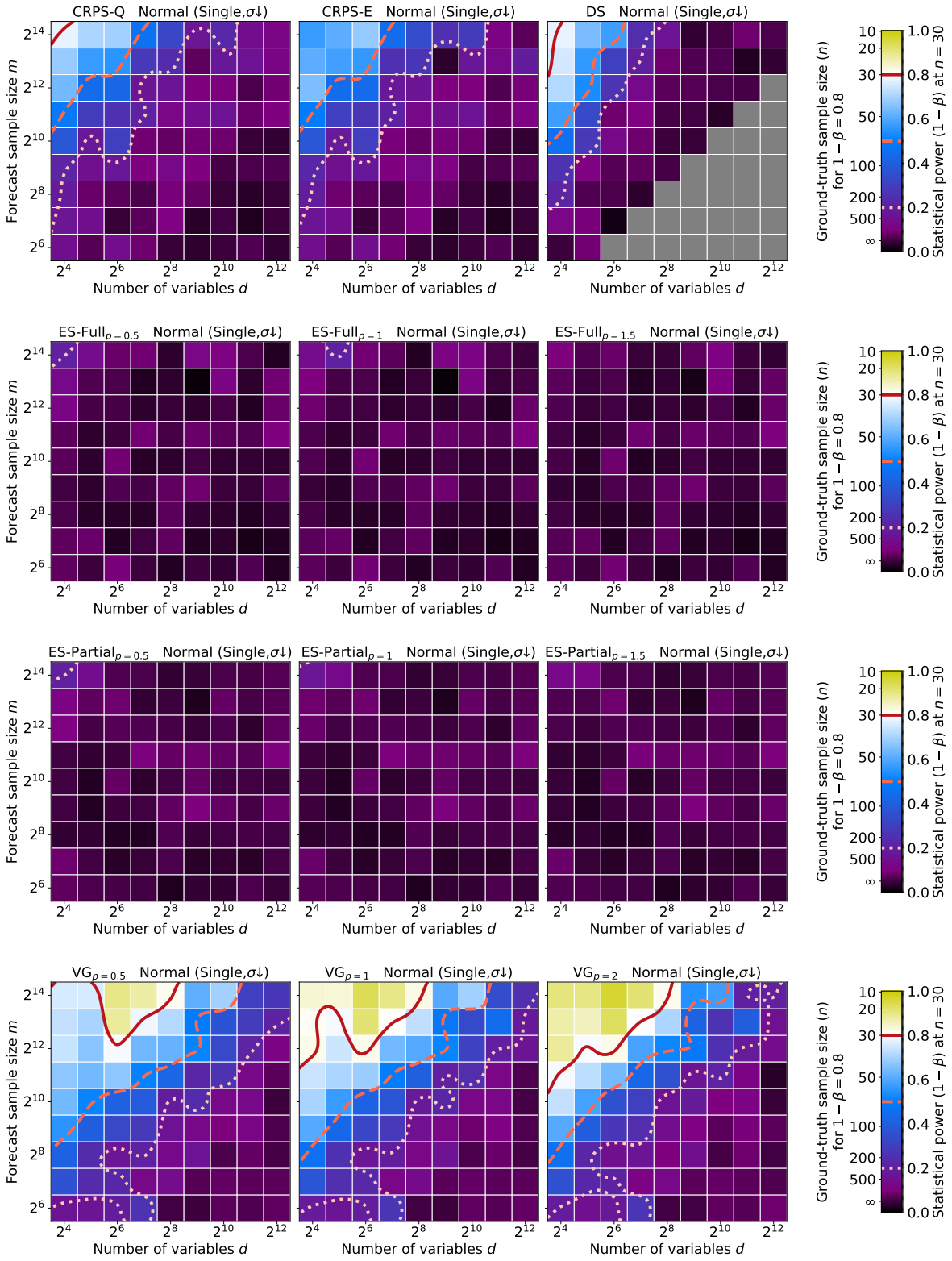

Figure 12. Statistical power of all scoring rules at correctly detecting that forecast's marginal standard deviation is lower than the ground-truth's one for a single dimension, for Normal distribution marginals. DS cannot be computed when $d \geq m$, so the corresponding area is greyed out.

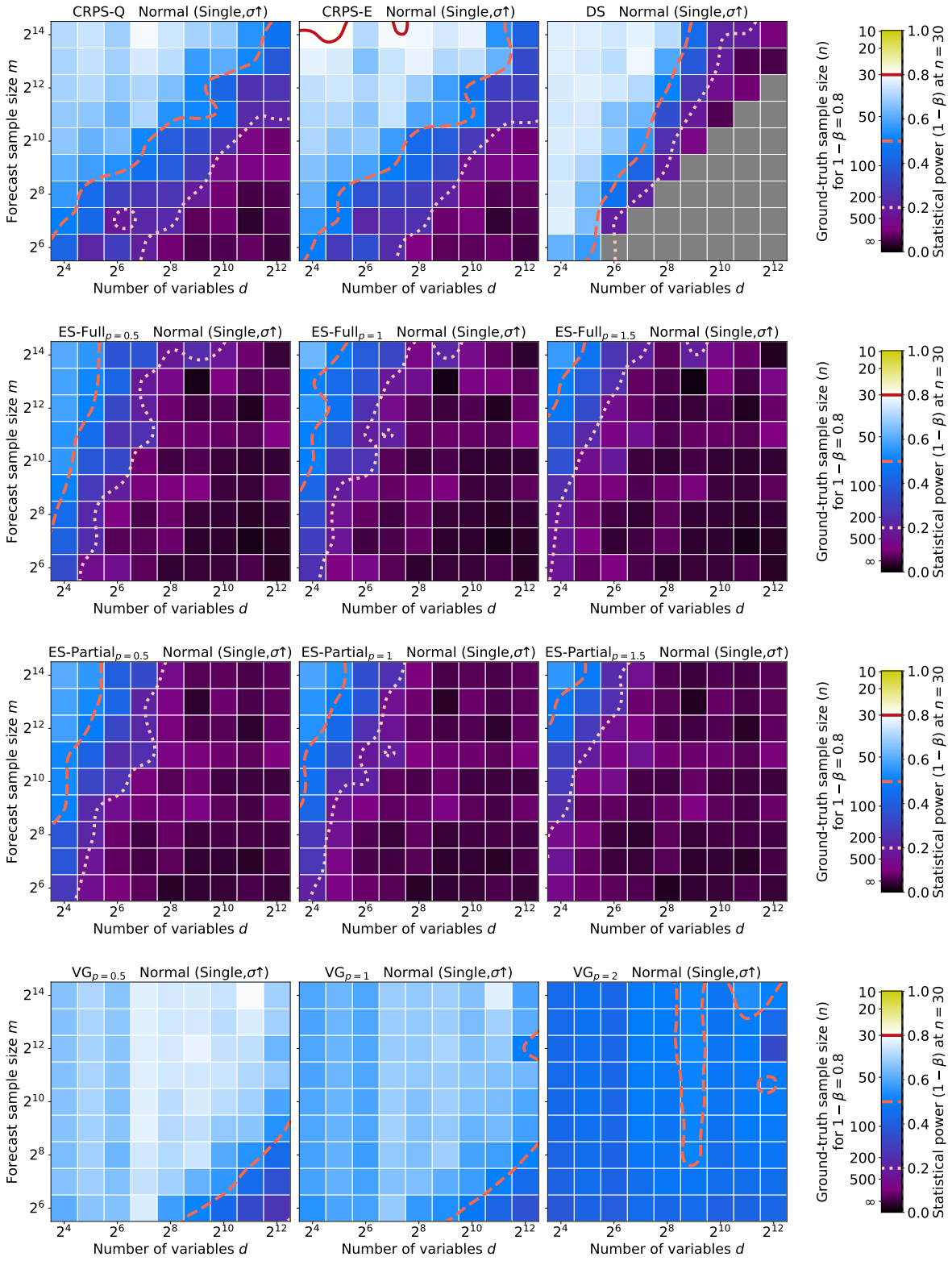

Figure 13. Statistical power of all scoring rules at correctly detecting that forecast's marginal standard deviation is higher than the ground-truth's one for a single dimension, for Normal distribution marginals. DS cannot be computed when $d \geq m$, so the corresponding area is greyed out.

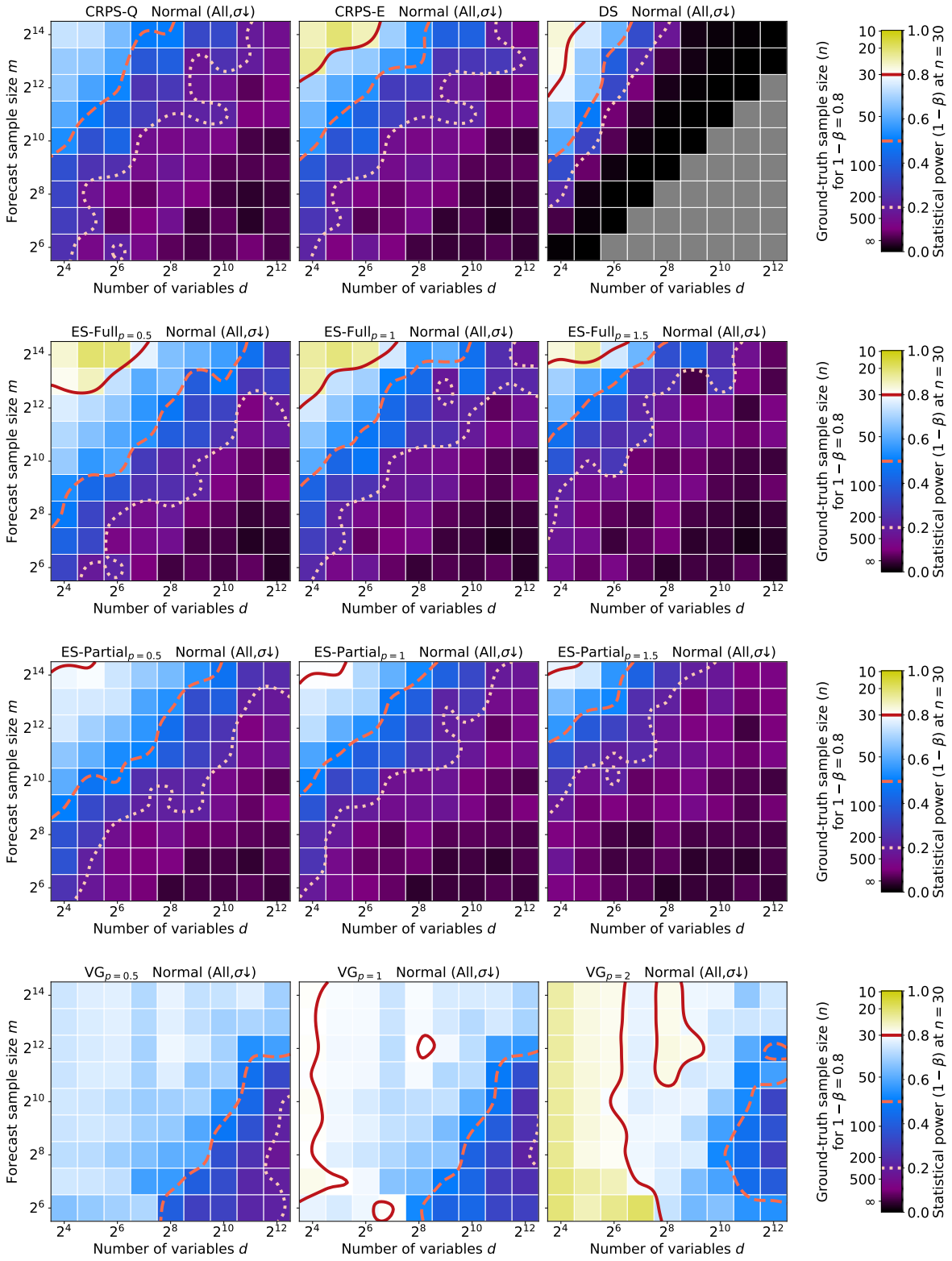

Figure 14. Statistical power of all scoring rules at correctly detecting that forecast's marginal standard deviations are lower than the ground-truth's ones for all dimensions, for Normal distribution marginals. DS cannot be computed when $d \geq m$, so the corresponding area is greyed out.

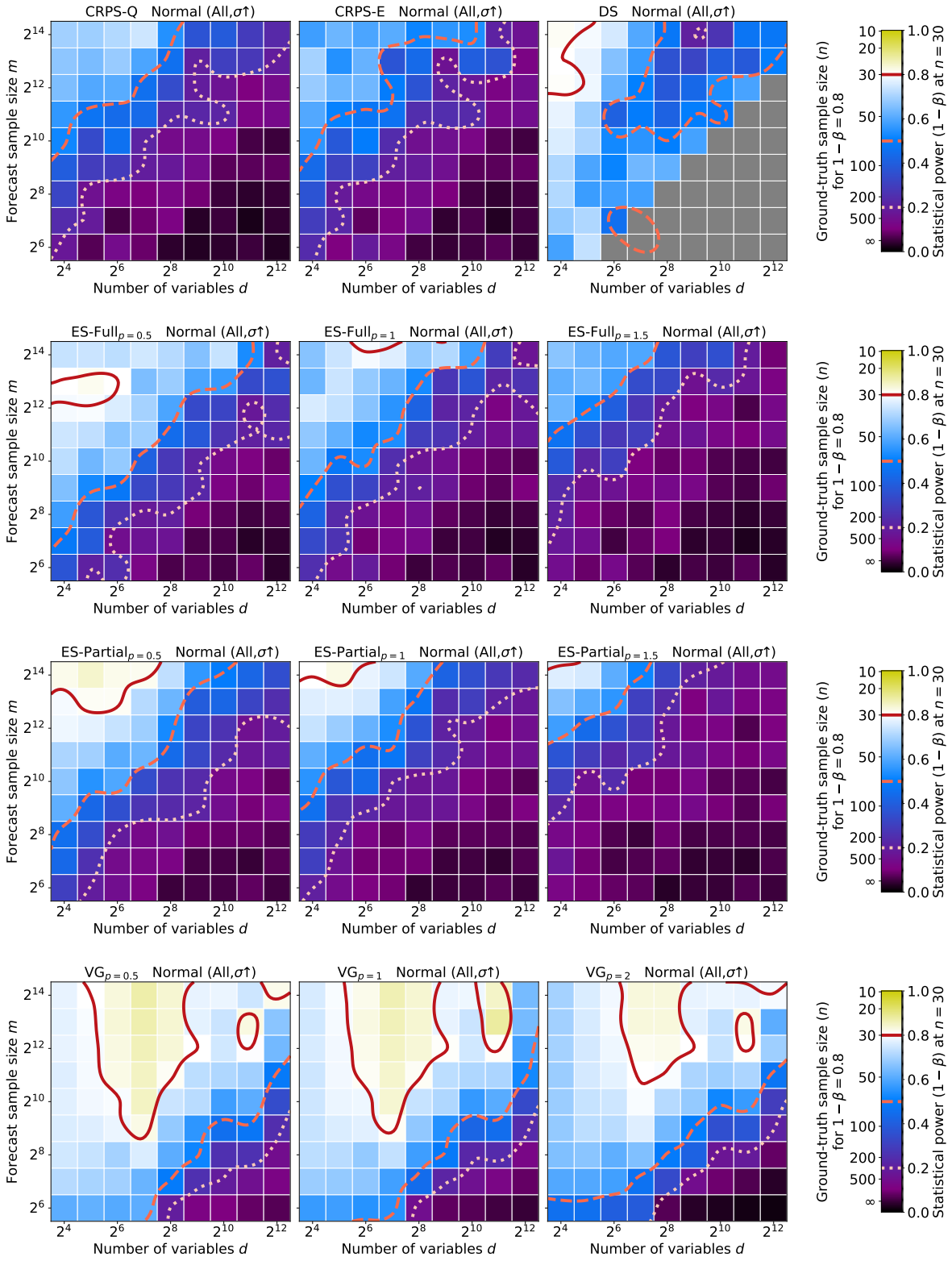

Figure 15. Statistical power of all scoring rules at correctly detecting that forecast's marginal standard deviations are lower than the ground-truth's ones for all dimensions, for Normal distribution marginals. DS cannot be computed when $d \geq m$, so the corresponding area is greyed out.

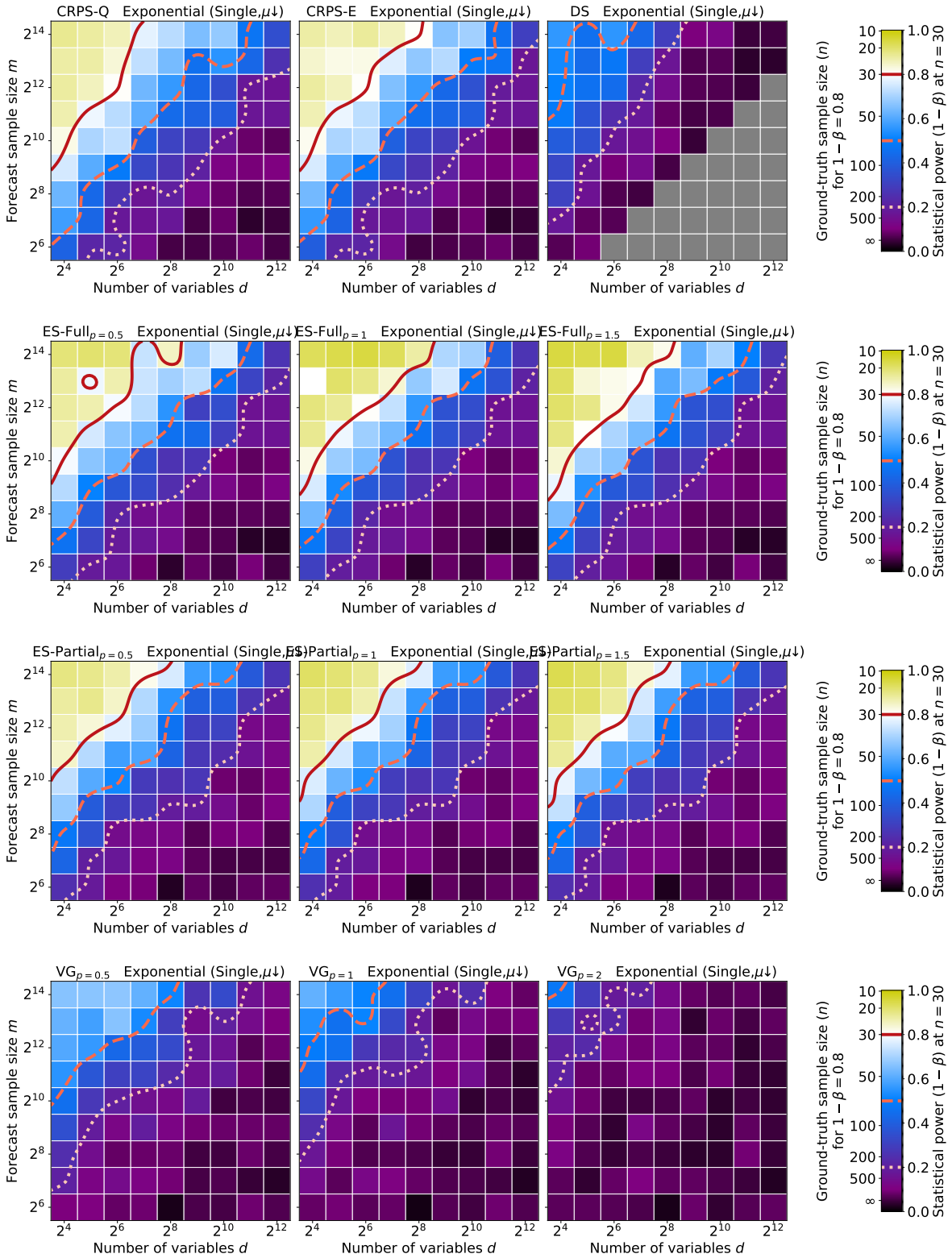

Figure 16. Statistical power of all scoring rules at correctly detecting that forecast's marginal mean is lower than the ground-truth's one for a single dimension, for Exponential distribution marginals. DS cannot be computed when $d \geq m$, so the corresponding area is greyed out.

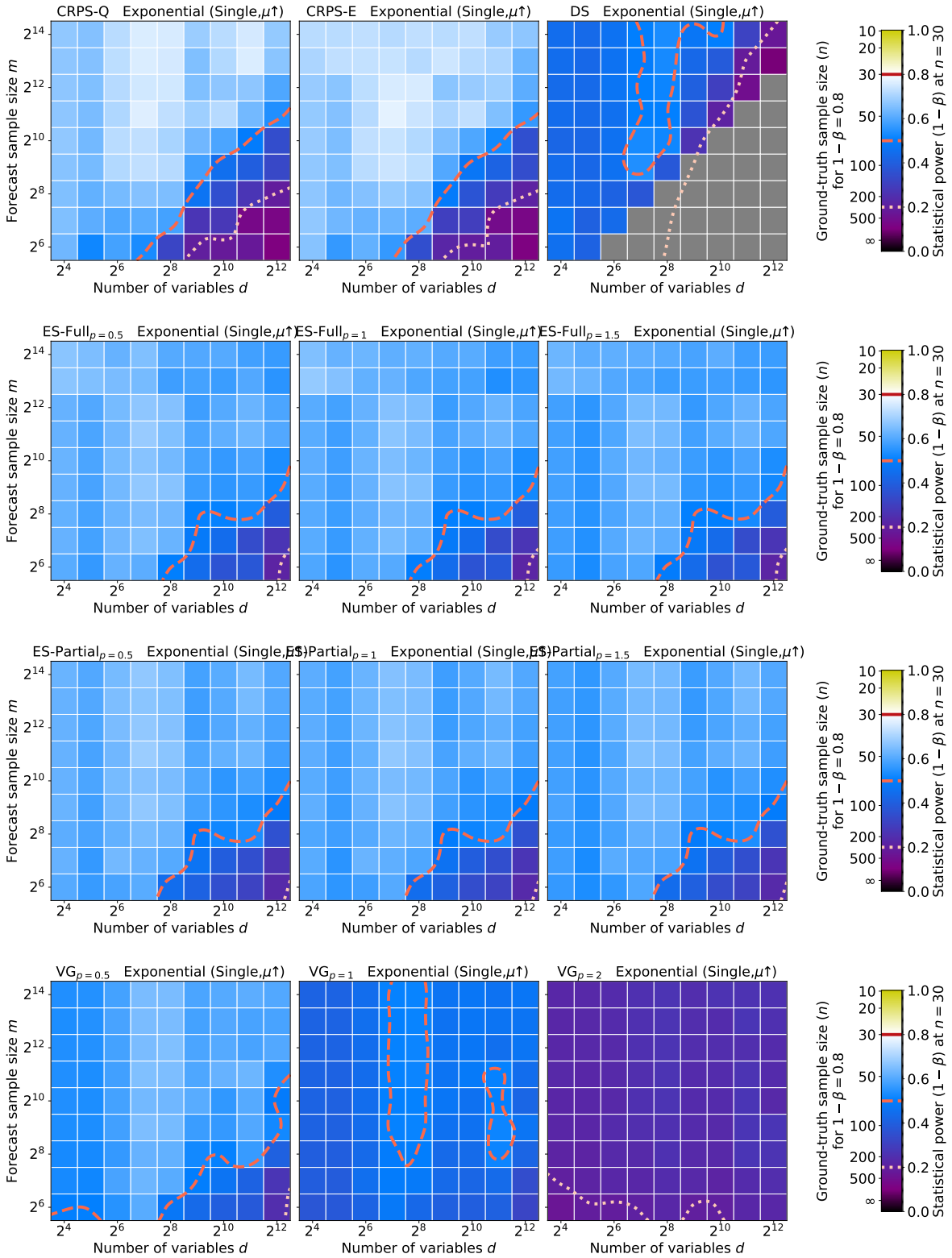

Figure 17. Statistical power of all scoring rules at correctly detecting that forecast's marginal mean is higher than the ground-truth's one for a single dimension, for Exponential distribution marginals. DS cannot be computed when $d \geq m$, so the corresponding area is greyed out.

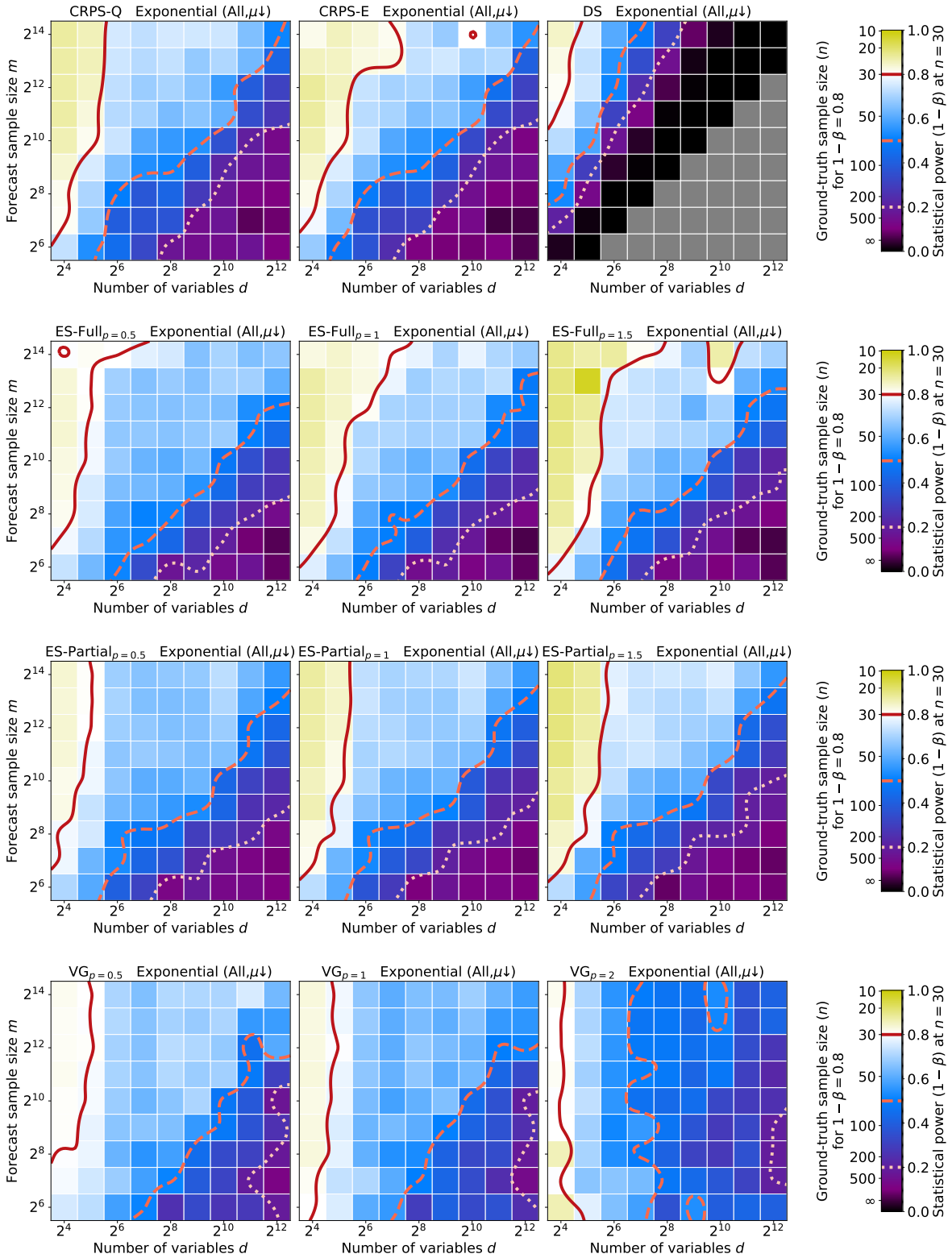

Figure 18. Statistical power of all scoring rules at correctly detecting that forecast's marginal means are lower than the ground-truth's ones for all dimensions, for Exponential distribution marginals. DS cannot be computed when $d \geq m$, so the corresponding area is greyed out.

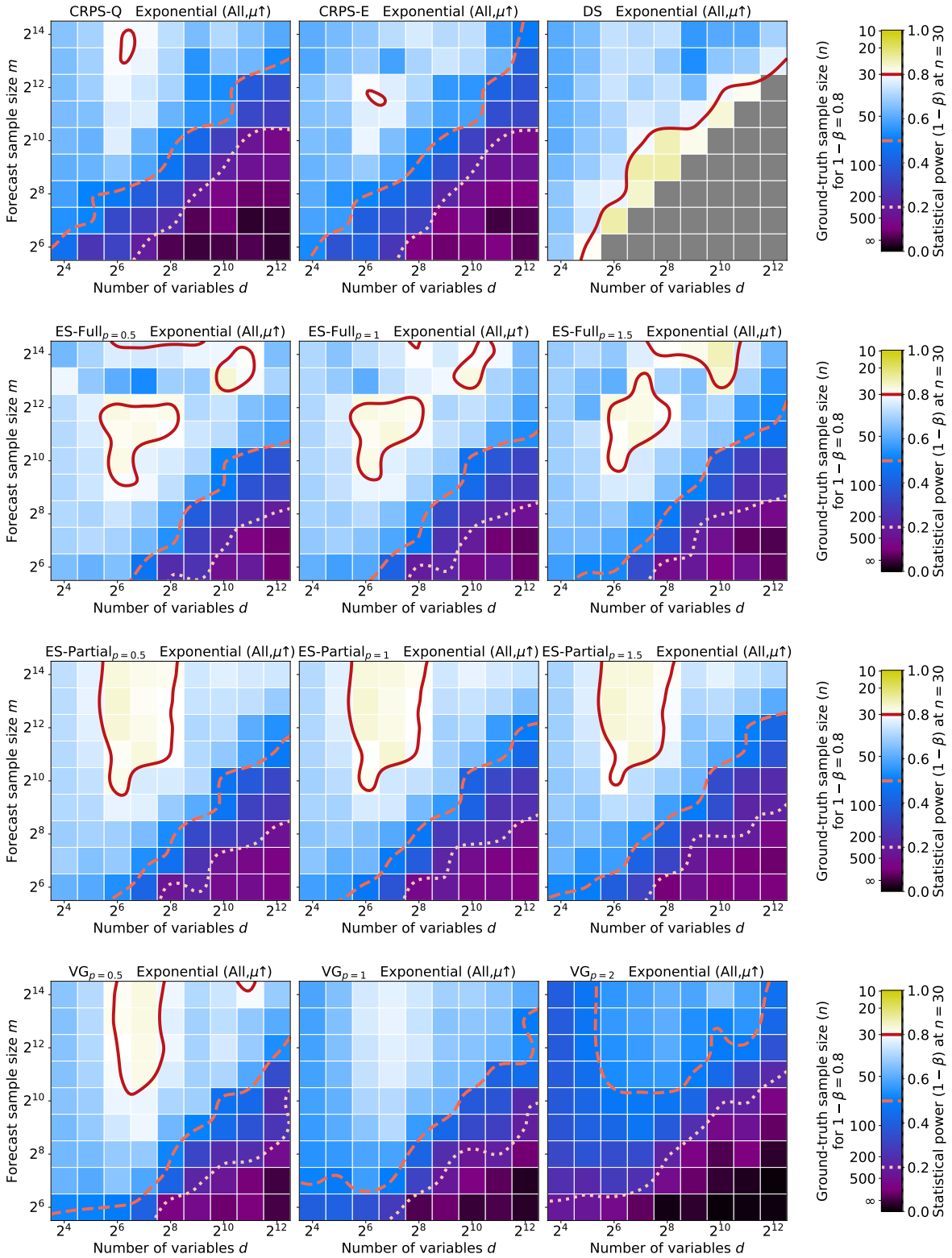

Figure 19. Statistical power of all scoring rules at correctly detecting that forecast's marginal means are higher than the ground-truth's ones for all dimensions, for Exponential distribution marginals. DS cannot be computed when $d \geq m$, so the corresponding area is greyed out.

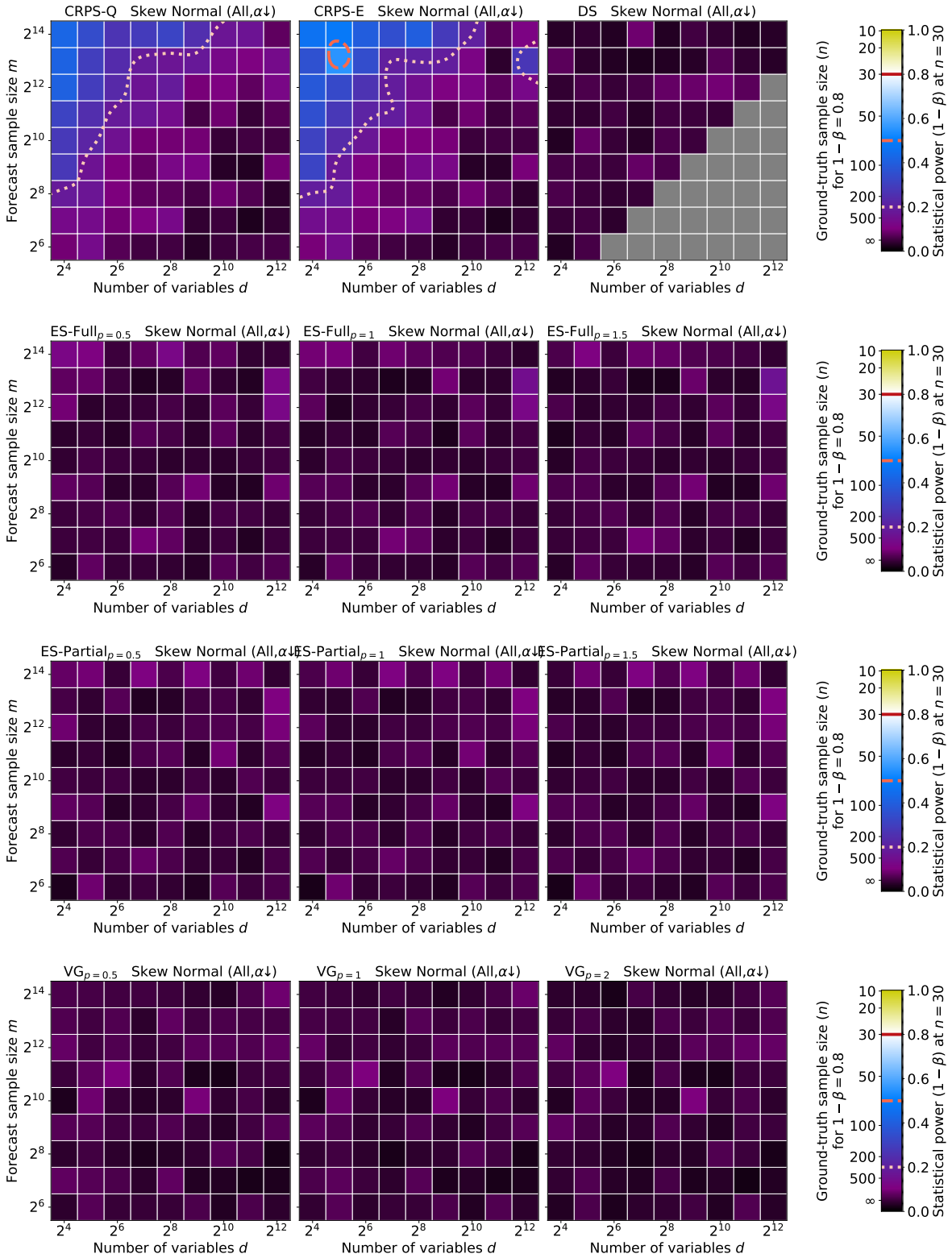

Figure 20. Statistical power of all scoring rules at correctly detecting that forecast's marginal skewness are lower than the ground-truth's ones for all dimensions, for Skew Normal distribution marginals. DS cannot be computed when $d \geq m$, so the corresponding area is greyed out.

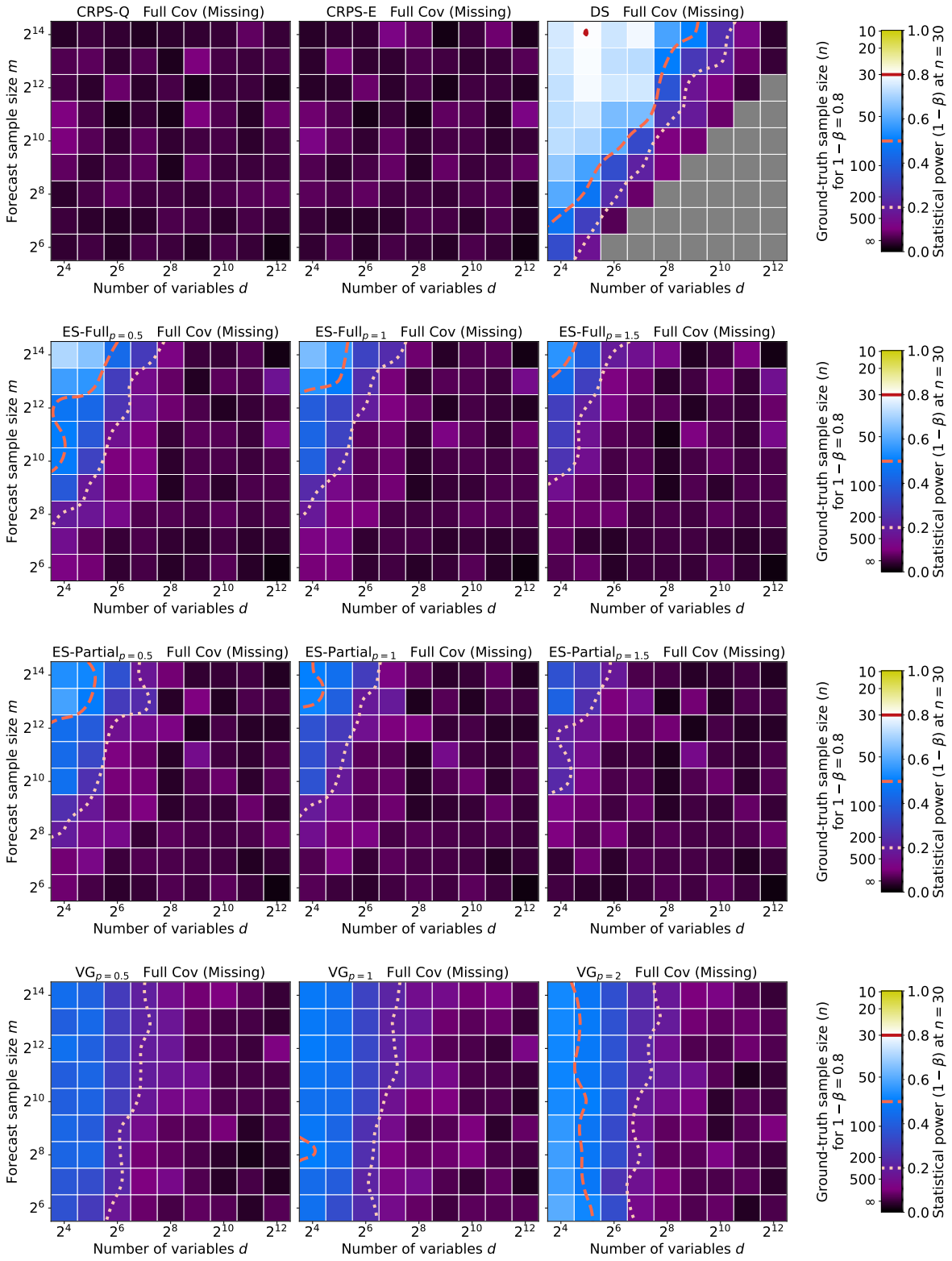

Figure 21. Statistical power of all scoring rules at correctly detecting that all variables are independent in the forecast, while they are all positively correlated in the ground truth, for Normal distribution marginals. DS cannot be computed when $d \geq m$, so the corresponding area is greyed out.

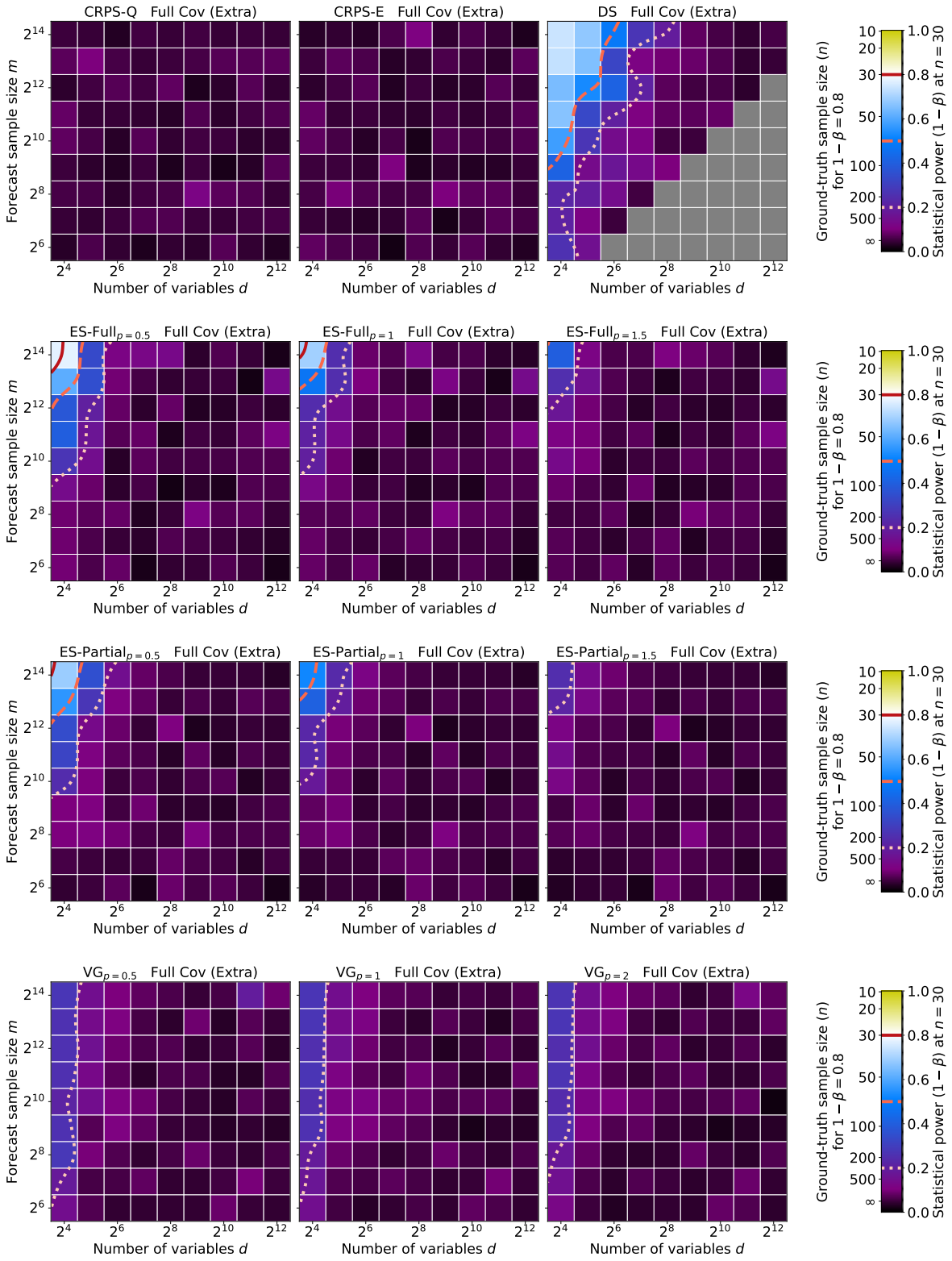

Figure 22. Statistical power of all scoring rules at correctly detecting that all variables are positively correlated in the forecast, while they are all independent in the ground truth, for Normal distribution marginals. DS cannot be computed when $d \geq m$, so the corresponding area is greyed out.

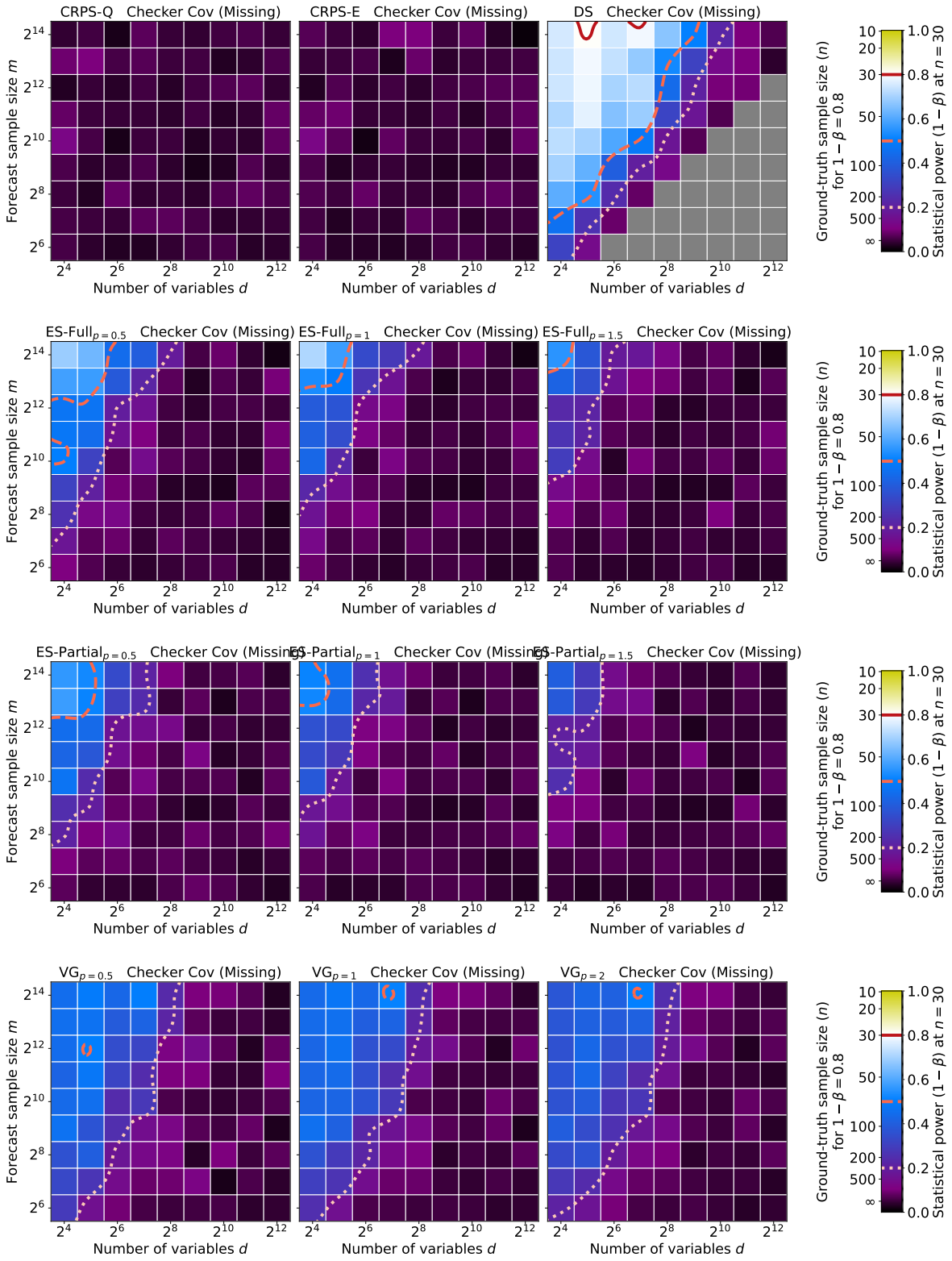

Figure 23. Statistical power of all scoring rules at correctly detecting that all variables are independent in the forecast, while they are all either positively or negatively correlated in the ground truth, for Normal distribution marginals. DS cannot be computed when $d \geq m$, so the corresponding area is greyed out.

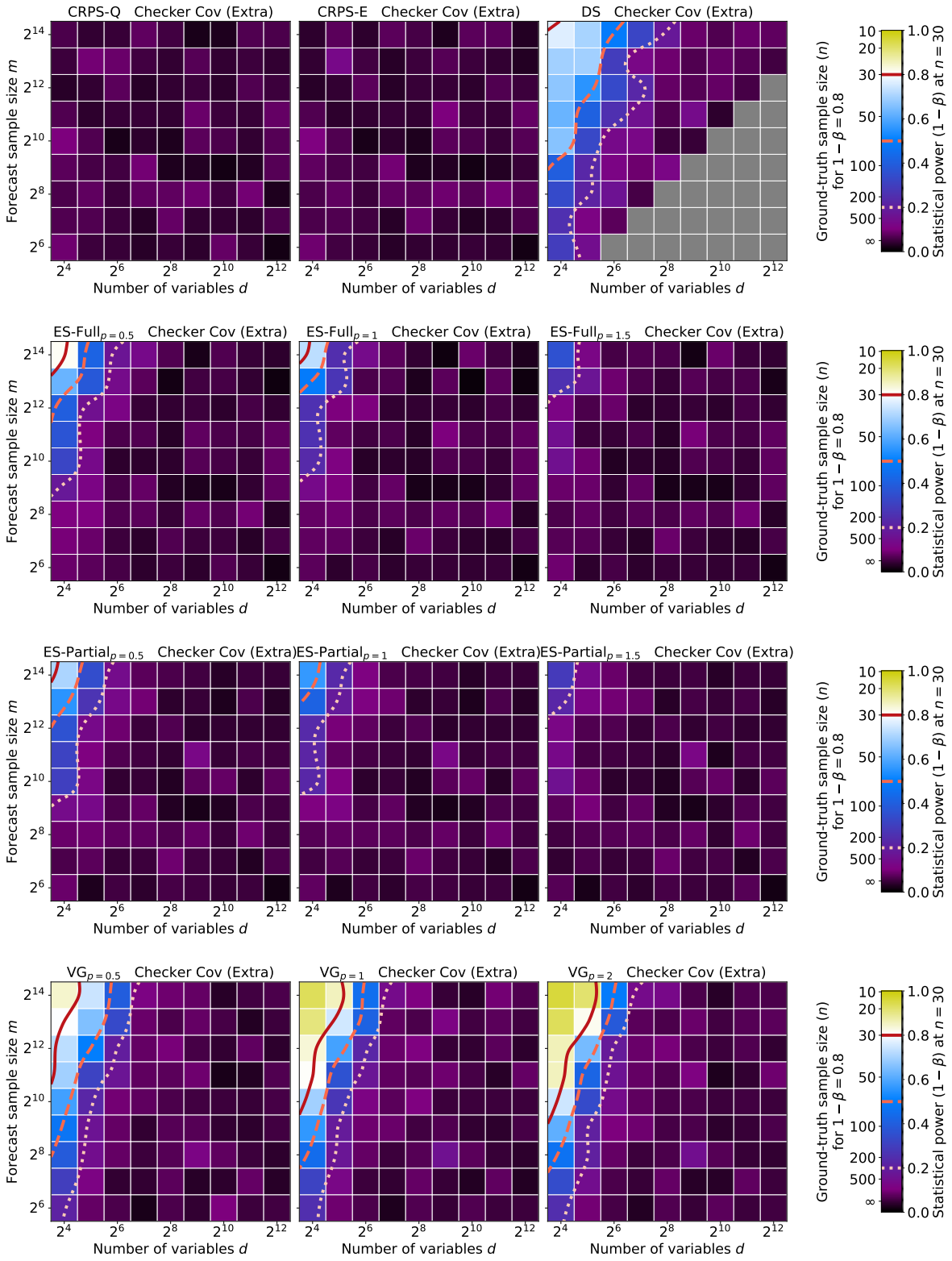

Figure 24. Statistical power of all scoring rules at correctly detecting that all variables are either positively or negatively correlated in the forecast, while they are all independent in the ground truth, for Normal distribution marginals. DS cannot be computed when $d \geq m$, so the corresponding area is greyed out.

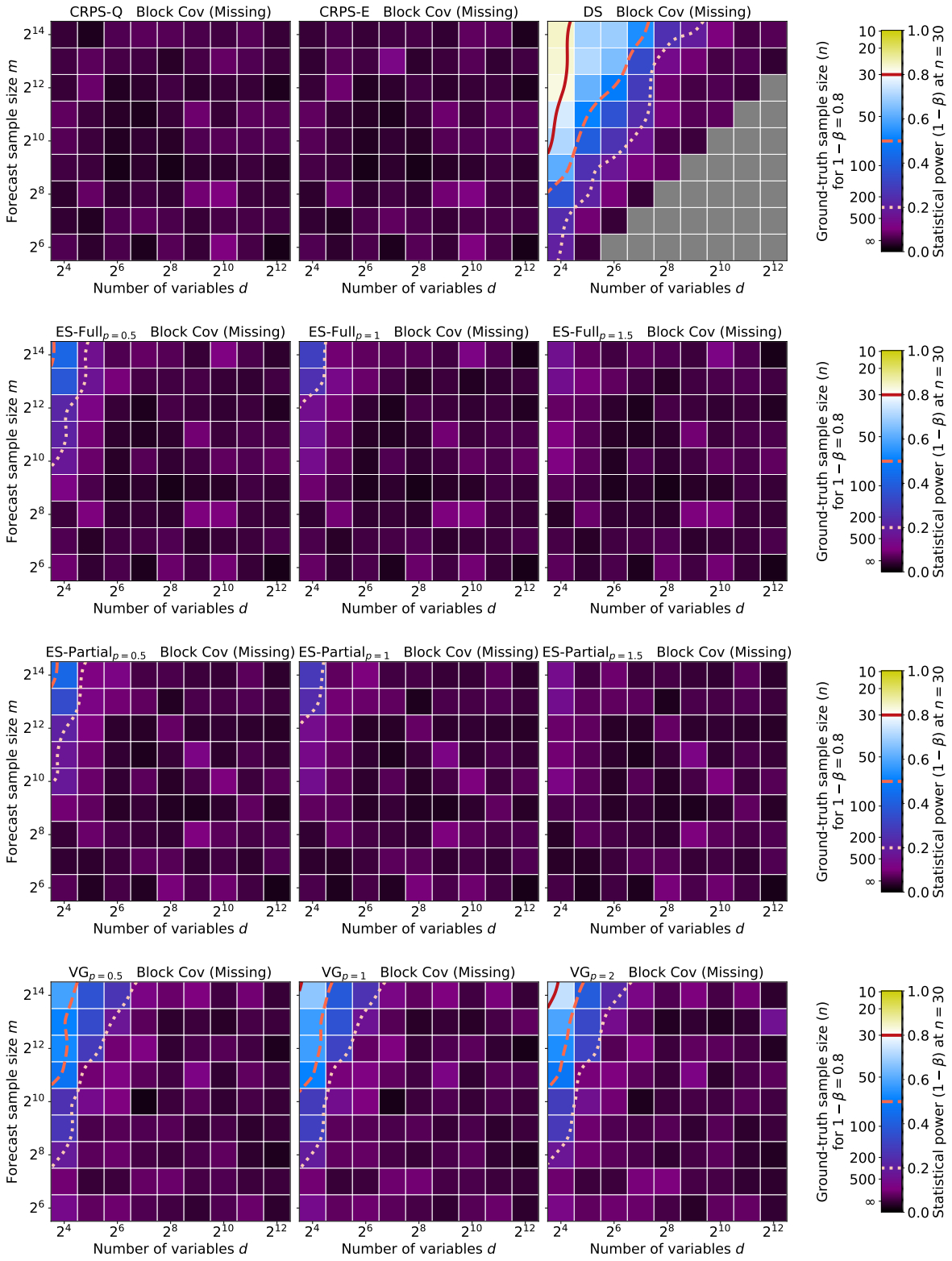

Figure 25. Statistical power of all scoring rules at correctly detecting that all variables are independent in the forecast, while each pair are positively correlated in the ground truth, for Normal distribution marginals. DS cannot be computed when $d \geq m$, so the corresponding area is greyed out.

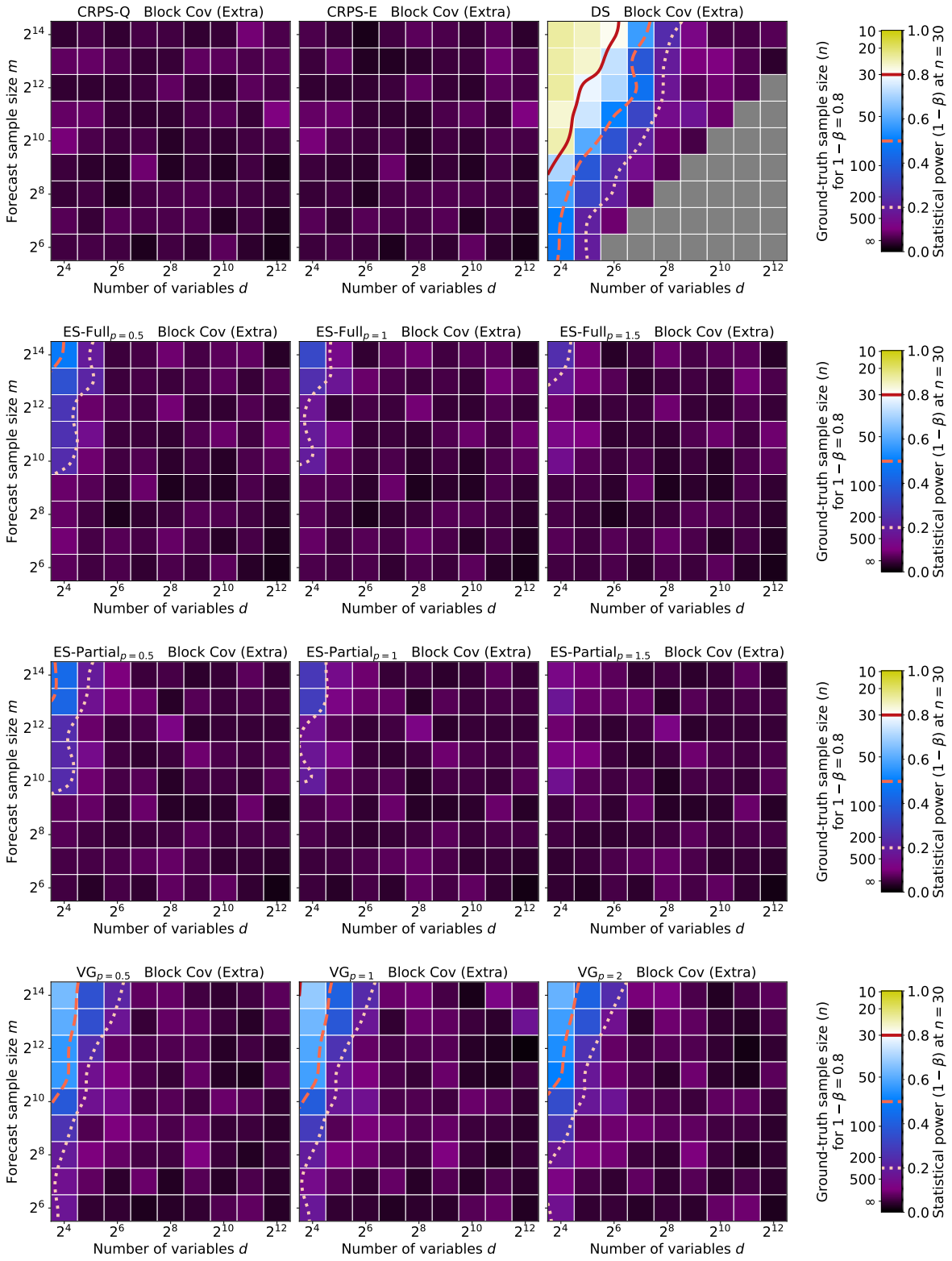

Figure 26. Statistical power of all scoring rules at correctly detecting that all pairs of variables are either positively correlated in the forecast, while they are all independent in the ground truth, for Normal distribution marginals. DS cannot be computed when $d \geq m$, so the corresponding area is greyed out.

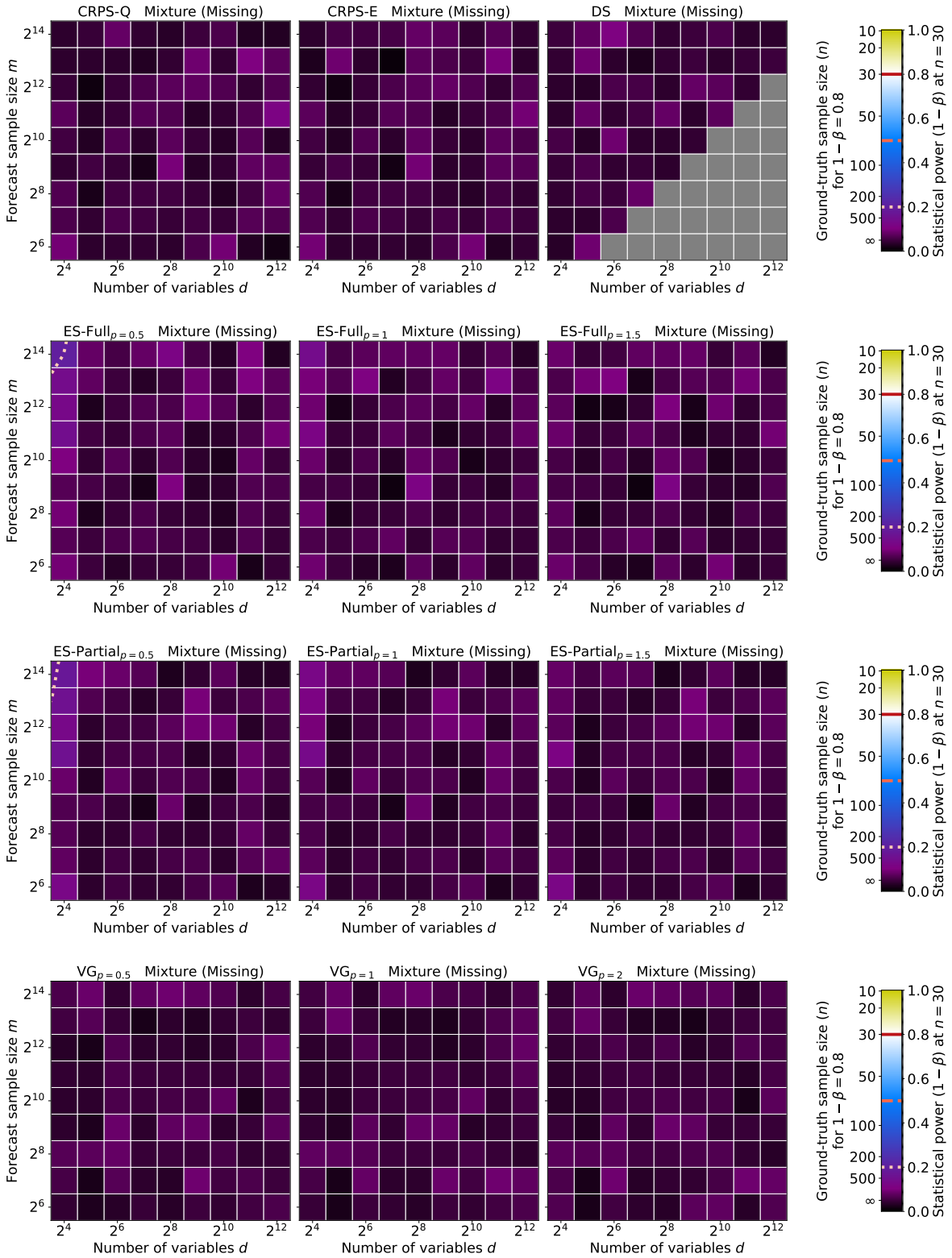

Figure 27. Statistical power of all scoring rules at correctly detecting that the forecast's distribution only contains a single mode, while the ground-truth's one contains two. DS cannot be computed when $d \geq m$, so the corresponding area is greyed out.

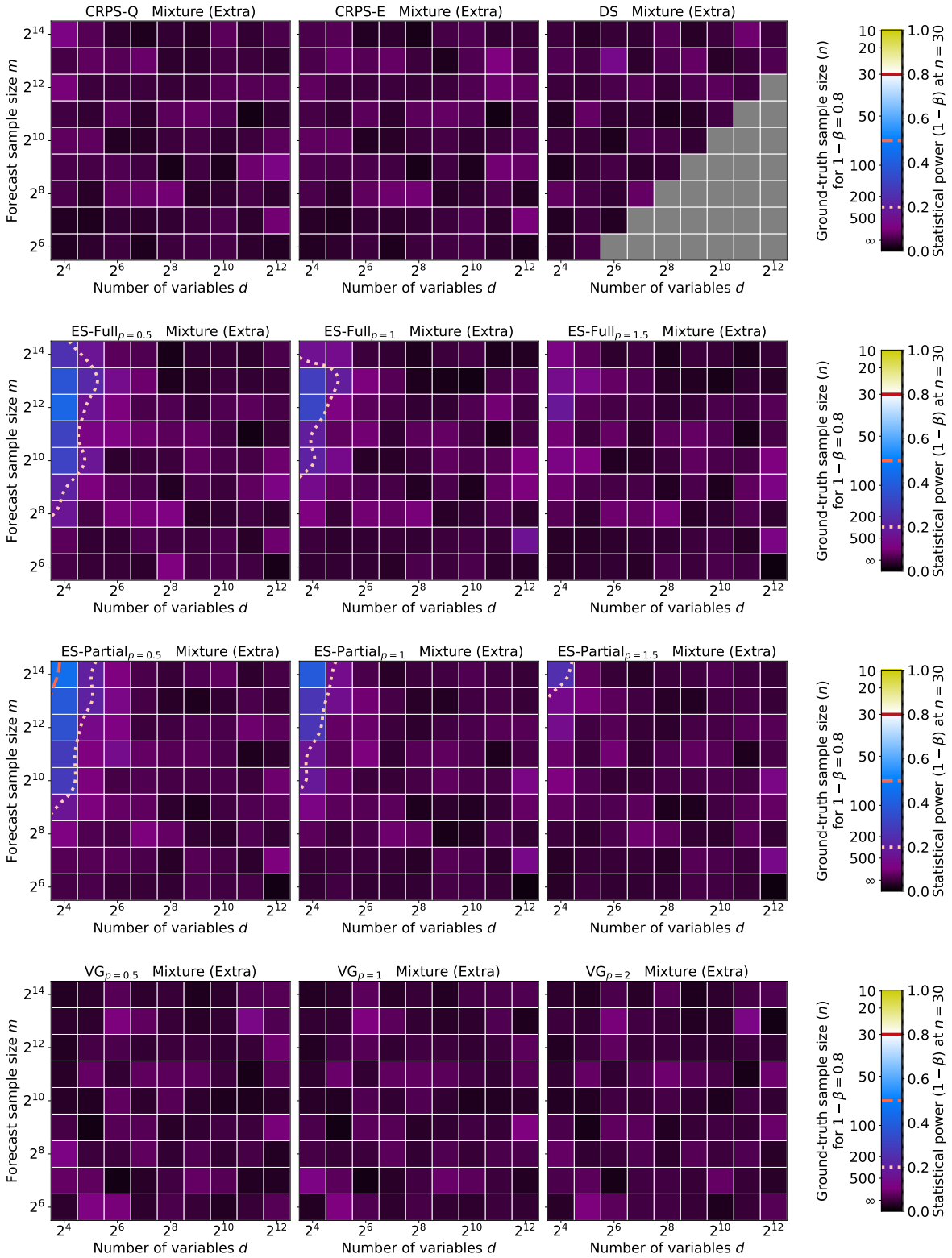

Figure 28. Statistical power of all scoring rules at correctly detecting that the forecast's distribution contains two modes, while the ground-truth's one only contains a single one. DS cannot be computed when $d \geq m$, so the corresponding area is greyed out.

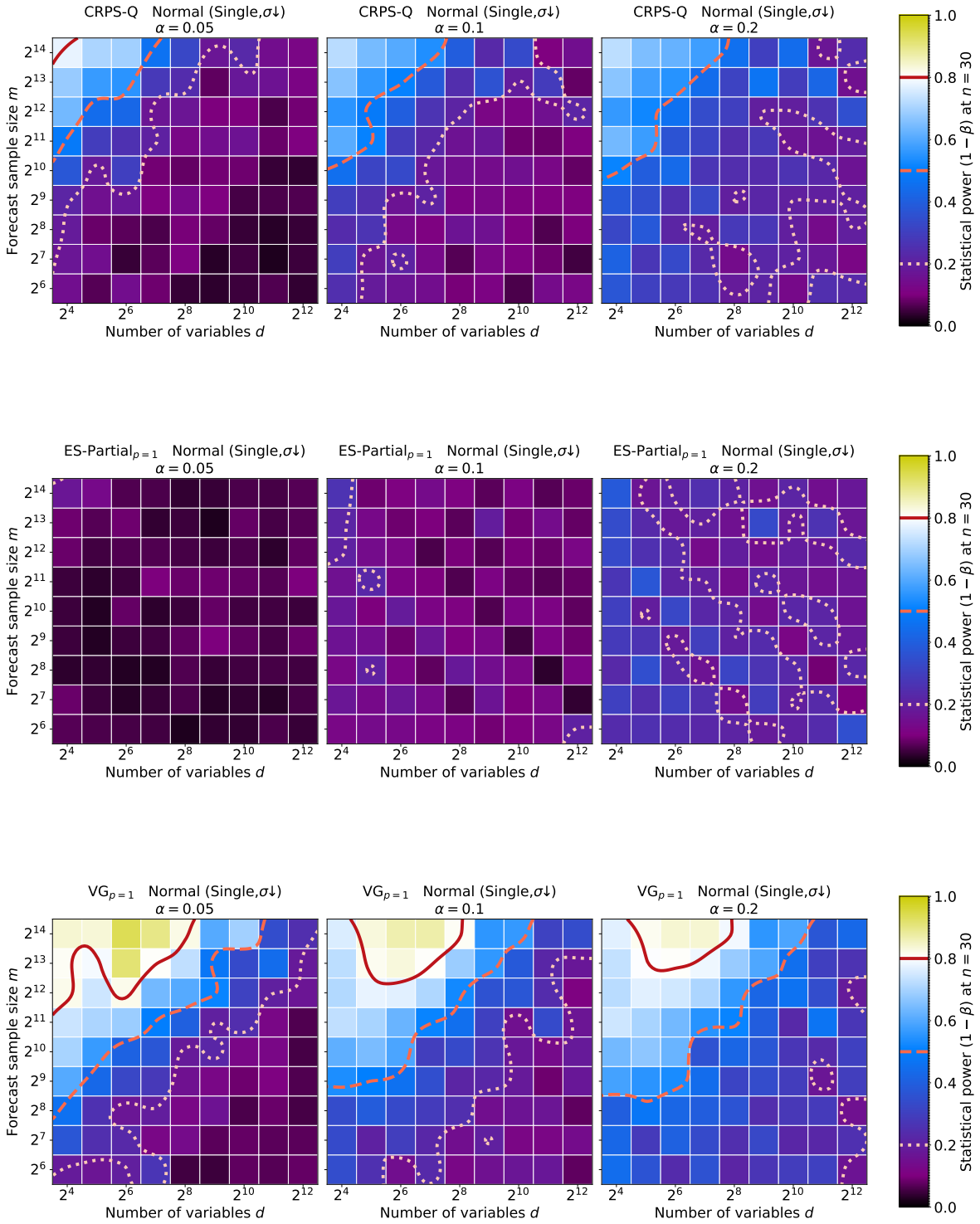

Figure 30. Statistical power of a subset of scoring rules at correctly detecting that forecast's marginal standard deviation is lower than the ground-truth's one for a single dimension, for Normal distribution marginals. The significance level α is varied for both the NLL tuning and the scoring rules under test.

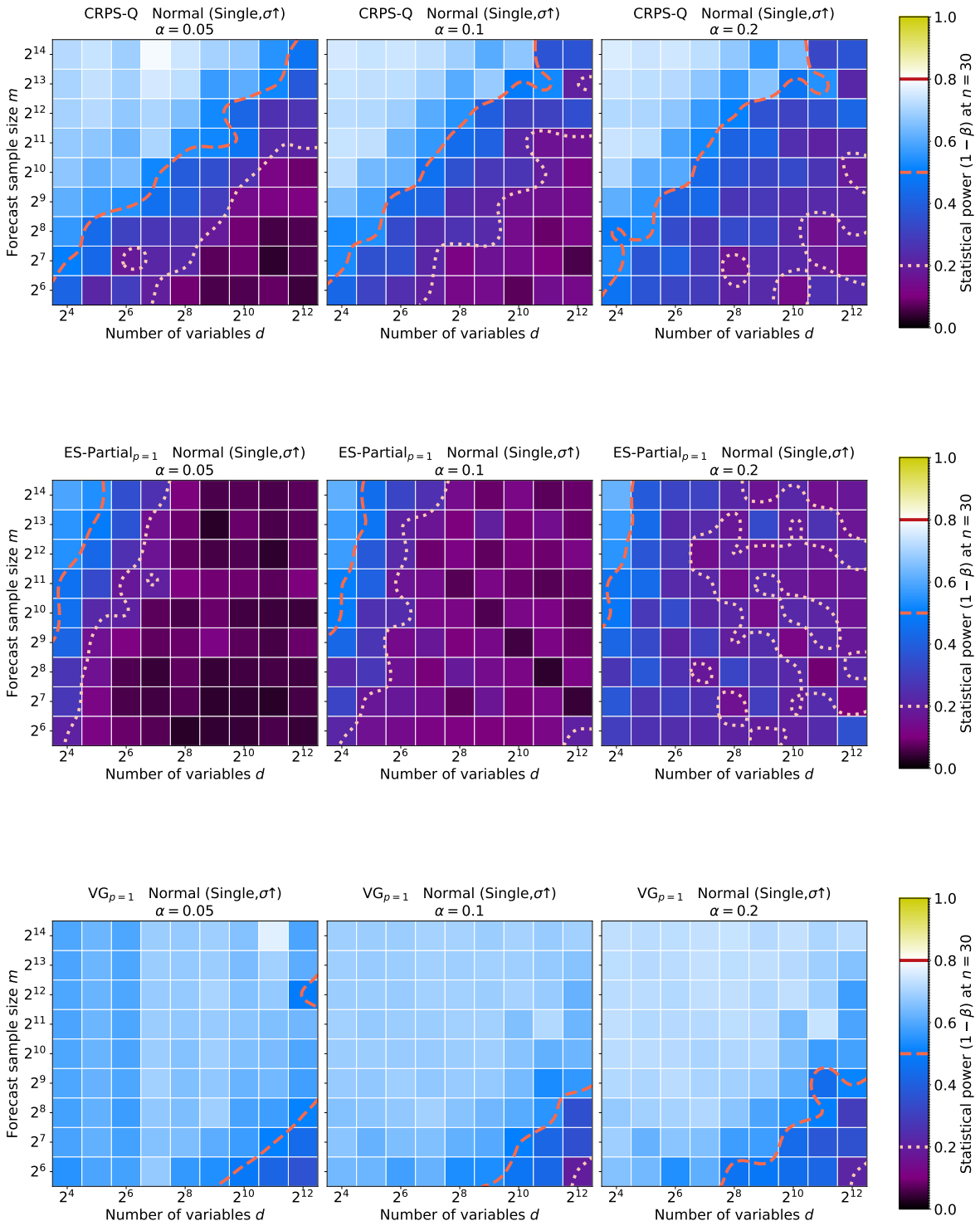

Figure 31. Statistical power of a subset of scoring rules at correctly detecting that forecast's marginal standard deviation is higher than the ground-truth's one for a single dimension, for Normal distribution marginals. The significance level α is varied for both the NLL tuning and the scoring rules under test.

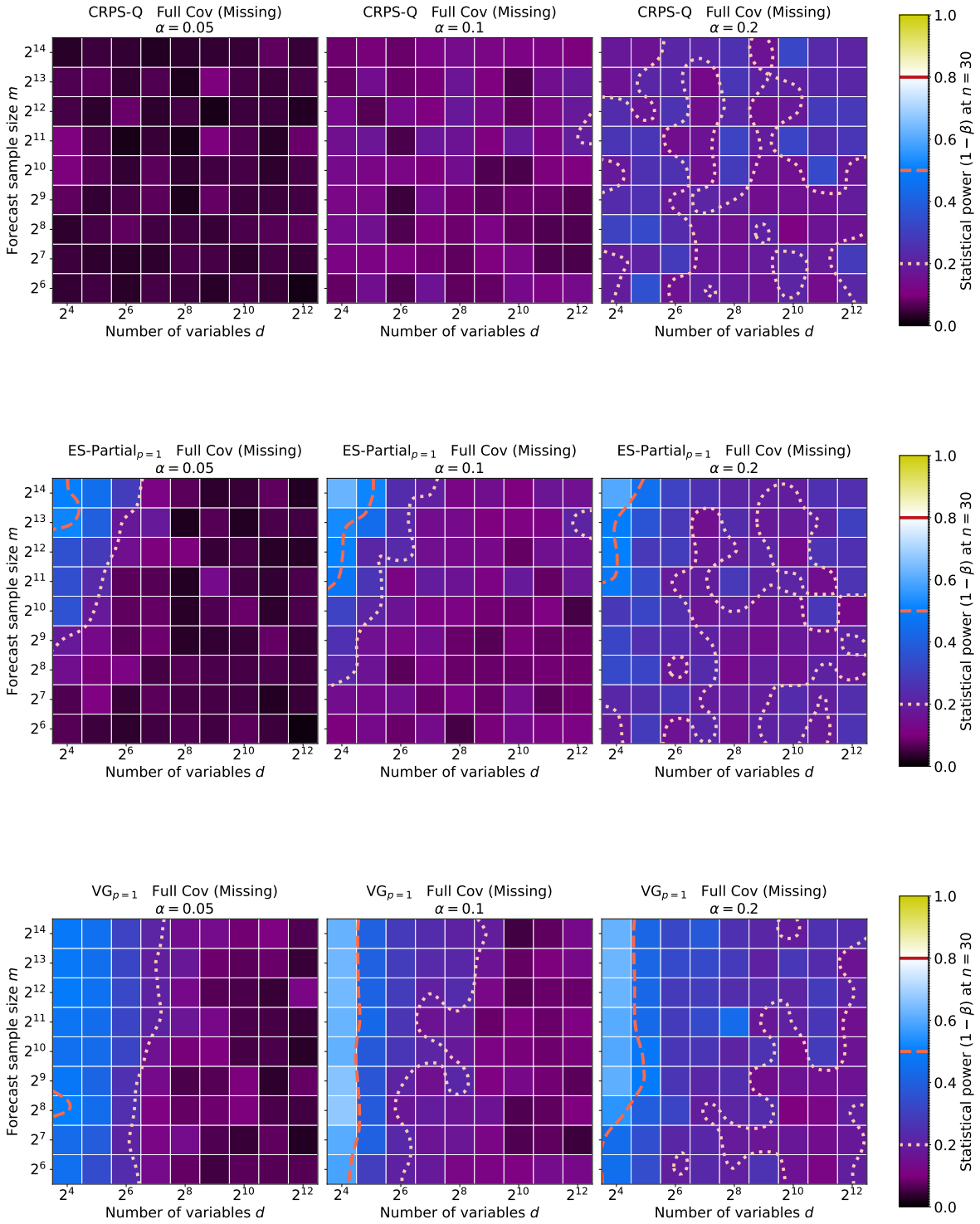

Figure 32. Statistical power of a subset of scoring rules at correctly detecting that all variables are independent in the forecast, while they are all positively correlated in the ground truth, for Normal distribution marginals. The significance level α is varied for both the NLL tuning and the scoring rules under test.

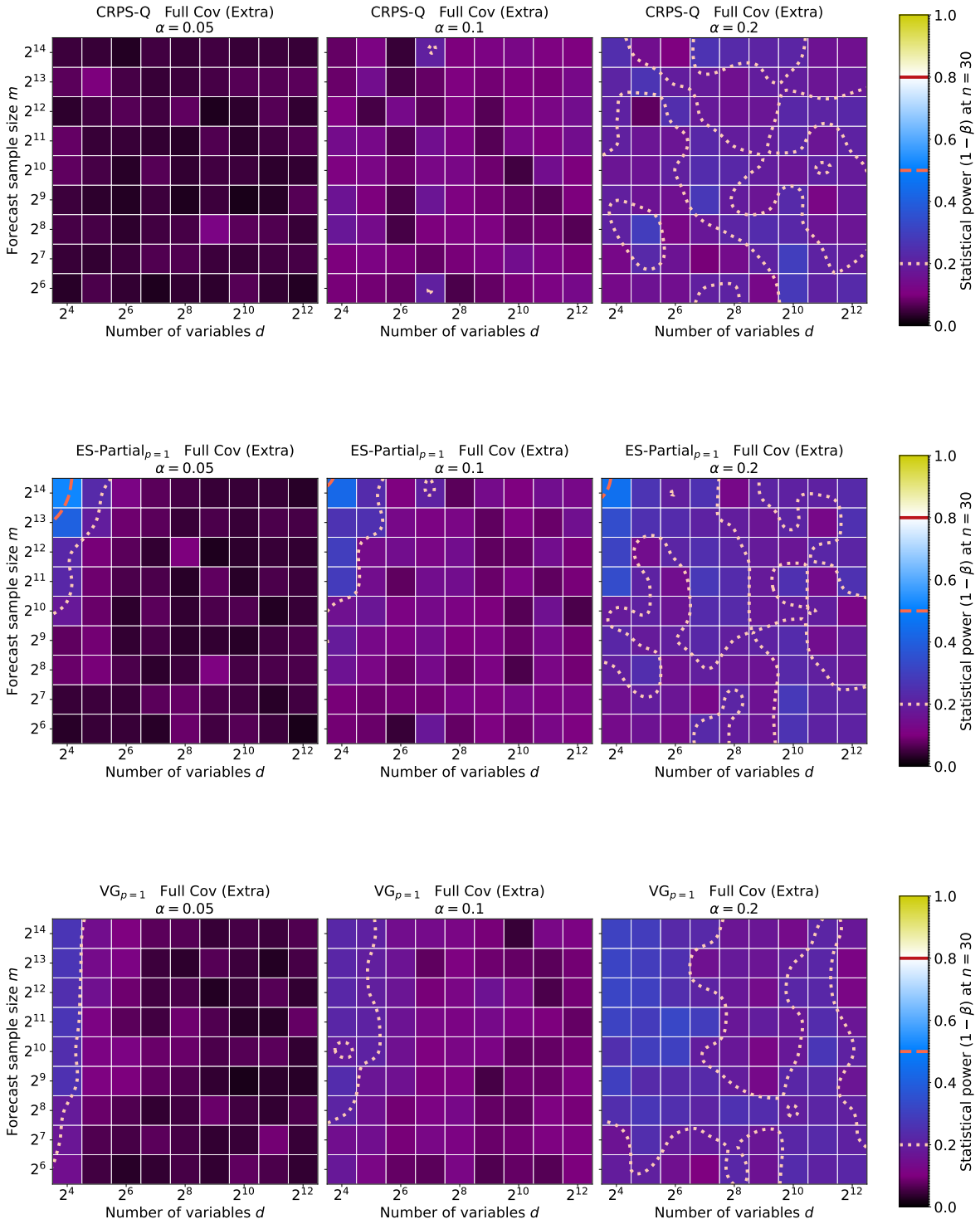

Figure 33. Statistical power of a subset of scoring rules at correctly detecting that all variables are positively correlated in the forecast, while they are all independent in the ground truth, for Normal distribution marginals. The significance level α is varied for both the NLL tuning and the scoring rules under test.

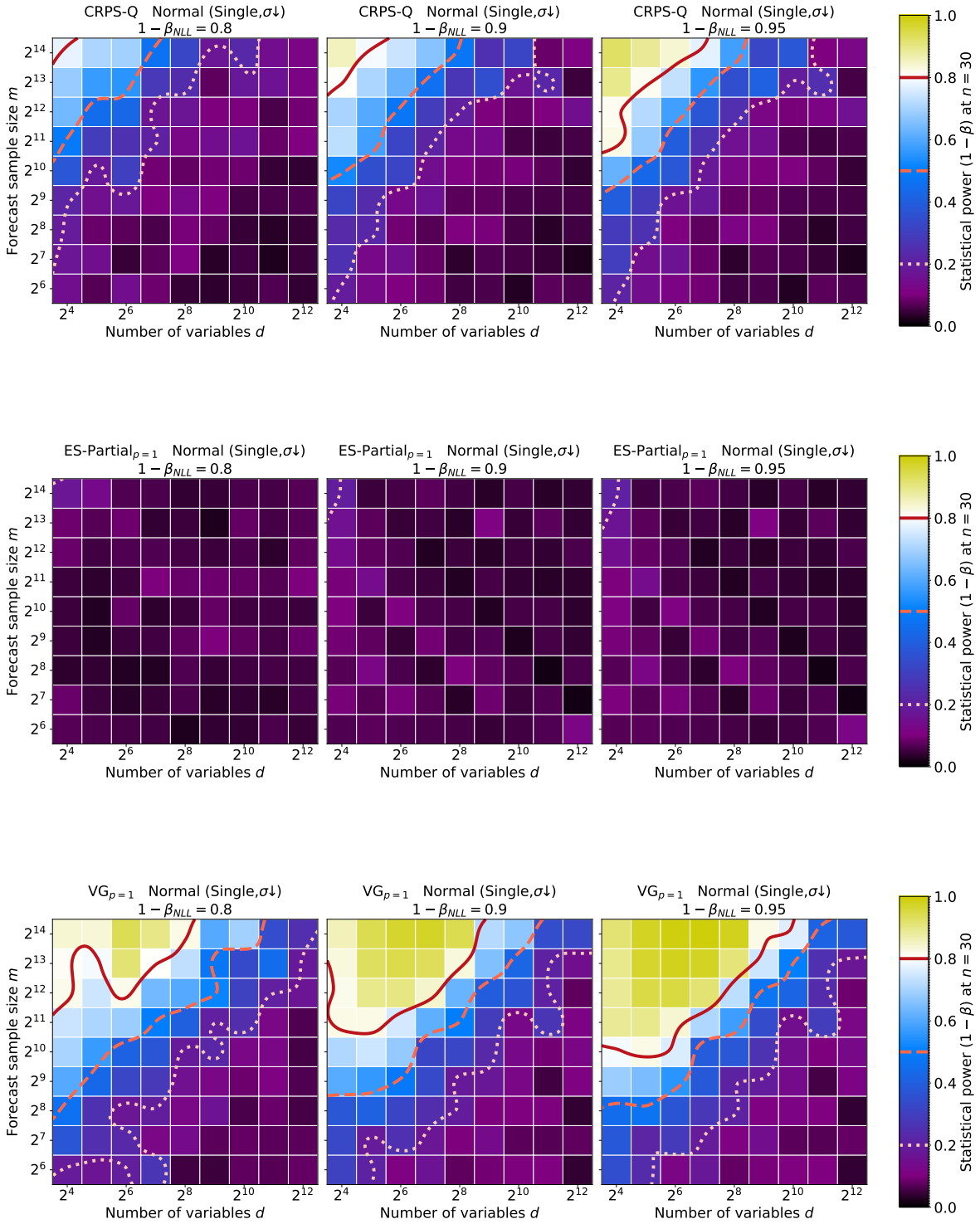

Figure 34. Statistical power of a subset of scoring rules at correctly detecting that forecast's marginal standard deviation is lower than the ground-truth's one for a single dimension, for Normal distribution marginals. The target statistical power $1-\beta_{NLL}$ is varied for the NLL tuning.

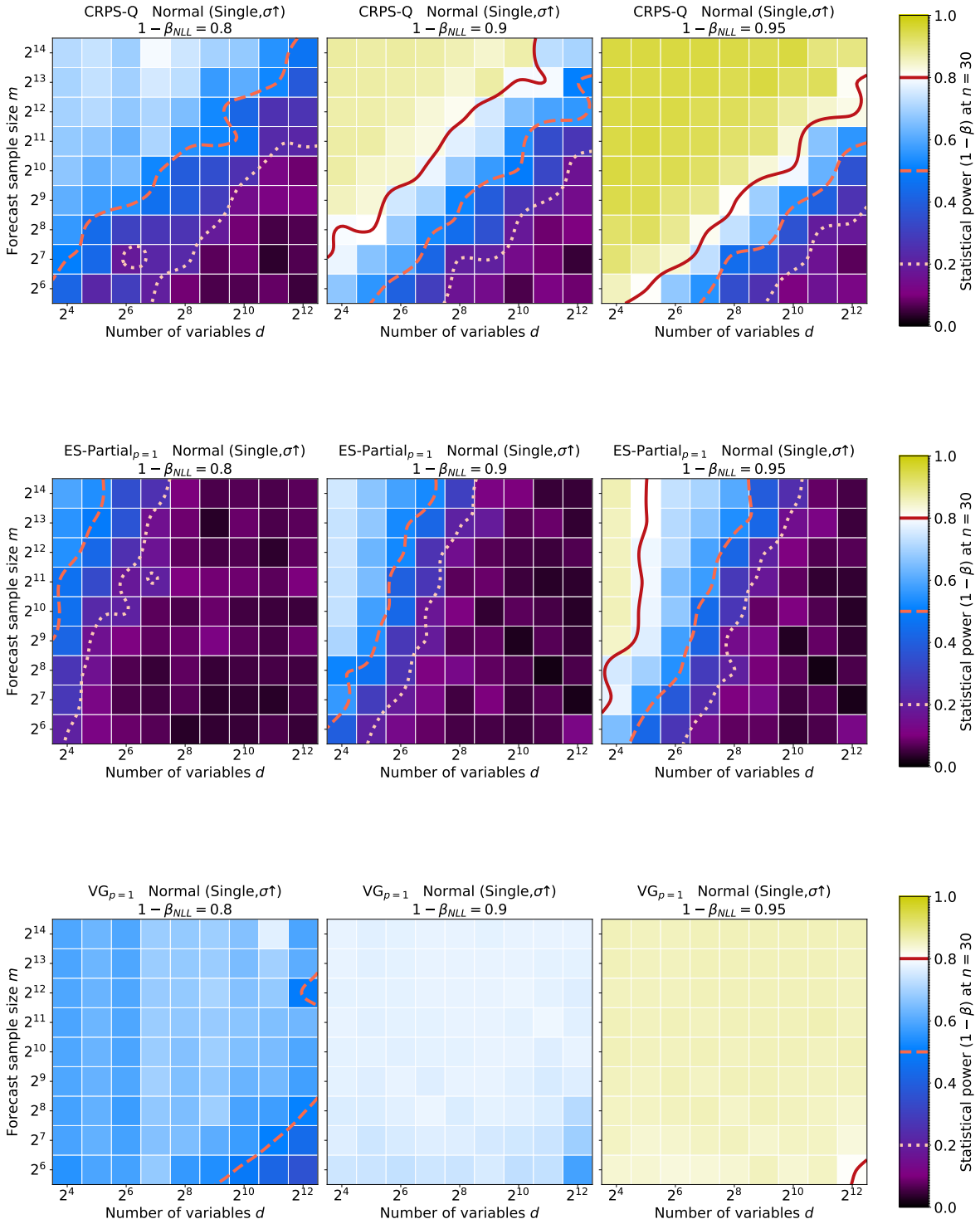

Figure 35. Statistical power of a subset of scoring rules at correctly detecting that forecast's marginal standard deviation is higher than the ground-truth's one for a single dimension, for Normal distribution marginals. The target statistical power $1-\beta_{NLL}$ is varied for the NLL tuning.

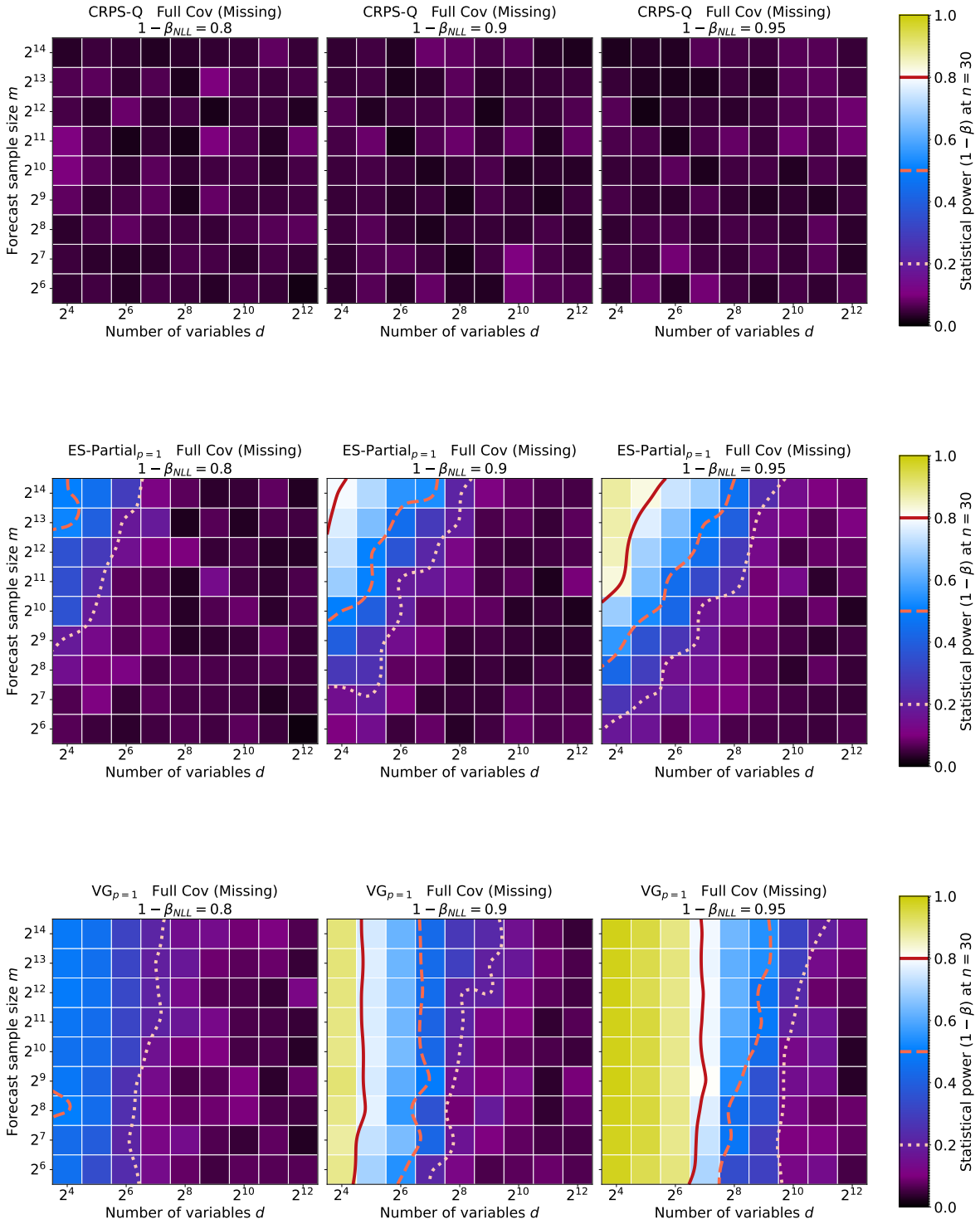

Figure 36. Statistical power of a subset of scoring rules at correctly detecting that all variables are independent in the forecast, while they are all positively correlated in the ground truth, for Normal distribution marginals. The target statistical power $1-\beta_{NLL}$ is varied for the NLL tuning.

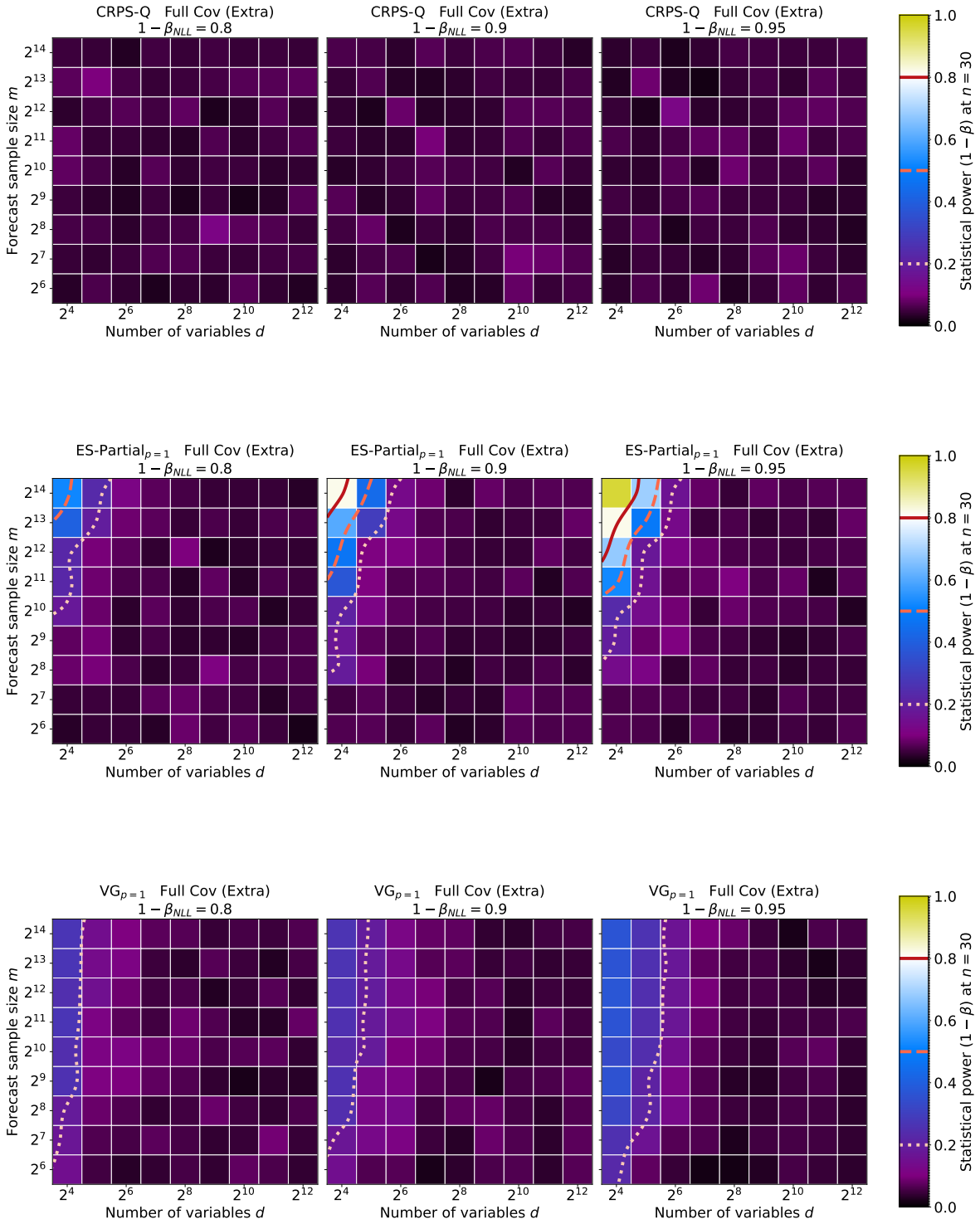

Figure 37. Statistical power of a subset of scoring rules at correctly detecting that all variables are positively correlated in the forecast, while they are all independent in the ground truth, for Normal distribution marginals. The target statistical power $1-\beta_{NLL}$ is varied for the NLL tuning.