



UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE TECNOLOGIA E GEOCIÊNCIAS
COMISSÃO NACIONAL DE ENERGIA NUCLEAR
CENTRO REGIONAL DE CIÊNCIAS NUCLEARES DO NORDESTE
PROGRAMA DE PÓS-GRADUAÇÃO EM TECNOLOGIAS ENERGÉTICAS E NUCLEARES

VALENTIN PASCHOAL PERRUCCI

**METODOLOGIA PARA PRÉ-PROCESSAMENTO BASEADO EM
COMPLEMENTARIDADE ESPECTRAL E COMBINAÇÃO DE MODELOS DE
DOWNSCALING ESTATÍSTICO COM VISTAS À PREVISÃO OPERACIONAL
EÓLICA E SOLAR DE CURTO PRAZO**

Recife

2024

VALENTIN PASCHOAL PERRUCI

**METODOLOGIA PARA PRÉ-PROCESSAMENTO BASEADO EM
COMPLEMENTARIDADE ESPECTRAL E COMBINAÇÃO DE MODELOS DE
DOWNSCALING ESTATÍSTICO COM VISTAS À PREVISÃO OPERACIONAL
EÓLICA E SOLAR DE CURTO PRAZO**

Tese apresentada ao Programa de Pós-Graduação em Tecnologias Energéticas e Nucleares da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de Doutor em Ciências.
Área de concentração: Fontes renováveis de energia

Orientadora: profa. Dra. Olga de Castro Vilela

Coorientador: Prof. Dr. Alexandre Costa

Recife, 2024

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Perruci, Valentin Paschoal.

Metodologia para pré-processamento baseado em complementaridade espectral e combinação de modelos de downscaling estatístico com vistas à previsão operacional eólica e solar de curto prazo / Valentin Paschoal Perruci. - Recife, 2024.

138f.: il.

Tese (Doutorado) - Universidade Federal de Pernambuco, Centro de Tecnologia e Geociências, Programa de Pós-Graduação em Tecnologias Energéticas e Nucleares, 2024.

Orientação: Olga de Castro Vilela.

Coorientação: Alexandre Costa.

1. Energia eólica; 2. Energia solar; 3. Previsão operacional; 4. Downscaling estatístico; 5. Combinação de modelos; 6. Complementaridade espectral. I. Vilela, Olga de Castro. II. Costa, Alexandre. III. Título.

UFPE-Biblioteca Central

VALENTIN PASCHOAL PERRUCI

AVALIAÇÃO DE COMPLEMENTARIDADE ESPECTRAL EM MODELOS DE DOWNSCALING ESTATÍSTICO E DE COMBINAÇÃO COM VISTAS À PREVISÃO OPERACIONAL EÓLICA E SOLAR DE CURTO PRAZO

Tese apresentada ao Programa de Pós-Graduação em Tecnologias Energéticas e Nucleares da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de Doutor em Ciências. Área de concentração: Fontes renováveis de energia.

Aprovada em: 30/08/2023

BANCA EXAMINADORA:

Dra. Olga de Castro Vilela (Examinador Interno)
Universidade Federal de Pernambuco – UFPE

Dr. Gustavo de Novaes Pires Leite (Examinador Externo)
Instituto Federal de Pernambuco – IFPE

Dr. Tsang Ing Ren (Examinador Externo)
Universidade Federal de Pernambuco – UFPE

Dra. Dóris Regina Aires Veleda (Examinador Externo)
Universidade Federal de Pernambuco – UFPE

Dr. Rodrigo Martin Alonso Suarez (Examinador Externo)
Universidad de la República - UdelaR

AGRADECIMENTOS

Agradeço à minha família pelo apoio, conforto e incentivo dado em todos os momentos; A toda a equipe do Centro de Energias Renováveis (CER-UFPE), que acolhem com grande carinho, estrutura e paciência a todos os interessados, simpatizantes, envolvidos e amantes da pesquisa científica e do conhecimento em geral;

A todos os funcionários do Programa de Pós-Graduação em Tecnologias Energéticas e Nucleares;

Às seguintes entidades cujo apoio financeiro se deu durante a realização do Doutorado: PROJETOS de P&D ANEEL da IBITU.INTELIPREV, NEO.PROGFALHAS, CHESF.CASA NOVA e à CAPES pelo apoio financeiro para realização deste trabalho de pesquisa; Aos Professores Alexandre Costa e Olga Vilela, pela orientação, confiança e apoio.

RESUMO

O presente trabalho de tese se insere no contexto da recente expansão dos setores de energia eólica e solar no Brasil e no mundo, que tem demandado soluções mais acuradas para fins de previsão operacional. Para tal, costuma-se empregar saídas de modelos de previsão numérica do tempo de larga escala, cuja principal limitação está associada à baixa resolução espacial. Para o aumento da resolução espacial (*downscaling*), modelos empíricos permitem melhor descrição de fenômenos a nível local a partir da associação com sinais observados. Com vistas ao aprimoramento geral da acurácia e robustez de previsões obtidas no âmbito de *downscaling* estatístico, propõe-se a criação de metodologia que combina múltiplas camadas de modelos estatísticos, dentre os quais se incluem modelos baseados em complementaridade espectral, conceito proposto para a avaliação simultânea de similaridade entre sinais nos domínios de tempo e frequência. Tal procedimento difere de métodos tradicionais, como em análises de correlação ou de erro quadrático médio, os quais realizam operações sob a perspectiva do domínio do tempo ou frequência separadamente. Dessa forma, no contexto da modelagem estatística, os métodos propostos para a análise de complementaridade espectral permitem a quantificação da “importância” de previsões em diferentes escalas temporais e instantes de tempo. Nos estudos de caso, foram avaliadas previsões para velocidade do vento, irradiância solar e potência gerada em centrais eólica e solar localizadas na Região Nordeste do Brasil. Os resultados mostraram importantes aprimoramentos obtidos a partir da plataforma de modelagem proposta em termos da robustez das soluções finais. Em geral, os modelos de combinação apresentaram aprimoramento na acurácia de 20% em média, podendo chegar até 50% em relação a outros modelos avaliados. Por outra parte, os novos modelos baseados em complementaridade espectral trouxeram melhorias expressivas em relação à qualidade geral das previsões, apresentando ganhos na performance geral em comparação com modelos similares.

Palavras-chave: Energia eólica; Energia solar; Previsão operacional; *Downscaling* estatístico; Combinação de modelos; Complementaridade espectral.

ABSTRACT

This thesis is part of the recent expansion of Brazil's wind and solar energy sectors worldwide, which has demanded more accurate solutions for operational forecasting. To this end, outputs from large-scale numerical weather prediction models are usually used, the main limitation of which is their low spatial resolution. In order to increase spatial resolution (downscaling), empirical models allow for a better description of phenomena at a local level by associating them with observed signals. To generally improve the accuracy and robustness of forecasts obtained in the context of statistical downscaling, we propose the creation of a methodology that combines multiple layers of statistical models, including models based on spectral complementarity, a concept developed in this work that is based on the simultaneous assessment of similarity between signals in the time and frequency domains. This procedure differs from traditional methods, such as correlation or mean square error analysis, which perform operations from the perspective of the time or frequency domain separately. Thus, in the context of statistical modelling, the methods proposed for analyzing spectral complementarity make it possible to quantify the "importance" of forecasts on different time scales and time instants. In the case studies, forecasts were evaluated for wind speed, solar irradiance and power generated at wind and solar power plants in Brazil's northeast region. The results showed important improvements obtained from the proposed modelling platform. In general, the new models based on spectral complementarity significantly improved the overall quality of the forecasts. The combination models showed an improvement in accuracy of 20% on average and up to 50% compared to the other models evaluated.

Keywords: Wind energy; Solar energy; Operational forecasting; Statistical downscaling; Model combination; Spectral complementarity.

LISTA DE ABREVIATURAS E SIGLAS

AN	Análogos
CBCL	Combinação baseada em Clustering
CBL	Combinação Linear
CL	Clustering
DSR	Downward Shortwave Radiation
EMD	Empirical Mode Decomposition
EN	Elastic Net
ENR	Elastic Net Regression
FFT	Fast Fourier Transform
GCM	General Circulation Model
GFS	Global Forecasting System
IBL	Interpolação Bilinear
MLR	Multilinear Regression
MRMR	Minimum-Redundancy-Maximum-Relevance
PCA	Principal Component Analysis
SCW	Spectral Complementarity Weights
SSWT	Synchrosqueezed Wavelet Transform
STFT	Short Term Fourier Transform
SWM	Stepwise Regression “Modificado”
SWR	Stepwise Regression
TWC	Transformada Wavelet Contínua
WRF	Weather Research and Forecasting Model

SUMÁRIO

1	INTRODUÇÃO	14
1.1	Caracterização do problema	14
1.2	Objetivos	16
2	CONCEITOS PRELIMINARES.....	19
2.1	GCMs: Modelos de Circulação Geral da Atmosfera	19
2.2	Técnicas de aumento na resolução espacial (<i>downscaling</i>)	23
2.3	Modelos baseados em classificação e regressão de padrões sinóticos.....	26
2.4	Técnicas de pré-processamento dados e otimização de modelos.....	28
2.5	Relevância, redundância e complementaridade entre preditores	32
2.6	Complementaridade espectral: análise no domínio da frequência.....	34
3	REVISÃO BIBLIOGRÁFICA.....	39
4	METODOLOGIA E MODELOS	47
4.1	Inicialização	49
4.2	Configuração Geral (CG).....	50
4.3	Algoritmos de modelos	53
4.3.1	IBL – Interpolação Bilinear	53
4.3.2	MLR – Regressão Multilinear	55
4.3.3	PCA+MLR – Regressão Multilinear de Componentes Principais	57
4.3.4	AN – Regressão Multilinear de Análogos	58
4.3.5	ENR – “ <i>Elastic Net Regression</i> ”	60
4.3.6	SWR, SWM – “ <i>Stepwise Regression</i> ”	61
4.3.7	MR – Máxima Relevância Mínima Redundância	62
4.3.8	CL – Classificação e Agrupamento de Padrões Sinóticos	63
4.4	Modelos baseados em complementaridade espectral.....	67
4.4.1	Complementaridade espectral global	68
4.4.2	Exemplo: complementaridade espectral global.....	70
4.4.3	Complementaridade espectral local	74
4.4.4	Exemplo: Complementaridade espectral local	77
4.5	Parametrização de Modelos	80
4.6	Teste: Avaliação de resultados.....	87
5	RESULTADOS	91
5.1	Previsões para velocidade do vento.....	96
5.2	Previsões para Irradiância Global Horizontal (GHI).....	103
5.3	Previsões para geração de potência eólica	108

5.4	Previsões para geração de potência solar fotovoltaica	113
5.5	Resultados Gerais.....	116
6	CONCLUSÕES.....	124
7	PERSPECTIVAS DE DESENVOLVIMENTOS FUTUROS.....	128
	REFERÊNCIAS.....	130

1 INTRODUÇÃO

O aumento da participação das fontes renováveis nas últimas décadas é notável não apenas no Brasil, como também de maneira global (LEE, J.; ZHAO, F., 2021). No Brasil, as políticas de incentivo à geração distribuída impulsionaram a busca pela integração das fontes renováveis na matriz elétrica nacional (REN 1059/2023). Naturalmente, em decorrência da natureza intermitente de variáveis atmosféricas de interesse como o vento e a irradiância solar, surgem desafios à transição para um sistema que suporte o despacho em larga escala de fontes renováveis (e.g., geração de energia solar fotovoltaica e eólica). Em síntese, o aumento da previsibilidade relacionada a variáveis atmosféricas confere maior estabilidade ao processo geral de tomada de decisão no setor de geração de energia elétrica por meio de fonte renováveis (INMAN; PEDRO; COIMBRA, 2013; KIRCHMEIER; LORENZ; VIMONT, 2014; LOPES *et al.*, 2007; ZHOU *et al.*, 2010). Nesse sentido, a crescente participação de fontes renováveis deve ser apoiada por procedimentos de modelagem e descrição do comportamento de variáveis de interesse para o setor (variáveis atmosféricas ou diretamente relacionadas à operação de centrais).

Para a introdução dos temas específicos do trabalho de tese ora apresentado, divide-se a primeira seção nas seguintes subseções: Caracterização do problema (1.1) e Objetivos (1.2). Na primeira seção, faz-se uma avaliação do contexto atual relacionado com métodos voltados para a previsão do comportamento de variáveis de interesse para os setores de energia eólica/solar.

1.1 Caracterização do problema

A descrição do comportamento das variáveis atmosféricas de interesse costuma recorrer a simulações numéricas produzidas por modelos de circulação geral da atmosfera (GCM – General Circulation Model), os quais costumam cobrir todo o globo terrestre por períodos da ordem de dezenas de anos (e.g., ERA-5, MUÑOZ-SABATER *et al.*, 2021; GFS, NCEP, 2023). Entretanto, os GCMs costumam apresentar baixas resoluções espaciais, uma vez que se torna proibitivo em termos de custo computacional a realização de simulações globais do comportamento atmosférico com resoluções espaciais mais altas que aquelas tipicamente oferecidas pelos centros de previsão meteorológica. Portanto, técnicas de aumento de resolução

espacial, ou *downscaling*, são utilizadas com vistas à obtenção de um maior detalhamento da variabilidade espacial dos estados atmosféricos. Tais modelos podem ser de natureza analítica (dinâmica), empírica (estatística) ou híbrida. O primeiro grupo de modelos se baseia em soluções numéricas para as equações das leis de conservação que regem a dinâmica atmosférica. O segundo, por sua vez, se realiza por meio do ajuste empírico entre dados observacionais e saídas de modelos atmosféricos. Por fim, os modelos híbridos possuem características mistas entre os dois primeiros grupos.

Como a maior parte dos estudos sobre técnicas de *downscaling* estatístico são voltados para a descrição ou previsão de variáveis ambientais (precipitação, temperatura, velocidade do vento em superfície etc.; BAGHANAM *et al.*, 2020), novos estudos têm centrado a atenção em aplicações voltadas para previsão de potência gerada em centrais eólicas e solares. Por um lado, sinais de potência dependem diretamente de variáveis ambientais como velocidade do vento (e.g., geração de potência eólica) ou irradiância solar (e.g., geração de potência fotovoltaica), por outro lado, deve-se considerar, ainda, variações de ordem eletro-eletrônica associadas com o funcionamento do sistema de geração (i.e., perturbações alheias ao comportamento atmosférico). Por essa razão, estudos diversos têm dado ênfase a modelos com maior capacidade de assimilação de informações, como técnicas baseadas em inteligência artificial.

Em face ao grande número de técnicas disponíveis, os modelos baseados em combinação (*ensemble prediction*) têm como vantagem principal a redução no risco associado à inabilidade dos modelos de *downscaling* estatístico em circunstâncias particulares (e.g., alta redundância e baixa relevância no conjunto de previsões usadas na como entrada dos modelos), conferindo robustez e acurácia às soluções combinadas frente às individuais. As principais técnicas adotadas consistem em bases homogêneas de modelos (e.g., *bagging ensemble*; *boosting ensemble*; ZHANG, Yuzhen; LIU, J.; SHEN, W., 2022). Ao longo dos últimos anos, no entanto, é possível observar um crescente número de trabalhos publicados sobre estratégias de combinação aplicadas a bases heterogêneas de modelos (*stacked ensemble*), as quais conferem aprimoramentos a partir da associação entre técnicas desde as mais simples, até mesmo aquelas que representam o estado-da-arte (FU *et al.*, 2023; NAITAM; MEGHANA; SRIVASTAV, 2023). No entanto, reporta-se que o aprimoramento provido pela combinação de modelos baseada em *stacked ensemble*

é sensível à diversidade e relevância, ou, em síntese, à complementaridade oferecida pelo conjunto de modelos combinados (ZHANG, Yuzhen; LIU, J.; SHEN, W., 2022). Além disso, a complementaridade é um fator fundamental para que se preserve a robustez dos modelos frente à redução na “complexidade” (ou número de parâmetros) dos modelos (KUMAR; CHEN, M., 2015). Ressalta-se que, neste trabalho, o conceito de complementaridade é aplicado em novas proposições para a quantificação da relevância e redundância entre modelos.

Para melhor adequar o conjunto de dados disponíveis ao treinamento de modelos estatísticos a um determinado modelo estatístico (e.g., combinação baseada em *stacked ensemble*), normalmente se aplicam métodos voltados para o pré-processamento de dados). Tais técnicas permitem maior detalhamento sobre as características intrínsecas do comportamento previsto, bem como sobre relações de complementaridade, as quais podem ser obtidas a partir de análise sobre potenciais similaridades entre a informação disponível e a informação desejada. Nesse sentido, há diversos estudos que demonstram procedimentos para a avaliação detalhada sobre as estruturas de covariâncias entre diferentes variáveis, tal como as técnicas baseadas em análise espectral (BUTTKUS, 2000). A partir de projeção para o domínio da frequência, base sobre a qual se descreve a estrutura de variações de um sinal qualquer em termos de ângulo de fase, frequência e amplitude de variação, diversos aspectos da relação entre variáveis podem ser ressaltados. No entanto, no que diz respeito à extração de atributos confiáveis a partir do domínio da frequência, ainda há desafios associados a modelos estatísticos aplicados à previsão operacional (a partir de saídas de modelos de larga escala).

1.2 Objetivos

A metodologia proposta se baseia em estratégia de *stacked ensemble*, a qual visa combinar diferentes camadas de modelos de *downscaling* estatístico e de combinação para o aprimoramento de previsões inicialmente providas por modelos de larga escala. O objetivo geral do presente trabalho visa o atendimento de demandas diversas no âmbito de geração de energia eólica/solar (e.g., melhor reprodução de valores extremos, do valor médio ou de variações em curto prazo). Mais especificamente, o objetivo do trabalho contempla três principais contribuições para

as práticas de previsão em “curto prazo” (até 48 horas à frente) da velocidade do vento, irradiância solar e potência gerada em centrais eólica e solar:

- i) Desenvolvimento e avaliação de metodologia integrada com três camadas de modelos, sendo a primeira voltada para a seleção de domínio (i.e., seleção de saídas de modelos de larga escala para pontos em torno do local de interesse), a segunda voltada para o *downscaling* estatístico (i.e., ajustes regressivos entre o comportamento em larga escala e observações), e a terceira, por fim, voltada para a combinação de previsões produzidas por modelos de *downscaling* estatístico;
- ii) Desenvolvimento de métodos objetivos de pré-processamento de dados para a configuração automática dos modelos de *downscaling* estatístico e de combinação com base na aplicação de índices de qualidade diversos em associação com técnicas de otimização supervisionada (i.e., orientadas pela qualidade das previsões frente aos valores observados);
- iii) Desenvolvimento e avaliação de novos modelos de *downscaling* estatístico e de combinação a partir de métodos de pré-processamento para modelos baseados em “complementaridade espectral”, conceito aplicado à avaliação de similaridade entre sinais inicialmente abordado em Perruci (2018), o qual é expandido no presente trabalho.

A hipótese avaliada neste trabalho, em síntese, se fundamenta em afirmação de que as soluções obtidas através da metodologia desenvolvida produzam previsões simultaneamente robustas e acuradas, conferindo ampla aplicabilidade, porém, preservando a capacidade de especialização dos modelos para casos específicos. Em mais detalhes, compõe-se a hipótese avaliada neste trabalho a partir de afirmações sobre as três partes do objetivo supracitado:

- i) O acoplamento entre as camadas mencionadas no primeiro item dos objetivos permite a obtenção de soluções acuradas, uma vez que as informações mais relevantes sobre área com ampla cobertura espacial (em torno do local de interesse) são contempladas, em múltiplas instâncias por modelos diversos, incluindo os mais aptos à generalização ou à especialização de eventos observados.
- ii) Já sobre o segundo item dos objetivos supracitados, avalia-se a hipótese de que as técnicas desenvolvidas neste trabalho de tese, voltadas para

pré-processamento de dados com base em otimização supervisionada, permitem mitigar riscos associados às decisões realizadas com base em dados sobre período distinto daquele que se pretende aplicar um modelo estatístico de previsão, conferindo maior robustez aos resultados finais.

- iii) Em relação ao terceiro item dos objetivos, neste trabalho, avalia-se a hipótese de que informações sobre a relevância e redundância entre previsões podem ser assimiladas a partir de atributos baseados em complementaridade espectral. Em tese, tais atributos conferem robustez e acurácia à combinação de modelos estatísticos em função de maior capacidade de detalhamento sobre a relação entre variáveis.

Em outro aspecto importante, o trabalho será realizado por meio de estudos de caso para localidades situadas na Região Nordeste do Brasil, região que concentra o maior potencial de geração nas modalidades eólica e solar em todo o país, motivo pelo qual também é a região que concentra a maior potência instalada (ABEEÓLICA, 2021; EPE, 2022; NASCIMENTO, 2017).

O trabalho está dividido em sete seções principais: 1) Introdução; 2) Conceitos Preliminares; 3) Revisão Bibliográfica; 4) Metodologia e Modelos; 5) Resultados; 6) Conclusões; 7) Perspectivas de trabalho futuros.

2 CONCEITOS PRELIMINARES

Esta seção está dedicada à apresentação de conceitos importantes para a contextualização da revisão de literatura e metodologia propostas (Seções 3 e 4). Na Seção 2.1, serão apresentados os modelos de circulação geral da atmosfera (GCMs), os quais são desenvolvidos por renomados centros de previsão meteorológica e disponibilizados gratuitamente para fins diversos (e.g., ECMWF, NASA, NOAA). A Seção 2.2 apresenta as técnicas de *downscaling*, adotadas para o aumento na resolução espacial dos modelos atmosféricos de larga escala. A Seção 2.3, por sua vez, discute a associação de técnicas de agrupamento (ou classificação) e regressão, as quais são frequentemente adotadas no âmbito de *downscaling* estatístico. Na Seção 2.4, apresenta-se uma discussão sobre técnicas pré-processamento de dados e de otimização na busca por soluções mais confiáveis, bem como a respeito da motivação para a aplicação de tais técnicas sobre dados de previsão numérica do tempo. Na seção 2.5, por sua vez, são resumidos os conceitos de relevância, redundância e complementaridade considerados neste trabalho de tese. Por fim, na Seção 2.6, apresentam-se conceitos relacionados com a análise espectral, relacionando-os com o objetivo deste trabalho.

2.1 GCMs: Modelos de Circulação Geral da Atmosfera

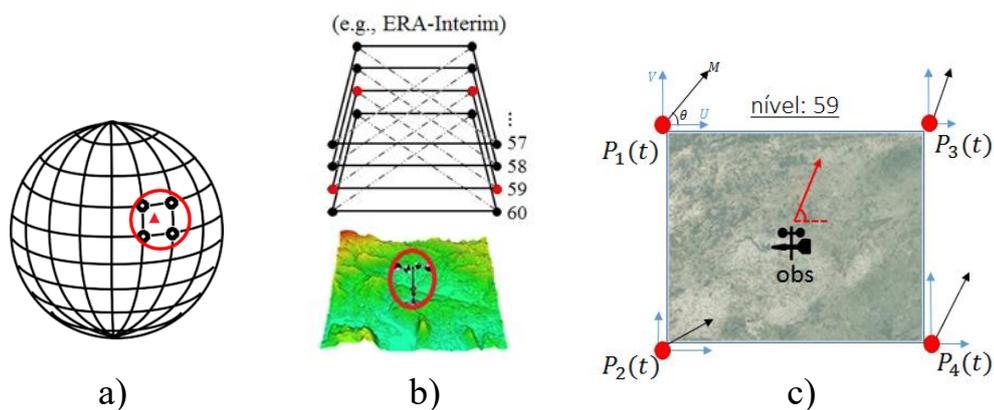
O comportamento de variáveis atmosféricas é frequentemente estudado por meio de abordagem analítica (modelos dinâmicos). De modo geral, o procedimento se baseia na solução de equações derivadas de leis de conservação de massa, energia e momento que regem um determinado sistema (e.g., AL-YAHYAI; CHARABI; GASTLI, 2010). A modelagem dinâmica da atmosfera pode ser convenientemente classificada em três grupos principais, definidos em relação à cobertura e escalas de variação espacial (e temporal) do domínio considerado: a) modelos de circulação geral da atmosfera¹ (macroescala); e.g., *General Circulation Models* (GCMs; DEE *et al.*,

¹ Resolução espacial típica da ordem de centenas de quilômetros. Normalmente cobre todo o globo terrestre e um período de dezenas de anos.

2011; KALNAY *et al.*, 1996); b) regional² (mesoescala); e.g., *Weather Research and Forecasting Model*, WRF; (CARVALHO *et al.*, 2014) c) local³ (microescala); e.g., modelos de camada limite planetária (LANDBERG; WATSON, 1994); fluidodinâmica computacional (SUMNER; WATTERS; MASSON, 2010); modelo de espectro solar SMARTS2 (GUEYMARD, 1995).

Os modelos de circulação geral da atmosfera (ou GCMs) simulam estados atmosféricos para todo o globo terrestre (Figura 1a) definidos através de sistema multivariado (e.g., temperatura, pressão, altura geopotencial, componentes meridional e longitudinal do vento, umidade relativa do ar, etc.) por meio de diagnósticos (análise ou reanálise) e prognósticos (previsão), os quais normalmente são disponibilizados em diversos níveis verticais, como exemplifica a Figura 1b. A Figura 1c, por fim, ilustra um estado atmosférico representado pela componentes U e V da velocidade do vento nos quatro pontos mais próximos a um hipotético local de interesse (“obs”), em um determinado nível vertical. Os pontos “P” indicam o estado composto pelo conjunto total de variáveis consideradas na definição de um padrão atmosférico para um determinado problema (note-se que relação de dependência de tal conjunto foi omitida para que se destaque a variabilidade temporal de tal conjunto em relação à dimensão de tempo t).

Figura 1 – Modelos de Circulação Geral da Atmosfera



² Resolução espacial típica da ordem de dezenas de quilômetros. Domínio centrado em regiões de interesse (e.g., estados ou regiões de países) e períodos específicos (normalmente alguns anos).

³ Resolução típica de centenas de metros. Sobre microrregião que cerca um local (ou ponto) de interesse (e.g., estação anemométrica ou solarimétrica, centrais de geração etc.).

Representação de modelos de circulação geral da atmosfera. a) Vista global, estratificação horizontal de domínio; b) Vista Lateral, estratificação vertical (níveis de modelo); c) Vista Superior; malha contendo os pontos mais próximos ao local de interesse (em um determinado nível). Fonte: autor.

Dados do tipo “análise” são essencialmente diagnósticos do comportamento atmosférico. O procedimento se baseia em processos de assimilação de dados observacionais que estão distribuídos de forma irregular no globo. A partir deste processo, uma nova base é produzida considerando um novo sistema distribuído de forma regular no espaço. Entretanto, ressalta-se, ainda, que há de se considerar as mudanças frequentes de versão de modelos “operacionais” (atualizados em tempo real), sendo este também um fator heterogeneidade em bases de dados de modelos atmosféricos. Portanto, dados de “reanálise” são tipos especiais de análise, para os quais se fixam as melhores parametrizações do modelo, o qual é executado para domínio regular com ampla cobertura temporal (ordem de dezenas de anos). Ao evitar descontinuidades ocasionadas das mudanças no procedimento, reanálises produzem sistemas homogêneos, o que tendem a produzir erros menores, entretanto, não estão disponíveis em tempo real em razão da necessidade de informações em tempo diferido para a produção de novas estimativas, além do alto esforço computacional envolvido (e.g., ERA-INTERIM, Merra, e Reanalysis 2).

Dados de previsão resultam de prognósticos realizados por modelos atmosféricos, os quais normalmente disponibilizam previsões de estados atmosféricos em tempo real. Por esse motivo, dados de previsão costumam ser aplicados para fins operacionais (e.g., GFS). Assim como dados de análise, os modelos dados de previsão operacional estão sujeitos às mudanças nas parametrizações dos modelos de larga escala. Porém, espera-se que os erros associados a aplicação desses dados sejam superiores aos dados de análise, tendo em vista o aumento natural da incerteza em relação aos estados futuros da atmosfera.

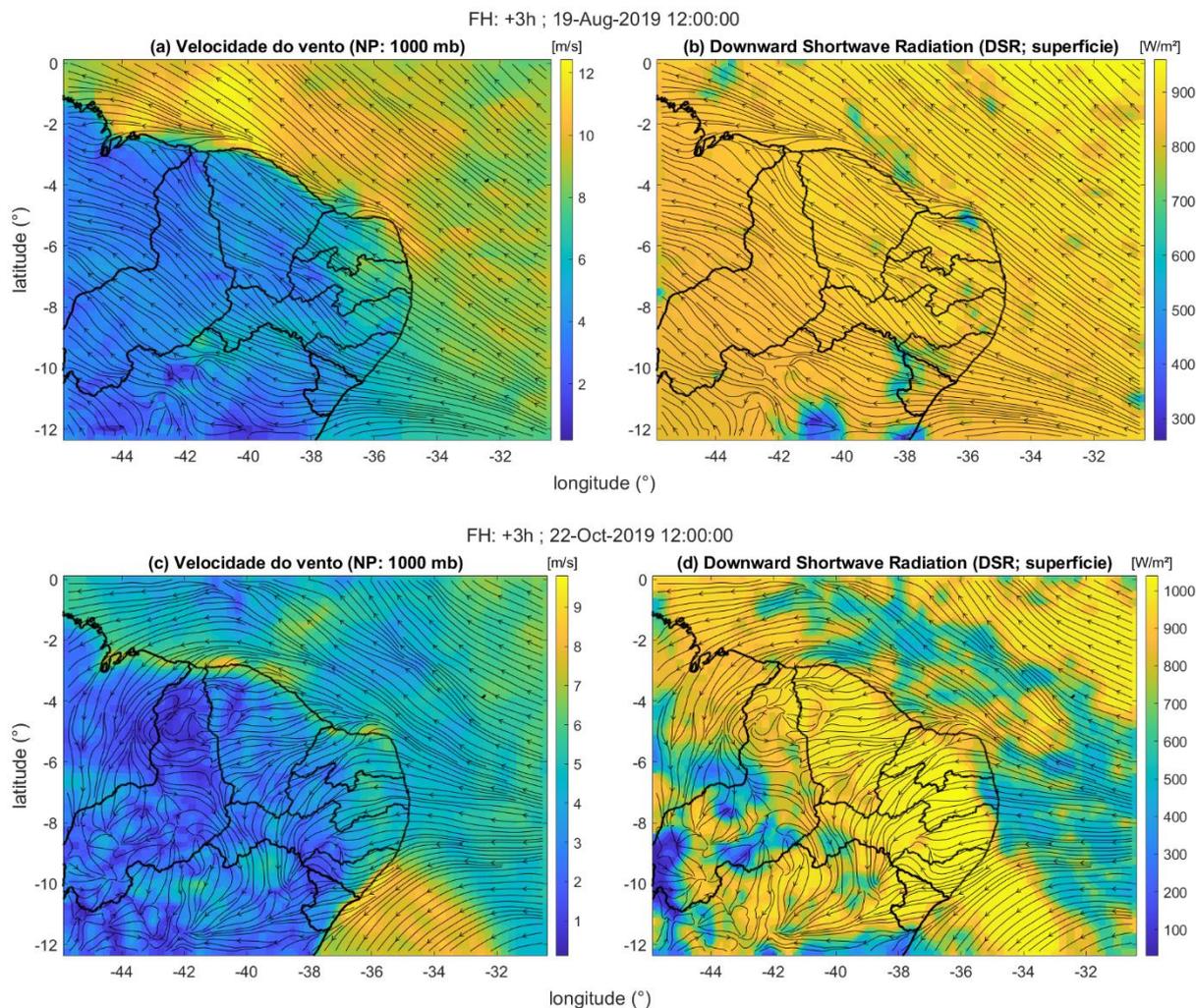
A variabilidade espacial e temporal no comportamento de grandezas atmosféricas costuma se intensificar com a aproximação ao nível superficial (da água ou solo), naturalmente elevando a incerteza sobre o subsistema. Em relação ao deslocamento de massas de ar, sendo essencialmente uma condição de contorno no sistema atmosférico, o nível superficial costuma apresentar maiores perturbações no comportamento do vento devido a fatores como: orografia, ocupação de terreno, efeitos causados por trocas térmicas, entre outros (WINSTRAL; JONAS; HELBIG, 2017). De maneira similar, a irradiância na superfície pode ser influenciada por fatores

como: sombreamento (obstáculos), passagem de nuvens, umidade relativa do ar, etc. (RIVINGTON *et al.*, 2008; TROCCOLI; MORCRETTE, Jean Jacques, 2014). Nesse contexto, incertezas relacionadas à modelagem do vento e irradiância solar (na superfície terrestre ou aquática) podem ser reduzidas através de descrições confiáveis dos fenômenos que se apresentam em regime local.

As limitações no uso de simulações de GCMs, entretanto, estão tipicamente relacionadas com as baixas resoluções produzidas (variabilidades espacial e temporal descritas em “amplos” intervalos). Informações sobre o comportamento de uma variável atmosférica, na microrregião (ou local) de interesse, podem não estar contempladas em malhas macroescalares (em malha com pontos geograficamente distantes entre si). Por esse motivo, simulações derivadas de modelos globais (e.g., GCMs) costumam ser adotadas em procedimentos voltados para o aumento da resolução espacial ou *downscaling*. Nesse procedimento, as simulações derivadas dos GCMs são empregados como entrada para modelos microescalares (dinâmicos e/ou estatísticos), os quais têm por finalidade a descrição de fenômenos pertinentes ao comportamento em regime local de determinadas variáveis atmosféricas (AL-YAHYAI; CHARABI; GASTLI, 2010; HORVATH; BAJIĆ; IVATEK-ŠAHDAN, 2011).

A Figura 2a e a Figura 2b mostram, respectivamente, padrões atmosféricos formado por previsões para horizonte de +3 horas para a velocidade do vento (nível de modelo em camada isobárica de 1000 mb) e de irradiância solar (DSR) para o dia 19-08-2019, às 12:00 (GMT-3). Por sua vez, a Figura 2c e a Figura 2d mostram, os mesmos tipos de padrões atmosféricos para o dia 22-10-2019, às 12:00 (GMT-3). A comparação entre os padrões que se apresentam nesses dias revela perfis mais homogêneos para o primeiro, especialmente em relação à irradiância solar e às linhas de corrente do da velocidade do vento. Tais exemplos servem para ilustrar que, a depender do local estudado, bem como o instante para o qual se realize a previsão de uma variável de interesse, a informação disponibilizada por modelos de larga escala pode ser de alta ou baixa variabilidade espaço-temporal. Entretanto, é importante ressaltar que a variabilidade do comportamento atmosférico descrito em larga escala deve estar relacionada com o comportamento local da variável de interesse, o que motiva a avaliação de tal conjunto de informações em termos da relevância e da redundância para a descrição de sinais observados.

Figura 2 – Exemplo de padrão sinóptico representado por saídas de previsões operacionais de modelo de larga escala (GFS).



As linhas de corrente mostradas no gráfico dizem respeito ao campo vetorial formado a partir das componentes zonal (U) e meridional (V) da velocidade do vento. Resolução espacial: 0,25° de latitude/longitude ou aproximadamente 27 km. Fonte: autor.

Na próxima seção (2.2), discorre-se sobre as técnicas de aumento de resolução espacial, ou *downscaling*, aplicadas a saídas de modelos de previsão numérica do tempo (Figura 2) para a descrição do comportamento observado para a variável de interesse em resolução espacial local (ordem de 100 metros).

2.2 Técnicas de aumento na resolução espacial (*downscaling*)

As técnicas de *downscaling* existentes costumam ser classificadas em dois grupos principais: estatístico e dinâmico. O primeiro grupo introduz efeitos locais por meio das observações, relacionando-as com simulações macroescalares. O segundo, por sua vez, consiste em modelos fundamentados em soluções analíticas que descrevem o comportamento local por meio de soluções numéricas para equações

baseadas nas leis que regem o sistema atmosférico. Além disso, é possível definir um terceiro grupo intermediário (ou híbrido) uma vez que as limitações e aptidões das abordagens estatística e dinâmica são potencialmente complementares (e.g., WALTON *et al.*, 2015).

As abordagens dinâmicas para as técnicas de *downscaling*, podem ser classificadas em dois grupos principais: a modelagem da camada limite planetária (PBL, *Planetary Boundary Layer*, e.g., LANDBERG; WATSON, 1994), que consiste no acoplamento entre macroescala (GCM, cuja resolução espacial é da ordem de centenas de quilômetros) e a microescala (ordem de centenas de metros); e o “aninhamento” de modelos regionais (*nested models* ou *Limited Area Models*, LAM; e.g., WILBY; WIGLEY, 1997) em saídas do GCM, i.e., modelos que se utilizam das saídas de um GCM como condições de contorno para descrever o comportamento do vento na mesoescala (cuja resolução é da ordem de dezenas de quilômetros), escala intermediária entre a macro e microescala. Uma vantagem associada ao *downscaling* dinâmico está relacionada à independência de observações para a realização de estimativas, uma vez que recorre apenas a relações físicas entre variáveis simuladas por modelos de circulação geral da atmosfera ou de modelos regionais (Seção 2.1). A simulação, entretanto, costuma apresentar maiores erros em relação às técnicas estatísticas.

No que diz respeito à abordagem estatística para a simulação do comportamento local da variável de interesse, é importante diferenciar as técnicas de *downscaling* estatístico de técnicas baseadas em séries temporais. A principal diferença entre os modelos baseados em *downscaling* estatístico e séries temporais pode ser entendida pelos tipos de preditores empregados em cada grupo de técnicas. Os modelos autorregressivos aplicam componentes temporais obtidas a partir da defasagem (“*lag*”) de uma série observacional, a qual é especificada com o objetivo de descrever a evolução temporal do sinal, ou passos futuros em função de estados anteriores. Nessa estrutura, é suficiente que os modelos baseados em séries disponham apenas dos períodos observados anteriores para que se realizem previsões (ZIEGEL; ENDERS, 1995; SSEKULIMA *et al.*, 2016).

As técnicas de *downscaling*, por outra parte, são mais flexíveis em relação às aplicações, uma vez que adotam geralmente apenas simulações numéricas como entrada, cuja disponibilidade independe de observações no local. Entretanto, observações são necessárias apenas para o treinamento dos modelos, os quais

associam simulações numéricas ao comportamento observado. Em referência à ciência econométrica, de forma geral, as técnicas de *downscaling* estatístico fazem uso de dados em “painel” (*panel data*; e.g., DAMODAR, 2004), os quais dispõem os diferentes preditores na forma de colunas em uma tabela cujas linhas determinam o instante de tempo (“*cross-sectional data*”). A esse respeito, normalmente, as técnicas regressivas adotadas para fins de *downscaling* associam informações concomitantes (ou padrões sinópticos⁴) entre diferentes variáveis predictoras e o preditando (observações), bem como suas evoluções temporais por meio da concatenação cronológica de estados. Nesse caso, normalmente as variáveis que compõem o espaço de preditores são derivadas de simulações (diagnóstico ou prognóstico) de modelos macroescalares ou microescalares.

Em relação às estratégias de *downscaling* voltadas para a descrição do comportamento do vento e irradiância solar na superfície, existe um extenso número de proposições na literatura, sendo tipicamente reportados menores erros associados em comparação com a abordagem dinâmica (INMAN; PEDRO; COIMBRA, 2013; WILBY, R. L. *et al.*, 1998; WILKS, 2011). Isso normalmente ocorre devido ao ajuste de modelos estatísticos, os quais apreciam informações sobre o regime local obtidas diretamente de dados observados. Após o ajuste dos modelos estatísticos, simulações numéricas produzidas por GCMs podem ser aplicadas a períodos para os quais não se possui observações. Assim, ao serem associadas ao comportamento observado, são capazes de simular o comportamento local a partir do comportamento macroescalar (PERRUCCI, 2018; WILBY, R. L. *et al.*, 1998; WILKS, 2011). Em ferramentas de *downscaling* estatístico (os quais podem incluir múltiplas estratégias), é possível combinar os resultados de um grupo de modelos de *downscaling* estatístico com vistas ao melhor aproveitamento de aptidões de abordagens distintas. A combinação pode ser realizada priorizando um aumento na acurácia⁵ ou robustez⁶ dos modelos (SÁNCHEZ, 2008; TIMMERMANN, 2006).

⁴ Instantâneo do conjunto de variáveis atmosféricas concomitantes e distribuídas espacialmente, as quais definem um estado (ou instantâneo) do sistema considerado.

⁵ Combinação voltada para o aprimoramento do resultado em relação às entradas e a um determinado critério de avaliação.

⁶ Combinação que visa preservar, de forma adaptativa, a solução (entrada) que apresenta os melhores resultados em relação a um critério de avaliação.

No que diz respeito ao presente trabalho de tese, em última instância, serão avaliados modelos baseados em regressão para o *downscaling* de previsões geradas por modelo de larga escala (GFS). Da mesma forma, modelos regressivos são aplicados à combinação das previsões geradas pelos modelos de *downscaling* estatístico. Nesse sentido, a próxima seção (2.3) discorre sobre os principais tipos de arquitetura de parâmetros aplicados a modelos baseados em regressão no contexto de *downscaling* estatístico.

2.3 Modelos baseados em classificação e regressão de padrões sinóticos

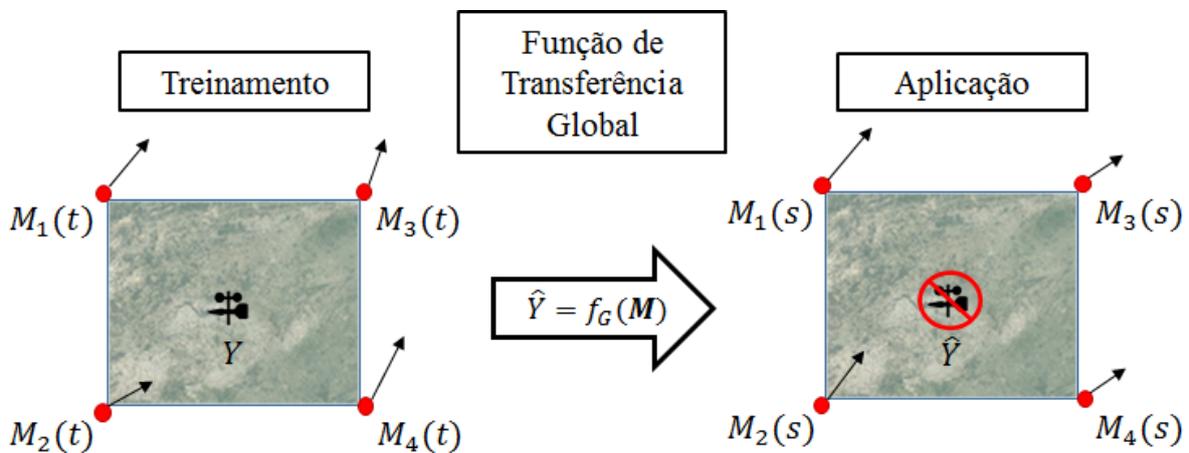
Técnicas de regressão têm como objetivo o estabelecimento de parâmetros para um determinado modelo estatístico. A arquitetura de parâmetros em modelos estatísticos pode ser dividida em (WILBY, R. *et al.*, 2004): regressão global, relação obtida a partir de um único ajuste realizado para todas as amostras disponíveis para a inferência de parâmetros (Figura 3); regressões locais, relações obtidas a partir de diferentes ajustes realizados para subconjuntos de amostras disponíveis para a inferência dos parâmetros (Figura 4; ENKE; SPEKAT, 1997). Note-se que, em relação à arquitetura de parâmetros, os termos “global” e “local” nesta seção se referem ao período adotado para o ajuste, estando relacionado com a variabilidade temporal, diferentemente do significado atribuído em outras partes do texto, em que, estando associadas ao regime em que se avalia o comportamento de uma variável, refere-se à variabilidade espacial dos dados atmosféricos.

O ajuste de parâmetros global busca resumir a relação entre as variáveis ao longo de todo o período disponível. Na Figura 3, um período de treinamento (t) é escolhido para relacionar a magnitude do vento (M) nos quatro pontos mais próximos às observações, obtidas no local de interesse, por meio de função de transferência única (f_G). Ao ser ajustado o conjunto de parâmetros da função f_G , essa mesma relação pode ser aplicada para estimar a magnitude do vento local através de simulações do GCM em um instante de tempo ou período distinto s (independentemente da disponibilidade de observações).

A Figura 4 apresenta o caso de ajustes locais (ou especializados), o qual é bastante similar ao caso discutido na Figura 3 (ajuste global). A diferença entre as duas abordagens se dá em relação ao número de funções de transferências

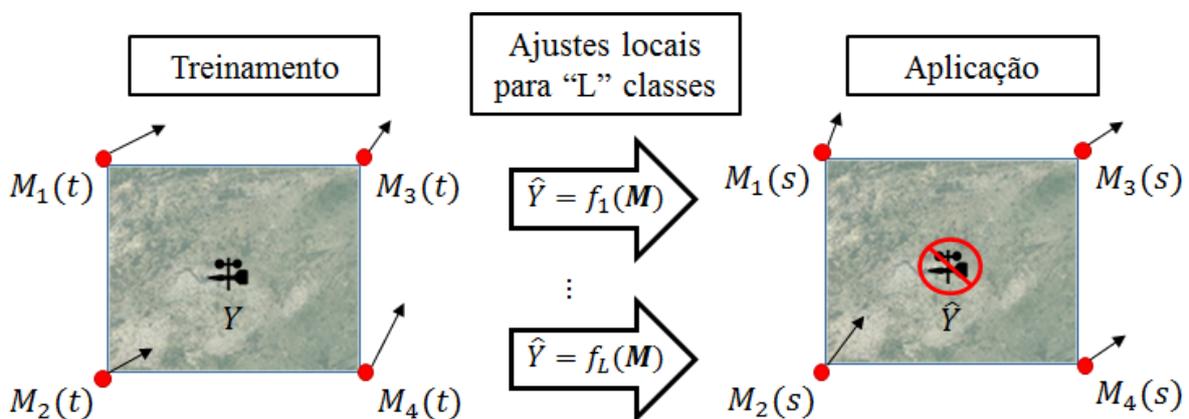
ajustadas. Nesse caso, técnicas de agrupamento são empregadas para caracterizar L classes, ou conjunto de instantâneos para os quais são ajustadas funções de transferência especializadas. A estimativa \hat{Y} para um instante s pode ser obtida através da classificação do padrão sinóptico $\mathbf{M}(s) = [M_1, M_2, \dots, M_m]_s$ em relação à distância para os centroides dos L grupos (formados durante o processo de treinamento). A seleção da função de transferência local pode ser realizada por meio da minimização da distância entre os centroides (em espaço m -dimensional) dos grupos formados e o padrão no instante s .

Figura 3 – Arquitetura de Parâmetros Globais.



Fonte: autor.

Figura 4 – Arquitetura de Parâmetros Locais.



Fonte: autor.

Na próxima subseção (2.4), considerando o contexto das aplicações para os setores de energia eólica/solar, realiza-se uma discussão sobre aspectos importantes

associados aos métodos empíricos – como aqueles baseados em pré-processamento e otimização – e as características das previsões numéricas do tempo produzidas por modelos de larga escala.

2.4 Técnicas de pré-processamento dados e otimização de modelos

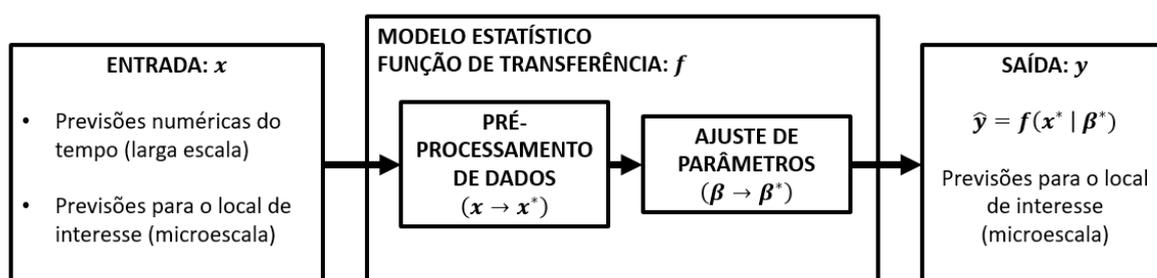
O comportamento de variáveis atmosféricas possui aspectos determinísticos (relacionados com leis físicas que regem os fenômenos e relações estatísticas significativas) e estocásticos (incertezas intrínsecas ao sistema atmosférico, WANG, H. Z. *et al.*, 2016). Naturalmente, a previsibilidade sobre o comportamento dessas grandezas está relacionada com a variabilidade espacial (localização geográfica) e temporal, cuja representação se dá em diferentes escalas (ou resoluções⁷). Tal como se descreve por meio de modelos de larga escala, o comportamento atmosférico se expressa por meio da sucessão de padrões sinópticos cujos perfis se apresentam com aspecto homogêneo ou heterogêneo (em função dos processos atmosféricos que atuam na região), sendo a partir de tal sucessão que são inferidas as relações entre as saídas de modelos de larga escala e o comportamento local observado. Neste contexto, deve-se atentar ao fato de que, por um lado, as previsões numéricas do tempo (dados de entrada) possuem determinada habilidade preditiva sobre o comportamento local da variável em questão (e.g., velocidade do vento). Por outro lado, a previsibilidade inerente ao comportamento local da variável constitui fator determinante para a acurácia de modelos de previsão (KUMAR; CHEN, M., 2015). A acurácia e robustez de modelos estatísticos, portanto, se constituem a partir da relação entre as variáveis assimiladas pelo modelo adotado, bem como na manutenção de tal relação em períodos futuros.

De maneira geral, neste trabalho, ao se referir ao *downscaling* estatístico e à sua combinação, o termo “modelo” representa um “bloco” da metodologia que engloba uma função de transferência e algoritmos de pré-processamento de dados e de ajuste de coeficientes, como indicado de forma simplificada na Figura 5. O bloco de pré-processamento tem como objetivo a seleção ou transformação de variáveis de entrada

⁷ Resoluções espaciais: menor distância entre dois pontos em relação às direções de latitude, longitude e altitude; resolução temporal: menor intervalo de tempo em série temporal (i.e., intervalo de integração).

(x) para um ou mais subconjuntos x^* , formado(s) no intuito de melhor adequar os dados disponíveis à função de transferência f (como descreve na Seção 2.3). Os parâmetros associados à função de transferência (β) são ajustados por meio soluções numéricas ou estatísticas, das quais resultam o conjunto de parâmetros otimizados β^* . Por fim, a saída do modelo pode ser obtida da aplicação da função de transferência, com os parâmetros ótimos, sobre o conjunto de variáveis de entrada pré-processado: $\hat{y} = f(x^* | \beta^*)$. Note-se que – para os modelos mais simples – o bloco de pré-processamento pode ser omitido, de forma que a função de transferência é ajustada diretamente sobre o conjunto de entrada original: $\hat{y} = f(x | \beta^*)$.

Figura 5 – Fluxograma simplificado de modelos de *downscaling* estatístico e de combinação.



Fonte: Autor.

No contexto de modelos regressivos, as técnicas tipicamente aplicadas para fins de pré-processamento podem ser de três tipos principais: i) não supervisionadas, com critérios fundamentados em propriedades dos dados de entrada; ii) supervisionados, com critérios fundamentados na comparação entre a saída do modelo e a observação; ou iii) semi-supervisionados, os quais combinam técnicas supervisionadas e não supervisionadas. Por outra parte, o ajuste dos parâmetros do modelo é necessariamente realizado de forma supervisionada (e.g., por meio de minimização de gradientes de erro), podendo ser realizado também como parte intrínseca do bloco de pré-processamento (e.g., stepwise regression) (JOVIĆ; BRKIĆ; BOGUNOVIĆ, 2015).

Em relação às variáveis de entrada, normalmente, são adotadas (como entrada para modelos estatísticos) simulações numéricas relativas aos pontos de malha mais próximos do local de interesse. A partir de tais dados, é comum haver baixa variabilidade espacial em malhas de simulações numéricas derivadas de modelos macroescalares, o que pode resultar na escolha de preditores redundantes, sendo necessário o emprego de técnicas de compressão e/ou seleção de variáveis em domínios amplos para evitar problemas de sobreajuste de modelos e

multicolinearidade de entradas (GEORGE; JANAKI; PARAMESWARAN GOMATHY, 2016; LINDERSON; ACHBERGER; CHEN, D., 2004).

Para superar tais desafios, recomenda-se o emprego de séries temporais suficientemente longas para a obtenção de melhores ajustes com modelos regressivos (KUMAR; CHEN, M., 2015; MONAHAN, A. H., 2012). No entanto, a disponibilidade de dados observacionais de centrais eólicas/solares é um desafio às diversas abordagens estatísticas que se propõem para aplicações relacionadas, devido a questões de confidencialidade, tempo de operação etc. Dessa forma, ao se dispor apenas de intervalos “curtos” (i.e., da ordem de um ano), a escolha arbitrária ou inapropriada dos períodos de treinamento e avaliação pode resultar na “incongruência” entre os dados (ou períodos) usados para treinamento e o período pretendido para a previsão em tempo real. Tal incongruência potencialmente implica em vieses nas saídas dos modelos devido a mudanças sazonais no comportamento atmosférico (CHATFIELD, 2000; WILKS, 2011).

Por outra parte, alguns modelos podem se mostrar mais adequados para um problema específico, apresentando maior habilidade ao reproduzir determinadas características estatísticas da série temporal de referência (e.g., variância reproduzida, correlação linear etc.) em razão de sua capacidade de generalização ou especialização. Entretanto, há casos em que tais modelos podem ser subutilizados por não apresentarem desempenho satisfatório em relação ao critério de avaliação adotado, o qual não pode contemplar devidamente outra aptidão do modelo – o que, conseqüentemente, omite informações com potencial explanatório (GOUTTE, 1997). Conseqüentemente, a omissão de informações relevantes sobre a variável alvo em modelos estatísticos pode introduzir erros sistemáticos aos resultados (BUSENBARK *et al.*, 2022; CINELLI; HAZLETT, 2020; WÜTHRICH; ZHU, Ying, 2023)

A associação entre técnicas estatísticas e de aprendizagem de máquina tem sido estudada como solução para os desafios relacionados à modelagem estatística, com vistas ao aproveitamento das informações “favoráveis” e mitigação de efeitos causados por informações “desfavoráveis” à modelagem (FENG *et al.*, 2017; LAJNEF *et al.*, 2015; SALCEDO-SANZ *et al.*, 2017). Isso é possível uma vez que procedimentos de decisão automáticos ou semiautomáticos podem ser estruturados a partir de critérios de decisão objetivos. Tais procedimentos de decisão, ou “algoritmos de otimização” podem ser incluídos em qualquer etapa do procedimento geral de modelagem estatística, e podem ser adotados na identificação de

particularidades do sistema de treinamento de modelos e acionamento de processos que deem um tratamento adequado para casos que requerem técnicas especializadas.

Existem aspectos sobre o comportamento de variáveis atmosféricas que podem demandar maior sofisticação para avaliação e reprodução. Por outro lado, os modelos mais aptos à especialização, como os baseados em regressões não lineares e classificação, apesar de demonstrarem maior habilidade em reproduzir comportamentos atípicos (LYDIA *et al.*, 2016), também estão suscetíveis a problemas de sobreajuste (devido ao seu próprio potencial de especialização), resultando em soluções instáveis de baixa confiabilidade. Por esse motivo, em qualquer caso, há um limiar entre a reprodução adequada e inadequada de comportamentos atípicos, sendo esse um desafio intrínseco à busca por especialização de parâmetros e generalização de modelos (GÓMEZ; ROJAS, 2016).

Além disso, é importante destacar que os aprimoramentos pretendidos devem estar relacionados ao propósito da modelagem, uma vez que aplicações em escalas climatológica (tipicamente voltadas para um período histórico) e meteorológica (geralmente voltadas para estados futuros) costumam dar ênfase a aspectos distintos. Por exemplo, aplicações em escala climatológica costumam dar maior ênfase aos aspectos probabilísticos para fins de estimação de disponibilidade de recurso (informações de longo prazo; (KARTHIKEYA; NEGI; SRIKANTH, 2016; WANG, Jianzhou; HU; MA, K., 2016). Aplicações voltadas para a variabilidade temporal em períodos mais curtos costumam dar maior ênfase à evolução temporal das variáveis, i.e., aspectos determinísticos relacionados com variações curto e curtíssimo prazo (e.g., previsão operacional; TANG; BASSILL, 2018).

Para a seleção de variáveis, é comum a aplicação de técnicas de filtro, as quais são voltadas para a classificação (ordenamento) de variáveis candidatas ao ajuste, a partir da qual são eliminados aquelas que apresentam menor “relevância”. Em modelos estatísticos, a relevância das variáveis de entrada pode ser quantificada por meio de parâmetros que transportam informações sobre a relação de dependência entre um determinado preditor e o alvo (e.g., correlação). Nesse sentido, a classificação de preditores pela relevância é favorável, porém, não garante que todas as variáveis selecionadas tragam contribuições ao ajuste. Isto ocorre devido à incapacidade de técnicas de filtro em eliminar a redundância do sistema, somando-se o fato de que a função de transferência não é avaliada. Contudo, destaca-se em tais

métodos o baixo custo computacional, sendo aplicáveis a conjuntos com grande número de variáveis (CHANDRASHEKAR; SAHIN, 2014; TANG, J.; ALELYANI; LIU, Huan, 2014).

Ao considerar a função de transferência, há um aumento considerável no esforço computacional demandado, uma vez que o ajuste do modelo deve ser repetido para diversos subconjuntos de preditores para que a utilidade do conjunto seja avaliada. Os algoritmos que consideram a utilidade dos preditores, dada uma determinada função de transferência, são conhecidos como *wrappers* ou “empacotadores” (TANG, JILIANG; ALELYANI, SALEM; LIU, 2014). Em geral, o procedimento adotado envolve uma busca exaustiva em iterações que exploram distintas possibilidades de ajuste. Entretanto, ao considerar o custo computacional associado ao ajuste da função de transferência em um elevado número de possibilidades, costuma-se acoplar a tais algoritmos técnicas de filtro e outros métodos de busca (e.g., métodos heurísticos). Sendo assim, um direcionamento adequado à busca pode reduzir o custo computacional associado ao método.

Para o desenvolvimento de métodos robustos que sejam capazes de identificar informações importantes para o treinamento dos modelos, neste trabalho, parte-se da premissa de que a acurácia geral dos modelos estatísticos é favorecida por informações (variáveis/amostras) complementares em relação a aspectos diversos do sinal de referência (observações). Nas próximas seções (2.5 e 2.6), aprofunda-se a discussão sobre os conceitos de relevância, redundância e complementaridade em conjuntos preditores.

2.5 Relevância, redundância e complementaridade entre preditores

Como mencionado na seção anterior, a determinação de conjuntos de variáveis preditoras, bem como a escolha das amostras aplicadas à inferência e otimização de parâmetros configuram etapas essenciais para o treinamento de modelos estatísticos. Para que um modelo estatístico possa capturar diferentes aspectos do preditando (i.e., sinal observado), por um lado é importante que haja informações relevantes no conjunto de entrada, sendo benéfico, nessa perspectiva, que haja um amplo conjunto de variáveis e de amostras à disposição (da ordem de dezenas de variáveis). Por outro lado, informações redundantes tendem a produzir modelos instáveis, uma vez que o aumento no número de parâmetros, que não implica em informação complementar

para o modelo, tende a reduzir a significância estatística dos coeficientes ajustados (WILKS, 2011).

Tipicamente, aplicam-se métricas voltadas para a quantificação da similaridade (ou dissimilaridade) entre variáveis para a identificação de configurações mais apropriadas. A depender do contexto, a similaridade entre variáveis pode estar atrelada: i) à relevância, caso em que uma variável na camada de entrada do modelo é similar ao preditando; ii) à redundância, caso em que duas variáveis na camada de entrada do modelo são similares entre si. Em ambos os casos, tal avaliação normalmente se dá por meio da similaridade entre pares, sendo, por si só, insuficientes para a qualificação objetiva de um grupo de variáveis preditoras em relação ao preditando. Por essa razão, aqui, conceitua-se, a complementaridade entre sinais como a “compensação” entre duas ou mais variáveis relativa à similaridade entre tais variáveis e uma outra variável, a qual se tem por dependente. Dessa forma, pode-se afirmar que variáveis complementares (em relação a um determinado alvo) possuem alta relevância (i.e., similares ao sinal de referência) e baixa redundância (dissimilares entre si). Em suma, a complementaridade evidencia diferentes tipos de relevância (ou razões para que se considere um determinado preditor em um modelo regressivo).

No que se refere aos modelos regressivos, diversos resultados na literatura demonstram que o emprego de conjuntos compostos por preditores complementares em estratégias de *stacked ensemble* confere maior capacidade adaptativa sobre as limitações de soluções individuais em razão de compensação entre aptidões e limitações, permitindo melhor aproveitamento das informações disponíveis. Dessa forma, afirma-se que modelos que combinam preditores complementares produzem saídas com qualidade superior àquelas referentes às soluções individuais (DONG *et al.*, 2021; RIBEIRO *et al.*, 2022; WANG, Jujie *et al.*, 2022; ZHANG, Yuzhen; LIU, J.; SHEN, W., 2022).

Neste trabalho de tese uma nova abordagem é proposta a inferência de complementaridade em conjuntos preditores. Tal abordagem se fundamenta na hipótese de que a qualidade geral das previsões frente aos valores observados pode ser explorada a partir do domínio da frequência, representação alternativa que, no lugar de índice temporal, descreve os sinais em termos de frequência, amplitude e fase, evidenciando assim a estrutura inerente das variações. Na próxima seção (2.6), mais detalhes são fornecidos sobre a aplicação de técnicas de projeção para o

domínio da frequência considerando o contexto no qual o trabalho se insere, bem como o objetivo de avaliação da complementaridade espectral.

2.6 Complementaridade espectral: análise no domínio da frequência

Informações sobre os diversos fenômenos atmosféricos se manifestam em vários “níveis”, ou escalas que indicam a ordem de grandeza das variações em termos espaciais e temporais. Por exemplo, movimentos ciclônicos de massas de ar são fenômenos característicos da macroescala, cujas ordens de grandeza espacial e temporal são de milhares de quilômetros e alguns dias, respectivamente. Também podem ser mencionadas as variações sazonais (e.g., estações do ano) e os ciclos diurnos que se apresentam em variáveis como a velocidade do vento, irradiância solar e temperatura ambiente. O comportamento de tais variáveis em escala local é influenciado por diversos fatores, como a presença de obstáculos, corpos d’água, ocupação do terreno, dentre outros (SILVA, 2003). Por esse motivo, certos padrões cíclicos podem se apresentar de forma sutil ou complexa em função da intermitência ou previsibilidade de tais eventos, sendo dessa forma possivelmente incorporados aos parâmetros dos modelos como parte do ruído inerente aos dados em favor da robustez nos resultados (JAKOB THEMESSEL; GOBIET; LEUPRECHT, 2011; WILBY, R. *et al.*, 2004; ZORITA, Eduardo; STORCH, Hans VON, 1999).

No presente trabalho, avalia-se a hipótese de que as periodicidades características de eventos atmosféricos, tal como são descritos por modelos de larga escala, são importantes para a descrição de variáveis ambientais e de geração de energia eólica e solar (em resolução espacial local). Por essa razão, dedica-se esta seção à apresentação de conceitos e técnicas que são importantes para os métodos de quantificação da complementaridade espectral propostos, que, em tese, proporcionam melhor aproveitamento de informações subjacentes sobre as relações entre sinais atmosféricos. Argumenta-se que tais propriedades são particularmente importantes em se tratando de aplicações para a previsão operacional de variáveis de interesse ao setor de geração de energia eólica e solar. Tal afirmação se sustenta na possibilidade de ausência de períodos suficientemente longos de campanhas de dados observacionais, particularmente no caso de centrais recentes, tipicamente se limitando a períodos da ordem de um ano. Sabe-se, no entanto, que variáveis ambientais como irradiância solar e velocidade do vento apresentam sazonalidade

anual, de forma que períodos inferiores a um ano podem induzir tendência não estacionária na transição entre as amostras disponíveis ao treinamento dos modelos e as amostras relativas ao período de operação.

As transformações para o domínio da frequência permitem a discriminação de componentes periódicos em função de suas frequências características, o que estimula a criação de abordagens que preservam informações sobre a importância de variações evidenciadas na densidade espectral de determinada variável. Uma das técnicas mais aplicadas para essa análise é o operador linear conhecido como transformada de Fourier, que em algoritmo heurístico para o caso discreto recebe o nome de Transformada Rápida de Fourier (FFT, *Fast Fourier Transform*; BERGLAND, 1969; MADDEN; JULIAN, 1971; WELCH, 1967). O algoritmo FFT resulta em números complexos que transportam informações sobre a amplitude e fase de funções de onda harmônica que compõem o sinal.

Uma importante limitação em espectros globais é que se pressupõe um comportamento estacionário da variável em questão, sendo seu uso inadequado para a avaliação de fenômenos transientes (GURLEY, K.; KAREEM, 1999). Existem diversas variáveis que, assim como o vento próximo à superfície local, apresentam características não estacionárias, sobretudo em razão de processos atmosféricos típicos de determinados locais e épocas do ano, como a mudança entre períodos secos e chuvosos. Em outro cenário, quando o conjunto de amostras cobre período inferior a um ano, aspectos relacionados à sazonalidade quando isolados podem configurar comportamentos não estacionários aparentes, tratando-se de fato de uma incongruência entre os regimes pertinentes ao conjunto de amostras de treinamento e o período de teste (posterior). Para contornar esta limitação, foram propostas modificações ao algoritmo da Transformada Discreta de Fourier que resultaram em métodos conhecidos como a Transformada de Fourier de Curto Prazo⁸ (STFT, *Short-Term Fourier Transform*; e.g., (ALLEN, 1977; GRIFFIN; LIM, 1984) e Transformada de Fourier Fracional⁹ (*Fractional Fourier Transform*; e.g., (NARAYANAN; PRABHU, 2003). Em tais métodos, a informação temporal é preservada ao serem considerados

⁸ Técnica que consiste na aplicação da transformada de Fourier em curtos intervalos de tempo, sendo esses considerados estacionários. O método preserva informações da dependência temporal no espectro de frequências, o que permite realizar análises locais de tempo-frequência.

⁹ Generalização da transformada de Fourier clássica pela introdução de um termo arbitrário no expoente complexo que define uma rotação do plano tempo-frequência.

subperíodos separadamente. Entretanto, algumas considerações devem ser mencionadas em relação à técnica STFT, como o fato de que a resolução de tempo-frequência é constante para qualquer frequência\período analisado, não sendo indicado para a avaliação de comportamentos transientes nas amplitudes dos modos oscilatórios com períodos mais altos ou mais baixos do que aquele associado ao tamanho da janela temporal adotada.

Tal como define o princípio da incerteza de Heisenberg, as representações de uma determinada função nos domínios do tempo e frequência estão inversamente relacionadas em termos da “resolução” que pode ser preservada em cada dimensão (i.e., o nível de detalhes com que se consegue manter em cada eixo). Por exemplo, para a projeção em um periodograma, o período completo de um sinal qualquer é compreendido por meio da superposição de variações cíclicas estacionárias, de forma que se preserve maior resolução no domínio da frequência em troca da informação temporal.

Possuindo capacidade de preservar a informação temporal, no entanto, com melhor atribuição sobre as frequências representadas, a Transformada *Wavelet* Contínua (TWC) realiza projeção para o domínio da frequência a partir da nova base formada por funções *wavelets*, cuja característica principal diz respeito à representação finita ou localizada nos domínios do tempo e frequência. Tal propriedade contrasta com funções senoidais (aplicadas à FFT), as quais são localizadas apenas em relação à frequência (i.e., possuem energia “infinita” no domínio do tempo), tendo como resultado representações mais apropriadas (em espectrogramas) para análises de comportamentos transientes (LIU, Hui *et al.*, 2012; SIFUZZAMAN, M.; ISLAM, M. RAFIQ; ALI, 2009; VELEDA; MONTAGNE; ARAUJO, 2012). A principal motivação para a aplicação da TWC é que, diferentemente das técnicas baseadas em STFT, a “janela temporal” para avaliação do espectro em análises de tempo-frequência (espectrogramas) não são uniformes, ou seja, a avaliação espectral é realizada considerando múltiplas escalas de resolução, sendo janelas mais amplas aplicadas à descrição de baixas frequências e janelas mais curtas designadas a maiores frequências (JORDAN; MIKSAD; POWERS, 1997; TORRENCE; COMPO, 1998). Outra propriedade resulta do processo convolucional aplicado à transformada *wavelet*, o qual permite a representação do espectrograma em termos de frequências “instantâneas”, ou, sob perspectiva estritamente numérica,

com o número de instantes representados equivalente ao número de amostras do sinal original.

Em função da transformada, a representação da estrutura de fase e frequência dos sinais no domínio da frequência se dá em meio a pré-definições sobre a resolução de tempo e frequência com que se descreve o espectro do sinal. Por exemplo, para a FFT, as frequências são descritas com máxima resolução possível, implicando em mínima resolução temporal (do qual resulta representação do espectro em periodograma); para a STFT, a descrição de todas as frequências se dá a partir de uma mesma resolução temporal; para a transformada *wavelet*, variam-se as escalas de tempo e frequência para cada frequência descrita em cada intervalo do sinal.

Como alternativa a relações pré-estabelecidas de troca entre resoluções de tempo e frequência, abordagens “não paramétricas” baseadas na decomposição espectral do sinal foram propostas, como a decomposição em modos empíricos (*Empirical Mode Decomposition – EMD*; LI, X.; ZHANG; WANG, 2022). O EMD separa o sinal em harmônicos componentes fundamentais, denominados de “funções de modo intrínseco” (*Intrinsic Mode Function, IMF*), os quais isolam as informações mais relevantes do sinal em faixas de frequência representativas.

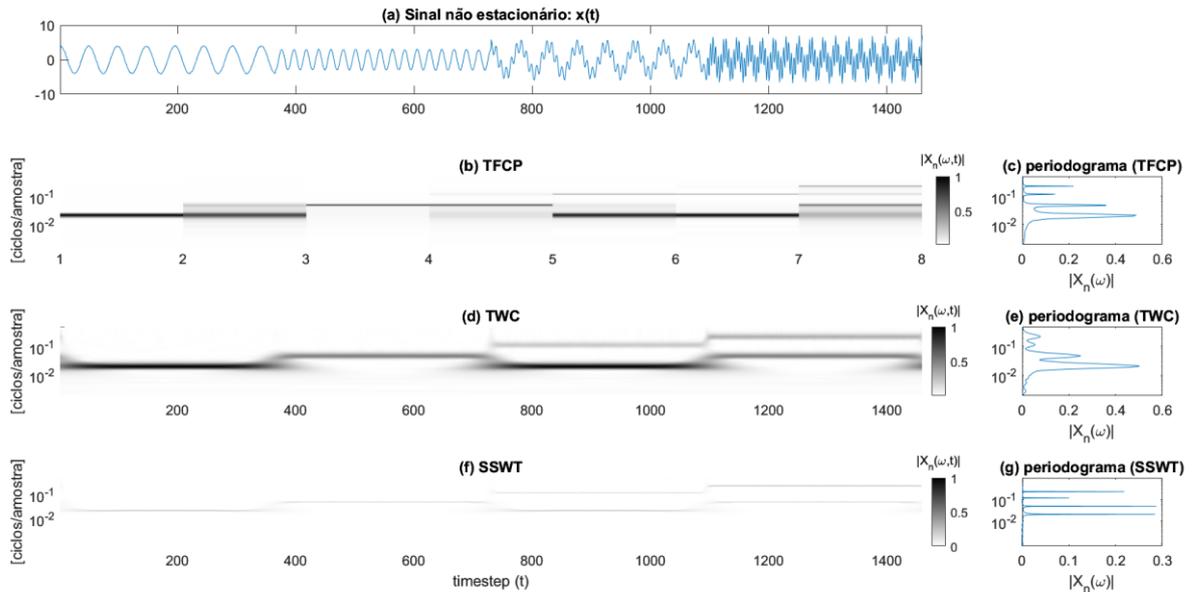
Unindo características da TWC e EMD, a *Synchrosqueezed Wavelet Transform* (SSWT) consiste na decomposição do sinal em “tipos de modos intrínsecos” (*Intrinsic Mode Types, IMT*), os quais, ao serem combinados, descrevem o sinal original de forma simplificada (Eq. 1)(DAUBECHIES; LU, J.; WU, H. T., 2011; MEIGNEN; OBERLIN; PHAM, 2019). A função φ_k é definida de forma que os IMT “isolem” os principais modos de frequência em cada instante de tempo t . A Figura 6 exemplifica um sinal sintético formado a partir da sucessão de quatro harmônicos sendo os dois primeiros harmônicos simples e os dois seguintes compostos. Constata-se que, em cada quarto do período em que se define $x(t)$, a técnica SSWT representa a energia dos sinais de forma precisa, reduzindo o espalhamento causado pelas limitações naturais da representação sobre o domínio da frequência. Destaca-se, no entanto, que tal abordagem não viola o princípio da incerteza, uma vez que se trata de uma simplificação por meio de redistribuição da energia do sinal na estrutura de frequências “instantâneas” do sinal.

$$x(t) = \sum_k IMT_k(t) \quad (1)$$

$$IMT_k(t) = A_k(t) \cos(\varphi_k(t)) \quad (2)$$

$x(t)$ – sinal original representado no domínio do tempo (t); IMT_k – k -ésimo IMT; A_k – amplitude do k -ésimo IMT; φ_k – função de separação do k -ésimo IMT.

Figura 6 – Comparação entre espectrogramas e periodogramas de sinal sintético não estacionário.



TFCP – Transformada de Fourier de Curto Prazo; TWC – Transformada Wavele Contínua; SSWT – *synchrosqueezed wavelet transform*. Fonte: autor.

A partir dos conceitos expostos ao longo da Seção 2, faz-se, a seguir, uma revisão bibliográfica detalhada sobre estudos dedicados ao *downscaling* estatístico aplicado a saídas de modelos de larga escala. Em tal levantamento bibliográfico, têm-se como ênfase as técnicas tipicamente adotadas para fins de pré-processamento e regressão de dados de previsão numérica do tempo (e.g., técnicas de seleção e extração de atributos, modelos lineares e não lineares). Por fim, ao final da próxima seção, um panorama geral é apresentado sobre o estado-da-arte em relação às metodologias para a combinação de modelos de *downscaling* estatístico para fins de previsão operacional.

3 REVISÃO BILIOGRÁFICA

Como discutido de forma preliminar na Seção 2.2, as técnicas empíricas de aumento de resolução espacial, ou *downscaling* estatístico, são amplamente empregadas em modelos de previsão numérica de tempo (*Numerical Weather Prediction*, NWP), e podem ser subdivididas em técnicas de regressão, classificação e geradores estocásticos (TANG, B. H.; BASSILL, 2018; WILBY, R.; WIGLEY, T., 1997).

As interpolações estão dentre as técnicas mais difundidas devido à baixa complexidade associada às suas parametrizações, e são frequentemente empregadas sob arquiteturas de parâmetros globais. Para fins de *downscaling* estatístico, a interpolação bilinear é habitualmente empregada como técnica de referência contra a qual são comparadas técnicas mais complexas, uma vez que é uma média ponderada em função apenas das posições geométricas das variáveis preditoras (ACCADIA *et al.*, 2003; BERNHARDT, M. *et al.*, 2010; PRYOR; SCHOOF; BARTHELMIE, 2005).

As técnicas baseadas em regressão apresentam um maior nível de sofisticação quando comparadas a interpolações, devido ao fato de possuírem procedimentos de otimização de parâmetros, os quais minimizam um gradiente de erro de ajuste entre os conjuntos preditando e preditor em questão. Dentre as técnicas mais simples, a regressão multilinear é frequentemente empregada em estudos de comparação de modelos, tendo em vista que apesar de serem de implementação relativamente simples, costumam apresentar resultados satisfatórios (CURRY; KAMP, VAN DER; MONAHAN, A. H., 2012; LI, Yi; DING, A. A., 2019; PAHLAVAN *et al.*, 2018). Outras abordagens relacionadas incluem: regressão linear não paramétrica (ROZAS-LARRAONDO; INZA; LOZANO, 2014); regressão sujeita a restrições (BÜRGER, 1996); regressão polinomial (GEORGE; JANAKI; PARAMESWARAN GOMATHY, 2016).

As redes neurais artificiais são bastante difundidas dentre as técnicas de regressão não linear, possuindo vantagens em relação a outras técnicas comuns de regressão (STORCH, VON, 2000). Um importante aspecto diz respeito à forma funcional (ou especificação do modelo), a qual é realizada em função dos próprios dados de treinamento e servem como aproximações de funções não lineares complexas. Em outras técnicas de regressão, define-se à priori a forma funcional do

modelo (e.g., linear, polinomial, etc.). A arquitetura em redes neurais normalmente é disposta em camadas, através das quais a informação é transmitida durante o processo de ajuste. Há dois tipos principais de rede: 1) propagação unidirecional (“*feed-forward*”); 2) retro propagativa (*backpropagation*; geralmente mais complexas).

Dentre as redes neurais do tipo *feed-forward*, destaca-se a conhecida como *perceptron* de múltiplas camadas (*Multi-Layer Perceptron*) devido à sua simplicidade (relativa a outros tipos de rede) e eficácia (HARPHAM; WILBY, Robert L., 2005; MAO; MONAHAN, A., 2018; OSMAN; ABDELLATIF, 2016). Outras abordagens mais complexas incluem redes voltadas para a descrição de transientes: rede neural temporal (COULIBALY; DIBIKE; ANCTIL, 2005; DIBIKE; COULIBALY, 2006; MENDES, D.; MARENGO, 2010) e rede neural recursiva (*recurrent neural network*, (MEDSKER, LARRY R.; JAIN, 2001). Em geral, as técnicas de regressão possuem um esforço computacional significativamente maior que técnicas lineares durante o treinamento, mas a aplicação dos parâmetros da rede, por sua vez, demanda baixo custo computacional.

Diversos modelos têm sido propostos através da aplicação de técnicas conhecidas como classificação de padrões sinópticos (WILBY, R.; WIGLEY, T., 1997). A técnica baseada em “análogos”, proposta inicialmente por Lorentz (1969), é bastante aplicada como um procedimento simples de classificação, servindo como modelo de referência para técnicas mais complexas (HORTON; BRÖNNIMANN, 2019; ZORITA, Eduardo; STORCH, Hans VON, 1999). O método na associação entre instantes diferentes com base em busca realizada em função da similaridade entre os padrões sinópticos, sendo comum o emprego da distância euclidiana (BARNETT; PREISENDORFER, 1978; TIMBAL; DUFOUR; MCAVANEY, 2003; WILKS, 2011). No caso mais simples, a função identidade (e.g., ZORITA e VON STORCH, 1999; WETTERHALL; HALLDIN; XU, 2005) ou regressão linear (e.g., HIDALGO, HUGO G.; DETTINGER, MICHAEL D.; CAYAN, 2008; PIERCE; CAYAN, 2016; TIPPETT; DELSOLE, 2013) são aplicadas como funções de transferência.

Quando providas de um banco de dados apropriado, as técnicas de análogos possuem a vantagem de serem capazes de superar estimativas obtidas com regressões “puras” em termos da variância reproduzida, uma vez que informações da variabilidade das observações são introduzidas na base de preditores considerada (PERRUCCI, 2018). Há limitações relacionadas com técnicas baseadas em análogos. A primeira é que essa abordagem demanda um esforço computacional relativamente

alto sobre extensos períodos de dados observacionais (GUTIÉRREZ *et al.*, 2004). A segunda reside na necessidade de períodos extensos para que os análogos sejam bem definidos, estando diretamente relacionada com a primeira (DOOL, VAN DEN, 1994).

As técnicas de agrupamento (*clustering*) são bastante aplicadas em modelos de arquitetura local, servindo como alternativas à técnica de análogos (GUTIÉRREZ *et al.*, 2004). O objetivo do método, em termos gerais, é a formação de classes que caracterizem subconjuntos de dados, os mais internamente homogêneos e externamente heterogêneos quanto forem possíveis, podendo ser considerado um processo natural de agrupamento “análogos”. Há dois tipos principais de *clustering*: Hierárquico e Não-Hierárquico (WILKS, 2011).

No *clustering* hierárquico, o agrupamento em classes é realizado de forma iterativa, determinando-se a partir de um determinado número de subgrupos (cada um contendo inicialmente um indivíduo) sucessivas fusões de subgrupos considerados mais semelhantes, assim, cada fusão reduz o número de subgrupos em uma unidade (*clustering* hierárquico aglomerativo; HART; GRAY; CLARK, 2015; MILLSTEIN *et al.*, 2019; WARD, 1963). Alternativamente, realiza-se o procedimento inverso, i.e., iniciando com um único grupo (contendo todo o conjunto de dados), realizam-se sucessivas partições dos elementos mais dispares em novos subgrupos. Essa última abordagem, conhecida como *clustering* dissociativo, é raramente empregada por ser menos apropriada para implementação e demandar um maior esforço computacional do que a anterior (WILKS, 2011).

Diferentemente dos métodos hierárquicos, as técnicas de *clustering* não-hierárquico permitem o intercâmbio de elementos entre classes durante o procedimento, o que, em tese, resulta numa configuração otimizada. Isso pode ser realizado a partir de um número de classes definidas à priori e, que em seguida, são caracterizadas pelos seus respectivos centroides (e.g., média aritmética), e reclassificadas em várias etapas até que uma configuração estável seja obtida (ENKE; SPEKAT, 1997; GUTIÉRREZ *et al.*, 2004; REICHENBERG; SIDDIQUI; WOGGIN, 2018). Outro exemplo de *clustering* hierárquico é a técnica conhecida como árvore de classificação e regressão, técnicas amplamente difundidas que permitem a avaliação de relações não lineares e com boa interpretabilidade, entretanto, possuem sensibilidade a *outliers* e podem apresentar instabilidade, ou alta variância nos

resultados de testes diferentes (ENKE; SPEKAT, 1997; ZORITA, E. *et al.*, 1995; ZORITA, Eduardo; STORCH, Hans VON, 1999).

Outras técnicas de classificação notáveis são as conhecidas como regras difusas (“*Fuzzy Rules*”). Essas se baseiam em sistemas de regras condicionais desenvolvidos para simular o processo de tomada de decisão humana para classificação de padrões sinópticos. As regras são elaboradas através de operadores lógicos do tipo “SE”, “E”, “OU” e “ENTÃO”, e associadas a atributos semi objetivos¹⁰. Esses atributos consistem em distribuições que normalmente são ajustadas de forma subjetiva, e aplicadas para quantificar a proximidade relativa, ou o “grau de pertinência” entre os elementos e as características definidos no sistema. A subjetividade introduzida no processo de parametrização de sistemas de inferência confere alta flexibilidade ao processo de classificação, incorporando conhecimentos informais (e.g., experiência em análises similares) aos critérios de decisão que se deseja reproduzir (BARDOSSY; BOGARDI; MATYASOVSKY, 2005).

Modelos probabilísticos baseados em *downscaling* são conhecidos como “geradores estocástico de tempo”, os quais podem ser voltados para estudos de previsão (BERNARDIN *et al.*, 2009; WILKS, 2011) e estudos de impactos causados por mudanças climáticas (SEMENOV; BARROW, 1997; WILBY, R. L. *et al.*, 1998). Esses modelos produzem séries temporais de várias variáveis (vento horizontal, precipitação, temperatura máxima e mínima diária, radiação solar e umidade) por meio de processo de inferência estocástica (HUTCHINSON, 1995; RICHARDSON, 1981).

Usualmente atribui-se aos modelos de *downscaling* estatísticos um conjunto de treinamento multidimensional, sendo comum nesses casos a introdução de informações linearmente dependentes (redundantes), as quais, além de não necessariamente implicarem em melhores resultados, demandam em maiores esforços computacionais. Neste sentido, costuma-se reduzir a dimensionalidade do sistema de preditores por meio de análise de componentes principais (“*Principal Component Analysis*”, PCA; e.g., BORDONI; STEVENS, 2006; JOLLIFFE, 1986), análise de correlações canônicas (e.g., BUSUIOC; TOMOZEIU; CACCIAMANI, 2008), ou ambas (HUTH, 1999). Estas técnicas formam novas bases para o sistema em

¹⁰ Por exemplo: SE “X” é alto, OU “Y” é baixo, ENTÃO “Z” é médio. Para essa regra, X e Y são entradas, Z é a saída, e alto, médio e baixo são atributos das variáveis, definidos subjetivamente.

questão sob as quais se costumam aplicar técnicas de regressão e classificação (por exemplo, MLR, NN, análogos, *clustering* etc.; WILKS, 2011a). Outras abordagens da redução no espaço de preditores se dá por meio de técnicas de filtro (as quais avaliam a relevância dos preditores em relação ao preditando): 1) seleção de atributos associada com teoria da informação mútua (AMJADY; KEYNIA; ZAREIPOUR, 2011); 2) decomposição espectral por meio de transformada *wavelet* (CATALÃO; POUSINHO; MENDES, V. M. F., 2011; WANG, Ning *et al.*, 2015); 3) decomposição em modos empíricos (EMD, *Empirical Mode Decomposition*; LI, X.; ZHANG, X.; WANG, S., 2022).

Em virtude do grande número de proposições aplicáveis à problemática de *downscaling* estatístico, estratégias de combinação têm sido adotadas para melhor aproveitar as vantagens de abordagens distintas (JUNG; BROADWATER, 2014; OSMAN; ABDELLATIF, 2016). Há dois objetivos principais para a combinação: aprimoramento e adaptação, as quais buscam, respectivamente, estimativas mais acuradas e robustas em relação às saídas individuais (COSTA *et al.*, 2008; HIBON; EVGENIOU, 2005; KANG, H. *et al.*, 2009; YANG, Y., 2004).

Há três principais grupos de técnicas de combinação de modelos: *bagging ensemble*; *boosting ensemble* e *stacked ensemble*. Os dois primeiros grupos, mais consolidados na literatura, dizem respeito a estratégias de combinação de base homogênea de modelos, tal qual árvores de decisão ou estruturas internas de redes neurais mais simples (e.g., *random forest* ou *multilayer perceptron neural network*). Ao longo dos últimos anos, no entanto, é possível observar um crescente número de trabalhos publicados sobre estratégias de combinação aplicadas a bases heterogêneas de modelos (*stacked ensemble*), as quais permitem a associação entre técnicas desde as mais simples, até mesmo aquelas que representam o estado-da-arte (FU *et al.*, 2023; NAITAM; MEGHANA; SRIVASTAV, 2023; ZHANG, Yuzhen; LIU, J.; SHEN, W., 2022).

O desenvolvimento de tais técnicas em áreas voltadas para a previsão de variáveis ambientais ocorre de forma concomitante ao aumento na demanda por previsões de variáveis como potência eólica e solar, as quais, além de possuírem relação de dependência com variáveis atmosféricas, estabelecem um grau de complexidade mais elevado em razão de perturbações de natureza elétrica, mecânica e de gestão (MARKOVICS; MAYER, 2022; ZHANG, Gang *et al.*, 2022). Nesse contexto, considerando a proximidade entre diferentes saídas de modelos de

downscaling estatístico (os quais descrevem o comportamento de uma variável para um mesmo local), técnicas lineares costumam oferecer resultados satisfatórios para a magnitude do vento e irradiância solar. Por outro lado, tais modelos oferecem baixo custo computacional, o que é favorável para a avaliação de um número maior de casos.

Em relação aos estudos voltados para a previsão em “curto prazo” (termo que aqui se refere aos horizontes de previsão de até 48 horas à frente) baseados em NWP, suas aplicações são similares àquelas tipicamente adotadas para fins de diagnóstico (e.g., análise), porém, existem desafios próprios à incerteza sobre estados futuros da atmosfera. Nesse sentido, estudos têm sido feitos com vistas à aplicação de modelos complexos baseados em inteligência artificial (GROCH; VERMEULEN, 2019). Por outro lado, outros autores têm se dedicado à correção de erros sistemáticos resultantes do processo de treinamento de modelos estatísticos (MA *et al.*, 2021). Por fim, é possível avaliar também a contribuição de diferentes bases de dados, as quais contêm conjuntos de variáveis endógenas e exógenas (BENAVIDES CESAR *et al.*, 2022).

Particularmente no caso de estudos voltados para a previsão de potência solar, em comparação com a aplicação de saídas de modelos de larga escala, maior ênfase tem sido dada aos modelos regionais de previsão numérica do tempo, como o WRF, WRF-Solar, HIRLAM, dentre outros (e.g., BACHER; MADSEN; NIELSEN, 2009; JIMENEZ *et al.*, 2016; PIERRO *et al.*, 2017). Tais modelos oferecem maior resolução espacial, sendo fundamental para a caracterização de padrões atmosféricos em maior nível de detalhes, especialmente em se tratando de variáveis com propriedades ópticas como a irradiância solar. No entanto, os modelos regionais demandam esforço computacional significativamente maior do que aquele requerido por modelos baseados em *downscaling* estatístico (aplicados a modelos de larga escala), os quais permitem assimilação de características do comportamento local às previsões de larga escala.

Apesar do crescente número de publicações com novas proposições para técnicas de previsão em curto prazo baseadas em saídas de modelos dinâmicos, até o momento, não há um consenso formado em relação a decisões importantes para o treinamento, seleção, e avaliação de modelos estatísticos em aplicações distintas (ADAM *et al.*, 2019; SCHULTZ *et al.*, 2021). Tal fato se deve à necessidade de relacionar diferentes aspectos inerentes à dinâmica atmosférica e à capacidade

preditiva de modelos físico-numéricos (e.g., variações sazonais e ocorrência de fenômenos típicos de escala global, regional ou local). Além disso, há poucas abordagens propostas para modelos de combinação de resultados produzidos por técnicas estatísticas aplicadas a saídas de modelos de larga escala de previsão numérica do tempo como o GFS. A maior parte dos estudos, no entanto, adotam métricas simples como o RMSE ou erro absoluto médio para a avaliação dos resultados. É importante ressaltar, no entanto, que soluções otimizadas são obtidas em relação à métrica adotada para a qualificação das saídas durante o treinamento dos modelos (i.e., função de perda ou *loss function*). Dessa forma, a integração entre diferentes configurações pode ser explorada por meio de experimentos voltados para o desenvolvimento de métodos baseados em *stacked ensemble*. Isso é importante para facilitar a associação de técnicas distintas, bem como para a flexibilização dos critérios aplicados aos modelos de *downscaling* estatístico (e.g., utilizando correlação, variância reproduzida, métricas de erro e outros índices de habilidade).

Para tornar modelos de combinação mais robustos e acurados, costuma-se adotar procedimentos de seleção de variáveis preditoras. A compatibilidade entre as variáveis regressoras candidatas e as observações pode ser avaliada por meio de parâmetros que resumem toda a informação temporal, como erros médios quadráticos ou correlação linear (KANG *et al.*, 2009; YU; LIU, Huan, 2003). Entretanto, tais parâmetros basicamente contêm informações relacionadas a centroides, que pouco dizem sobre os comportamentos periódicos sutis, como os de alta frequência (variações rápidas). Nesse contexto, a relevância de componentes periódicas pode ser mais facilmente explorada através de suas representações no domínio da frequência, ou seja, através de funções de densidade de energia espectral (JUNG; TAM, 2013; LEEDS; STEIN, M. L.; MOYER., 2013; PÉREZ; NAVARRO, M. E. J.; CANTOS, 2016). Como mencionado na Seção 2.6, diferentes modelos têm sido propostos a partir da aplicação de técnicas capazes de lidar com sinais complexos, com tendências não lineares e não estacionárias, tal como no caso das transformadas *wavelet* e técnicas de decomposição espectral (previamente discutidas na Seção 2.6). Entretanto, até o presente momento, em relação ao *downscaling* estatístico, a aplicação de tais técnicas tem centrado maior atenção em modelos de diagnóstico (e.g., BENESTAD *et al.*, 2019; HEGERMILLER *et al.*, 2017) não havendo estudos sobre aplicações de tais técnicas em modelos de previsão baseados em *downscaling*.

Por fim, como estado-da-arte, estudos recentes têm apontado para a necessidade de combinar modelos de diferentes tipos, incluindo as previsões numéricas do tempo (e.g., GFS e WRF), séries temporais (e.g., LSTM), bem como regressões aplicadas a conjuntos de entrada derivados de técnicas de engenharia de atributos (seleção e extração de atributos) (PENG, X. *et al.*, 2022; WANG, Y. *et al.*, 2021; ZHANG, Yu; LI, Yanting; ZHANG, Guangyao, 2020). Por outra parte, regressões com parâmetros dinâmicos (i.e., variáveis no tempo) também podem ser exploradas para maior adaptabilidade (DING, S.; LI, R.; TAO, 2021; JIANG *et al.*, 2020; SÁNCHEZ, 2008; VACCARO *et al.*, 2011). Como contribuição ao estado-da-arte, este estudo apresenta um novo método para a combinação de previsões a partir da otimização da complementaridade espectral entre os conjuntos de entrada e o alvo do modelo. Mais especificamente, avalia-se a hipótese de que a otimização da complementaridade espectral favorece o ajuste de modelos regressivos em relação à robustez e acurácia, tendo por pressuposto permite a identificação de informações com maior relevância e menor redundância.

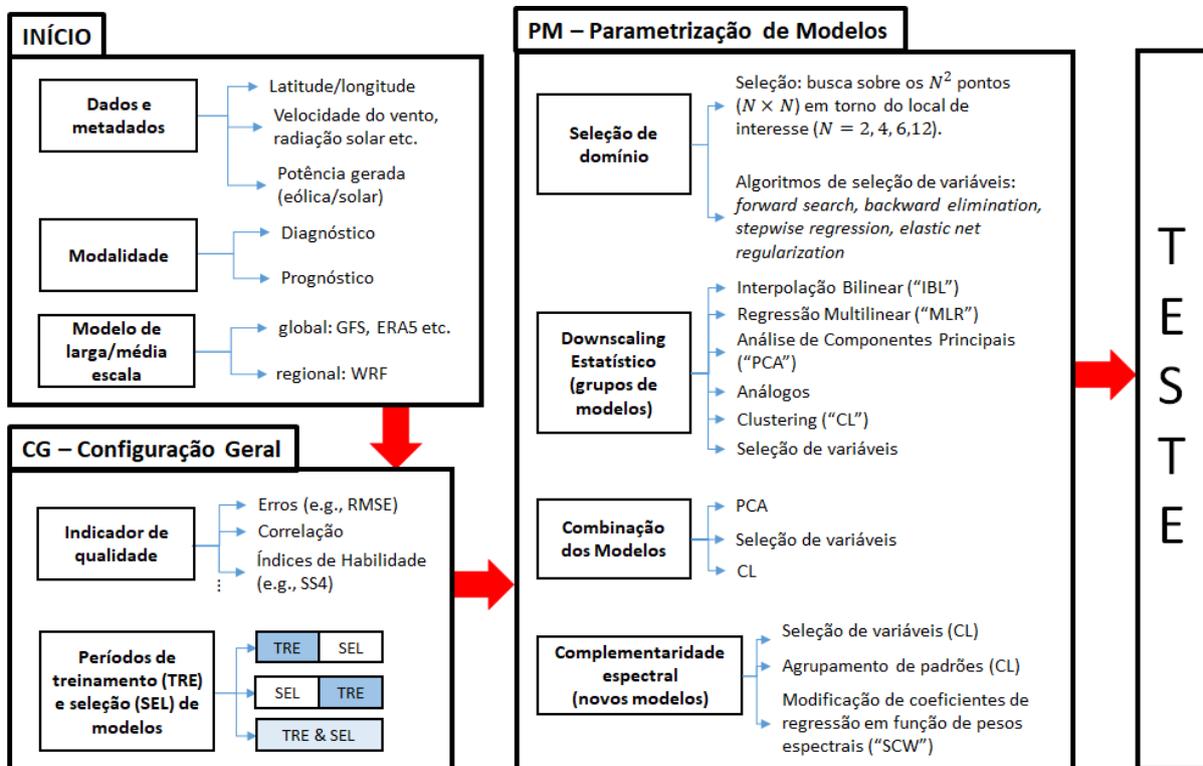
A próxima seção apresenta, de forma detalhada, a metodologia proposta no presente trabalho de tese, a qual consiste em plataforma (“*framework*”) de modelagem estatística que integra diferentes níveis de modelos, que incluem: seleção de domínio, *downscaling* estatístico e combinação de modelos. Partindo de configurações gerais da metodologia, a integração proposta se baseia na aplicação de métodos de otimização supervisionada (*wrappers*) voltados à configuração automática dos modelos em relação à seleção das variáveis, bem como de configurações diversas das técnicas adotadas em cada modelo. Além disso, os diferentes níveis da metodologia proposta permitem a intercomparação de modelos de *downscaling* estatístico e de combinação, bem como avaliação dos resultados em relação a diferentes aspectos da qualidade das previsões/estimativas. No intuito de investigar a aplicabilidade das projeções para o domínio da frequência em previsões operacionais, propõem-se novos métodos de pré-processamento de sinais para modelos regressivos com base na avaliação da complementaridade espectral entre sinais que descrevem o comportamento atmosférico em larga escala e observações relacionadas com tal comportamento aferidas no local de interesse.

4 METODOLOGIA E MODELOS

A metodologia apresentada nesta seção é estabelecida com vistas à combinação de modelos baseados em downscaling estatístico, que contemplam: i) diferentes configurações de treinamento; ii) modelos de *downscaling* estatístico baseados em seleção de domínio, regressão e classificação de padrões sinópticos (para que se possa avaliar aprimoramentos em diferentes níveis de complexidade); iii) combinação de modelos (com vistas à obtenção de resultados mais acurados e/ou robustos em comparação com saídas individuais); iv) modelos baseados em complementaridade espectral.

A Figura 7 apresenta o diagrama de blocos da metodologia adotada. Quatro blocos externos são definidos: Início (4.1), Configuração Geral (4.2), Parametrização de Modelos (Seção 4.5), Teste (Seção 4.6). Os detalhes sobre os algoritmos empregados para a implementação dos modelos aqui avaliados estão descritos na Seção 4.3.

Figura 7 – Metodologia de Downscaling Estatístico e combinação de modelos proposta.



Representação de plataforma de modelagem estatística desenvolvida para o *downscaling* estatístico e combinação de modelos. Fonte: autor.

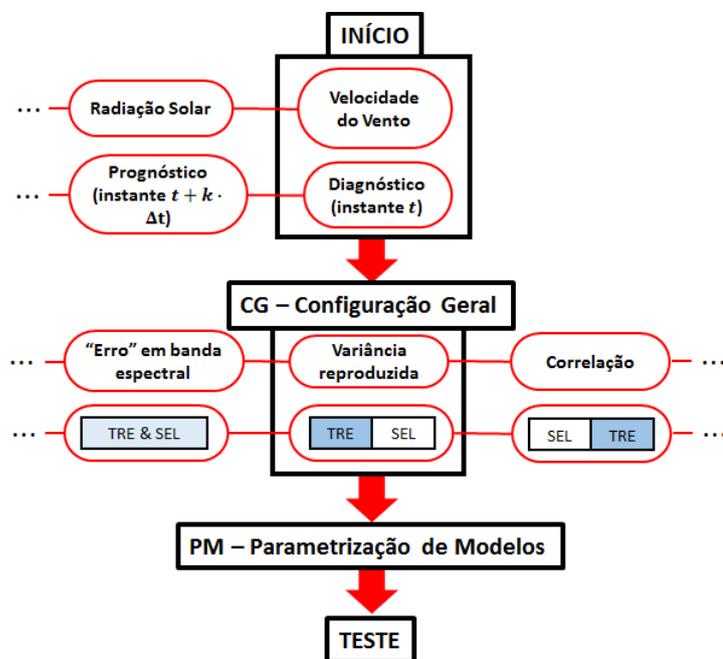
A partir de uma única inicialização, é possível realizar análises de sensibilidade sobre o sistema de modelagem em relação a diferentes configurações gerais, como ilustrado em exemplo na Figura 8 na qual se mostra um caso particular (hipotético) da metodologia (velocidade do vento estimada a partir de dados de reanálise para fins de descrição da variância observada, com divisão específica de treinamento e validação de modelos; mais detalhes na Seção 4.2). Através da diversificação de critérios de otimização no bloco de configurações gerais (aplicados no âmbito da inferência de parâmetros) dos modelos, serão estudadas as implicações de escolhas importantes sobre a avaliação de performance dos modelos propostos.

A Figura 8 exemplifica um caso particular da aplicação da metodologia proposta, em que são selecionadas configurações em cada bloco principal do diagrama apresentado na Figura 7. No exemplo, em uma inicialização para a modelagem da magnitude da velocidade do vento, a qual adota dados de previsão (ou de análise/reanálise) como entrada (provenientes de modelos de larga e/ou média escala), e como índice de qualidade das estimativas a variância reproduzida (ou razão entre desvios padrão das estimativas e observações). Mais detalhes sobre os blocos externos são apresentados nas Seções 4.1-4.6, a seguir.

Para o presente estudo, são consideradas apenas funções de transferência simples, como interpolação bilinear e regressão multilinear. Tal escolha se deve ao baixo esforço computacional e capacidade preditiva oferecida por essas técnicas para o problema de previsão por meio de simulações numéricas do tempo, ao passo que, no presente estudo, procura-se avaliar um número expressivo de casos com ênfase na comparação entre diferentes métodos seleção de variáveis (para modelos de ajuste global) e de agrupamento de padrões cujas amostras se aplicam à inferência de parâmetros (no que se refere a modelos de ajuste local). Informações sobre os modelos aplicados a este estudo, bem como informações gerais sobre os métodos de otimização adotados neste estudo se encontram nas Seções 4.3, 4.4 e 4.5.

Na etapa de inicialização, define-se o escopo da aplicação, que delimita a variável que se pretende descrever, se voltado para a previsão ou análise de recurso e, por fim, aplicado às simulações de larga e/ou média escala (Seção 4.1).

Figura 8 – Exemplo aplicação particular de sistema de modelagem estatística.



Representação de segmento (aplicação específica da ferramenta de *downscaling*) representado de forma genérica na Figura 7. Fonte: autor.

4.1 Inicialização

A variável de interesse, ou variável endógena, é a variável que será adotada como preditando (alvo) para o ajuste dos modelos propostos. Nesse estudo, será dada ênfase à magnitude da velocidade do vento e irradiância solar. Outras variáveis importantes também podem ser avaliadas, como a direção do vento ou potência gerada em central, sempre quando seja possível relacionar os dados observacionais a variáveis simuladas por GCMs. Entretanto, devido ao elevado número de graus de liberdade existentes na metodologia proposta, análises de sensibilidade devem ser restringidas em diversas etapas, para melhor aproveitamento de esforço computacional empreendido. É importante ressaltar que a magnitude do vento e a irradiância são variáveis que estão fortemente relacionadas com o potencial de geração em centrais eólicas e solares.

A modalidade do algoritmo diz respeito ao tipo de dados empregados para a composição de preditores. Em função da aplicação, a metodologia permite a utilização de dados de reanálise e previsão operacional, os quais são providos,

respectivamente, por meio de modelos diagnóstico e prognóstico (e.g., ERA-5 e GFS) (Seção 2.1). Em síntese, o bloco de inicialização (ou “INICIO”) diz respeito à definição do propósito da modelagem. Nesse texto, considera-se como propósito da modelagem a definição da variável de interesse (alvo do sistema de modelagem) e tipo de dado sobre o qual se deseja um aumento de resolução espacial (análise histórica ou previsão em curto prazo).

4.2 Configuração Geral (CG)

No bloco de Configuração Geral, são realizadas decisões relativas ao índice de qualidade e períodos para treinamento e seleção de variáveis e demais hiperparâmetros (métodos de otimização). Os índices de qualidade são parâmetros que transportam informações relevantes sobre a relação entre variáveis estimadas e as observações (alvo da modelagem) (CHANDRASHEKAR; SAHIN, 2014; KOHAVI, 1995; TANG, J.; ALELYANI; LIU, Huan, 2014). Nesse estudo, como critério de otimização de modelos, aplicam-se os índices de correlação de Pearson e raiz do erro médio quadrático (RMSe). A avaliação dos resultados dos modelos também será realizada a partir do índice de habilidade “SS4”, métrica que pondera entre a variância reproduzida e a correlação, e assume valores entre zero e um, respectivamente, do menor ao maior potencial explicativo (Eq. 3; TAYLOR, 2001):

$$SS4 = \frac{(1 + \rho)^4}{4(\hat{\sigma} + 1/\hat{\sigma})^2} \quad (3)$$

Em que: ρ – Índice de correlação de Pearson entre estimativas e observações;
 $\hat{\sigma}$ – desvio padrão das previsões normalizado pelo desvio padrão das observações.

Nesse estudo, também serão avaliados índices de habilidade globais e locais que indicam a similaridade entre espectros em relação à complementaridade espectral, ou seja, voltados para a avaliação da relevância e redundância espectrais em conjuntos de preditores. As transformadas podem ser aplicadas para estimar o espectro global (periodograma), bem como a evolução temporal do espectro (espectrograma). Mais detalhes sobre técnicas para análise sobre o domínio da frequência estão disponíveis nas Seções 2.6. Detalhes sobre os modelos que adotam tais índices estão disponíveis na Seção 4.4 (PERRUCCI, 2018).

A relação de dependência entre preditores e preditando pode sofrer variações em relação a parâmetros fixos, particularmente nos casos em que não há disponibilidade de dados observacionais em período suficientemente longo. Note, entretanto, que o potencial preditivo é quantificado pelo índice de qualidade considerado, cujo valor também pode ser avaliado em análise de sensibilidade para estudos sobre possíveis contribuições de classificadores distintos em procedimentos de seleção de variáveis (Seção 4.4).

Com vistas a avaliar a acurácia e a capacidade de generalização dos modelos preditivos, adotou-se um procedimento de validação cruzada baseado na divisão das séries temporais em dois conjuntos distintos: treinamento e teste. O período de treinamento é composto por dois subconjuntos: calibração, dedicado à inferência de parâmetros; e validação, aplicado para a otimização de hiperparâmetros dos modelos. Na Tabela 1 estão descritas as ações realizadas em cada um desses períodos. Em função da aplicação de técnicas baseadas em projeções para o domínio da frequência, os períodos de calibração e validação são formados pela separação do conjunto de treinamento em duas partes de igual tamanho e “contínuas”, ou seja, sem amostragem aleatória (i.e., validação cruzada baseada em *hold out*).

Ao serem definidos, para cada simulação realizada, a estação anemométrica/solarimétrica (ou sistema de geração de potência), mês de operação e horizonte de previsão, adotam-se amostras anteriores ao período de teste para a calibração e otimização dos modelos. Além disso, em relação aos períodos aplicados à calibração e validação dos modelos, duas configurações foram consideradas no âmbito de busca exploratória: 1) o período dedicado à calibração dos modelos individuais inclui período anterior ao aplicado para a validação dos modelos, cada qual contendo aproximadamente o mesmo número de amostras; 2) considera-se o inverso da configuração (1), i.e., aplica-se o período posterior para a calibração e anterior para a validação dos modelos individuais. Ressalta-se que a possível ocorrência de modelos sobreajustados ao período de calibração (i.e., com erros baixos para o período de treinamento, porém com menor capacidade de generalização para outros períodos) pode induzir os modelos de combinação ao erro durante o treinamento. Para mitigar tal problema, particularmente em relação aos modelos de combinação, invertem-se os conjuntos de calibração e validação após o treinamento dos modelos individuais. Destaca-se, ainda, que tais definições são consideradas através de parâmetros ajustáveis (i.e., graus de liberdade da metodologia proposta), os quais

permitem análises de sensibilidade sobre diversos critérios que são tipicamente definidos à priori. Contudo, neste trabalho, apenas um conjunto reduzido de configurações é avaliado, uma vez que a busca ampliada está fora do escopo da tese ora apresentada. Como perspectiva ao desenvolvimento da metodologia proposta, considera-se um espaço de busca expandido para configurações diversas, como índices de qualidade, resolução espacial do modelo de larga/média escala, número de amostras em conjuntos de treinamento e seleção etc.

Por último, para a avaliação final dos resultados, o período de teste aplicado contempla intervalo posterior àquele referente ao conjunto de treinamento, uma vez que se dedica à simulação da operação em tempo real dos modelos de *downscaling* estatístico. No presente trabalho, a performance dos modelos de *downscaling* estatístico aplicado ao GFS se concentrou em meses de baixa e alta geração eólica/solar.

Tabela 1 – Divisão de conjuntos de dados para calibração, validação e testes dos modelos de *downscaling* estatístico e de combinação aplicados à previsão do comportamento do vento local.

Conjunto	Aplicação	Período
Treinamento (calibração)	Inferência dos parâmetros dos modelos de <i>downscaling</i> estatístico	Primeira ou segunda metade do conjunto de treinamento.
Treinamento (validação)	Inferência dos hiperparâmetros dos modelos de <i>downscaling</i> estatístico através de testes de generalização (“wrappers”). É importante ressaltar que, como abordagem inicial, as estatísticas empregadas com vistas à otimização ¹¹ dos hiperparâmetros dos modelos de <i>downscaling</i> estatístico e de combinação foram o coeficiente de correlação de Pearson e o RMSE, respectivamente.	Parte complementar ao subconjunto de calibração
Teste	Avaliação da acurácia final dos modelos de <i>downscaling</i> estatístico sobre um período independente, i.e., um período que não foi empregado para qualquer tipo de inferência ou tomada de decisão por parte dos modelos. Este conjunto está dedicado à simulação da operação em tempo real dos modelos de previsão.	Meses de baixa e alta estação de geração eólica/solar.

¹¹ A metodologia permite a criação de diferentes parametrizações a partir das opções para os critérios de otimização dos modelos. Tais parametrizações podem ser combinadas em uma terceira camada de modelos, entretanto, tal abordagem está fora do escopo deste trabalho.

4.3 Algoritmos de modelos

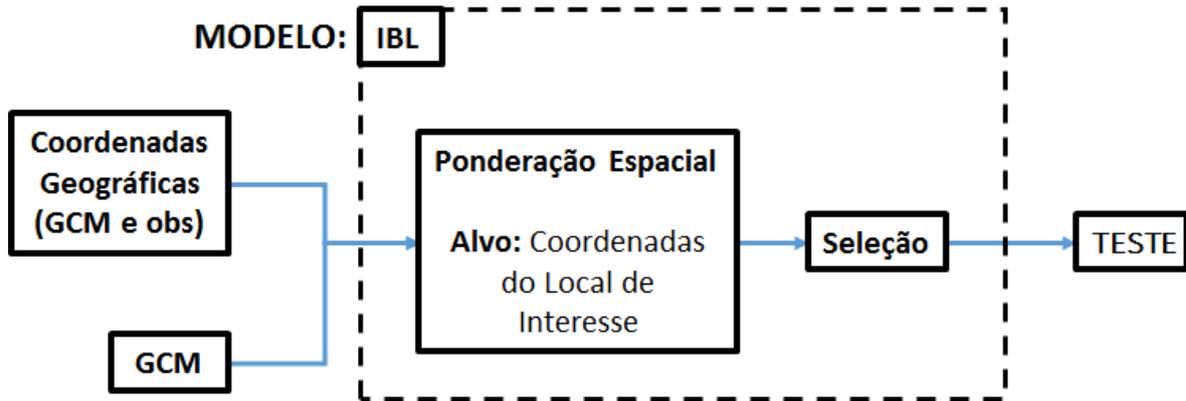
Nesta seção são apresentados o modelo de referência adotado e os algoritmos de otimização supervisionada aplicados aos modelos estatísticos desenvolvidos neste trabalho, os quais são voltados para a seleção de domínio, *downscaling* e de combinação de previsões de velocidade do vento, irradiância solar e geração de potência eólica e solar. Tais algoritmos permitem a integração da metodologia por meio de critérios objetivos, os quais são voltados para a avaliação da qualidade das previsões em relação a aspectos diversos (e.g., descrição do valor médio, posição e amplitude de variações “rápidas”, tendências sazonais etc.).

4.3.1 IBL – Interpolação Bilinear

A interpolação bilinear é aplicada para estimar a variabilidade espacial da região contemplada pela malha que contém os pontos (de saídas do GCM) mais próximos ao local de interesse (Figura 9). O fator de ponderação para um estado definido em um ponto da malha (situado nas extremidades) é inversamente proporcional à distância para o ponto de interesse (necessariamente situado no interior da malha).

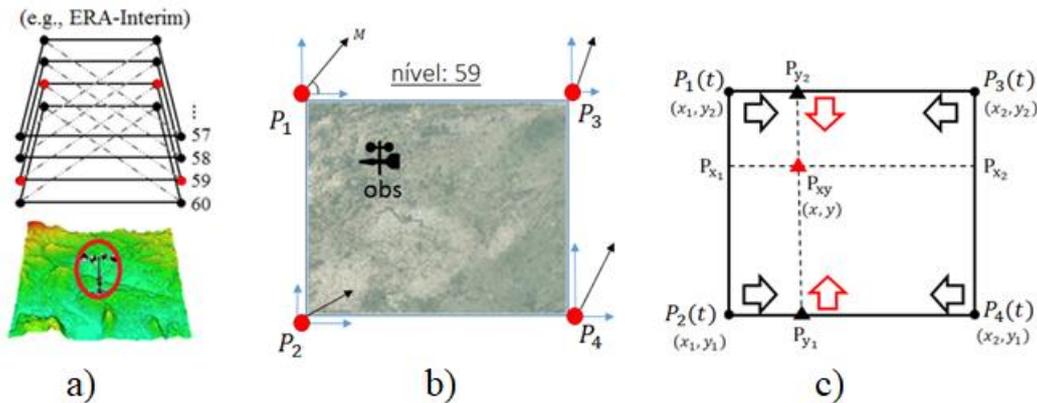
Nesse trabalho, o método considerado faz uso dos quatro pontos, em um mesmo nível de modelo do GCM (Figura 10a), mais próximos ao local de interesse considerando distâncias projetadas em plano horizontal (vista superior, Figura 10b). O procedimento de interpolação bilinear pode ser entendido por meio três interpolações lineares sucessivas. A Figura 10c ilustra o método no caso em que duas interpolações são realizadas a partir de pontos adjacentes em relação à direção longitudinal (para posição de interesse em relação à abcissa, x), formando estimativas, respectivamente, para os extremos inferior e superior P_{y_1} e P_{y_2} . Em seguida, realiza-se uma interpolação zonal entre os pontos P_{y_1} e P_{y_2} para gerar uma estimativa para o local de interesse.

Figura 9 – Diagrama do modelo IBL



Modelos de *downscaling* estatístico baseado em interpolação bilinear. O modelo reproduz saídas do GCM no local de interesse. Função de transferência depende apenas de coordenadas de pontos (da malha do GCM e do local de interesse). Fonte: autor.

Figura 10 – Ilustração de processo de interpolação bilinear.



a) Vista lateral: seleção de nível do modelo; b) Vista superior: associação com local de interesses por distâncias em plano horizontal; c) procedimento de interpolação. Fonte: autor.

Mais formalmente, em acordo com a Figura 10c, o estado de uma determinada variável pode ser representado por $f = f(x, y, t)$, em coordenadas x e y (longitude e latitude, respectivamente), em um determinado nível “ z ” (omitido). As duas primeiras interpolações longitudinais f_L para os pontos P_{y_1} e P_{y_2} podem ser descritas tal como nas Eqs. 4 e 5:

$$f(x, y_2) = f_L(P_{y_2}) = \left(\frac{x_2 - x}{x_2 - x_1} \right) f(P_1) + \left(\frac{x - x_1}{x_2 - x_1} \right) f(P_3) \quad (4)$$

$$f(x, y_1) = f_L(P_{y_1}) = \left(\frac{x_2 - x}{x_2 - x_1} \right) f(P_2) + \left(\frac{x - x_1}{x_2 - x_1} \right) f(P_4) \quad (5)$$

A estimativa final é dada pela interpolação zonal $f_Z(P_{xy})$, em função dos pontos previamente interpolados $f(P_{y_2})$ e $f(P_{y_1})$ (Eq. 6):

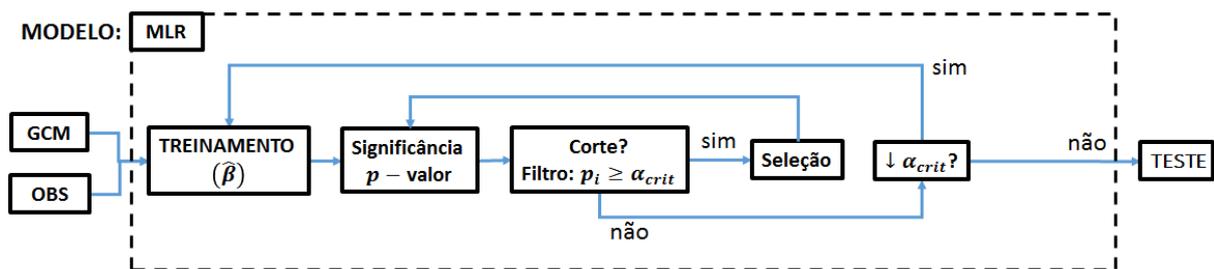
$$\hat{Y}(x, y) = f_z(P_{xy}) = \left(\frac{y_2 - y}{y_2 - y_1}\right) f(P_{y_1}) + \left(\frac{y - y_1}{y_2 - y_1}\right) f(P_{y_2}) \quad (6)$$

4.3.2 MLR – Regressão Multilinear

A regressão multilinear, no presente trabalho, é implementada com auxílio de um procedimento de otimização, como descrito no início da Seção 4.3.

O diagrama apresentado na Figura 11 ilustra o procedimento para ajuste e busca por especificação ótima. O bloco delimitado por linha pontilhada resume o processo que é repetido para o ajuste de todos os modelos lineares apresentados ao longo da Seção 4.3, bem como aplicado para seleção de variáveis preditoras (filtro) para modelos de combinação.

Figura 11 – Diagrama do modelo MLR



Modelo de *downscaling* estatístico baseado em regressão multilinear. Um processo de otimização de especificação de modelo linear também é demonstrado. Fonte: autor.

Para o ajuste do modelo, coeficientes são estimados para todos os preditores incluídos no domínio, os quais são obtidos de pontos de malha simulados pelo GCM (diagnóstico ou prognóstico, ver Seção 2.1). No bloco de treinamento, aplica-se o estimador do método de mínimos quadrados $\hat{\beta}$ (Eq. 7), dado por (WOOLDRIDGE, 2015):

$$\hat{\beta} = (X'X)^{-1}X'Y \quad (7)$$

em que: $X_{n \times p}$ – matriz contendo m preditores e coluna para parâmetro de interseção (β_0), totalizando $p = m + 1$ colunas; $Y_{n \times 1}$ – observações (alvo). Mais especificamente, as observações podem ser representadas pelo modelo linear considerando o resíduo $\epsilon_{n \times 1}$ (o qual se deseja minimizar):

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1m} \\ 1 & x_{21} & \dots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{nm} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

OU:

$$Y = X\hat{\beta} + \varepsilon;$$

Assim, o comportamento do vento é estimado pela seguinte expressão (Eq. 8):

$$\hat{Y} = X\hat{\beta} \quad (8)$$

No bloco seguinte, para fins de seleção de variáveis de entrada, um teste de significância é realizado para os coeficientes $\hat{\beta}$ por meio do teste t de Student, para a hipótese $H_0: \beta_i = 0$, contra a alternativa $H_1: \beta_i \neq 0$. Para o cálculo dos p-valores, em abordagem preliminar, considera-se o cálculo do erro padrão obtido na estimação dos coeficientes a partir da matriz de covariâncias entre preditores (Eq. 9):

$$SE(\hat{\beta}) = \sqrt{\text{diag} \left[\left(\frac{1}{t-p} \right) (\varepsilon'\varepsilon)(X'X)^{-1} \right]} \quad (9)$$

Na Eq. 9 o termo “diag” indica a extração da diagonal principal da matriz $p \times p$ resultante (variância do resíduo ε normalizada pelo número de graus de liberdade $t - p$ e pela matriz de covariância da matriz de preditores X). O valor esperado para o resíduo é considerado nulo independentemente dos preditores ($\mathbb{E}(\varepsilon | X) = 0$). Note que, na Eq. 9, considera-se como estimativa não enviesada para a variância dos coeficientes $\hat{\beta}$ o cálculo da soma do erro quadrático ($\varepsilon'\varepsilon$) ajustado pelo número de graus de liberdade. Em abordagem inicial, pressupõe-se inicialmente que o vetor de resíduo ε possui variância constante ou independente em relação aos valores de X (homoscedasticidade¹²). Na presença de heteroscedasticidade¹³, entretanto, as simplificações realizadas (omitidas nesse texto) não deveriam ser realizadas. Caso seja verificada a necessidade, outros estimadores de incerteza sobre a estimação dos parâmetros podem ser considerados em proposições futuras.

O estatístico de teste é calculado por: $t_i = \frac{\hat{\beta}_i}{SE_i}$, $i = 1, \dots, p$. Dessa forma, O p-valor é obtido a partir de distribuição t com $n - p$ graus de liberdade. Os p-valores são comparados com níveis de significância críticos (α_{crit}), os quais são estabelecidos

¹² $Var(\varepsilon|X) = \sigma^2$. Dispersão do resíduo constante em relação aos valores assumidos pelos preditores.

¹³ $Var(\varepsilon_i|x_{ij}) = \sigma_{ij}^2$. Dispersão do resíduo varia em função dos valores assumidos pelos preditores.

à priori. As variáveis cujos coeficientes satisfazem a relação $p_i \geq \alpha_{crit}$ são eliminadas, e a nova especificação do modelo é avaliada por meio de verificação no período reservado para seleção (Seção 4.2). Caso o resultado em relação ao índice de qualidade considerado for equivalente ou superior, mantém-se o corte, caso contrário, as variáveis são reincluídas na especificação do modelo. Caso uma especificação seja mantida após o corte, realiza-se novamente o teste de significância, para o mesmo valor de α_{crit} . O laço é interrompido se uma das três condições for satisfeita: 1ª) $p_i < \alpha_{crit}, \forall i$; 2ª) uma configuração anterior é reestabelecida; 3ª) Um número máximo de iterações for atingido (estabelecido pelo usuário).

Ao término do laço no bloco de corte de variáveis, o próximo bloco avalia se há um valor mais restritivo (menor) para α_{crit} dentre os estabelecidos pelo usuário. Em caso positivo, reinicia-se o processo a partir do conjunto de preditores completo, para novos testes de significância e corte, considerando valores de α_{crit} em ordem decrescente. O bloco de “seleção” na Figura 11 armazena a especificação que gera os melhores resultados em relação ao índice de qualidade adotado, preservando resultados de iterações passadas de α_{crit} (não mostrado na figura). Quando todos os valores de α_{crit} forem avaliados, a melhor especificação é estabelecida como forma final do modelo.

4.3.3 PCA+MLR – Regressão Multilinear de Componentes Principais

A técnica baseada em análise de componentes principais (PCA – *Principal Component Analysis*) se baseia em rotação de um sistema formado pelos preditores $X_{n \times m}$ para um novo sistema de base ortogonal, cujas direções dos eixos indicam direções com maior variância. A matriz de rotação para sistema de componentes principais é formada pelos autovetores da matriz de covariância (ou correlação, caso variáveis com ordens de grandeza diferentes sejam associadas), os quais estão dispostos nas colunas da matriz $P_{m \times m}$. Os autovalores indicam a contribuição de uma componente principal (combinação linear dos eixos do sistema original) para a variância do sistema original. Sendo assim, as m componentes principais são ranqueadas pelos respectivos autovalores.

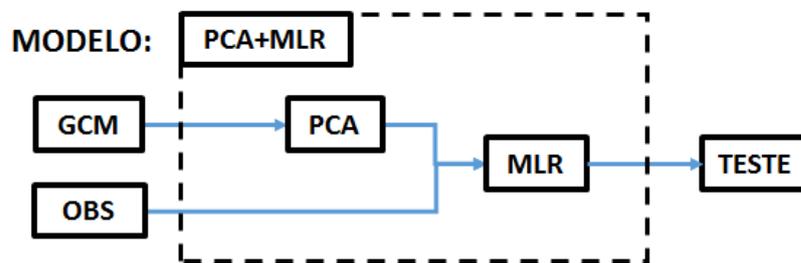
Realiza-se um corte de componentes com a menor contribuição (i.e., autovalores), preservando o menor número n que satisfaz a seguinte relação (JOLLIFFE, 1986):

$$E_R = \frac{1}{m^2} \cdot \sum_{j=1}^m \sum_{i=1}^m \frac{(R_{i,j} - X_{i,j})^2}{X_{i,j}^2} \leq 5\%$$

Na qual, $R_{t \times m} = X_{t \times m} P_{m \times n} P'_{n \times m}$ é a reconstrução do espaço por meio de projeção em n componentes principais; E_R – erro de reconstrução.

O método PCA é acoplado a outros modelos regressivos, por meio de pré-processamento dos conjuntos preditores, tal como se exemplifica na Figura 12 (como o modelos “MLR”, descrito na Seção 4.3.2). Os modelos que adotam tal procedimento de redução de dimensionalidade recebem a letra “p” em sua nomenclatura (e.g., MLRp).

Figura 12 – Exemplo de acoplamento de algoritmos: PCA+MLR



Exemplo de modelo de *downscaling* estatístico baseado em análise de componentes principais e regressão multilinear (modelo “MLRp”). Fonte: autor.

4.3.4 AN – Regressão Multilinear de Análogos

As técnicas de análogos são adotadas para compor conjuntos de preditores baseados na similaridade entre padrões sinóticos no instante sobre o qual se deseja realizar previsões e outros instantes sobre os quais se têm informações (armazenadas em bancos de dados). Informações observadas e de saídas de modelos de larga/média escala, bem como outras informações (e.g., dados de satélites) podem ser relacionadas por meio de instantes análogos. Neste estudo, porém, são adotados dados observacionais e saídas de modelo de larga escala em instantes análogos para o ajuste das funções de transferência dos modelos “AN”. Em síntese, o método é

aplicado como um procedimento de amostragem (não aleatória) com repetição para a reorganização temporal dos conjuntos de treinamento para modelos lineares em função da minimização de uma função de distância (WETTERHALL; HALLDIN; XU, 2005).

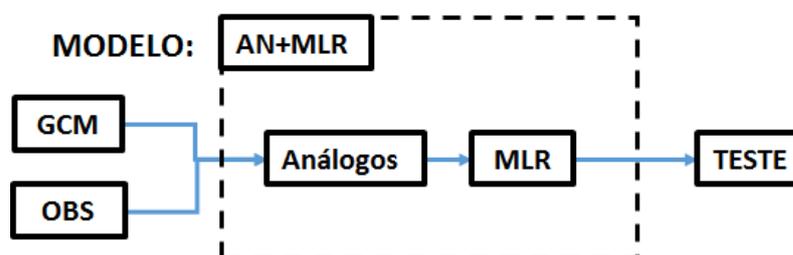
A Figura 13 apresenta o procedimento geral para o ajuste de modelos baseados em análogos. A distância euclidiana é empregada para ordenar instantes associados a padrões sinópticos em um conjunto de treinamento em relação à similaridade com um padrão sinóptico alvo. Na Figura 14a, o instante “s” é o instante alvo, e o instante “ t_a ” é dado como análogo por possuir menor distância em um sistema formado pelos preditores P_1 , P_2 , P_3 e P_4 . A distância euclidiana é calculada pela seguinte expressão:

$$d(t, s) = \|\mathbf{P}(t) - \mathbf{P}(s)\|$$

em que $\mathbf{P}(t) = [P_1, P_2, \dots, P_{m=4}]_t$ – padrão sinóptico definido pelo estado de preditores (P_1, P_2, \dots, P_m) em um instante t . O instante análogo será $t = t_a$, tal que, $d(t_a, s) = \min[d(t, s)]$. Desconsiderando o instante s (durante o treinamento ou seleção/validação), o modelo aplica parâmetros por amostras análogas do banco de dados que são associadas com o alvo por parâmetros lineares (Figura 14b). Entretanto, apenas são considerados para regressão os preditores remanescentes de procedimento de seleção resultante do algoritmo “MLR” (detalhado na Seção 4.3.2). As técnicas de análogos se baseiam na suposição de que se $P(s) \cong P(t_a)$, então, $\hat{Y}(s) \cong Y(t_a)$. As funções lineares são treinadas considerando funções de transferência para conjuntos de preditores formados por observações em instantes análogos, $Y(t_a)$; padrões sinópticos no instante análogo, $\mathbf{P}(t_a)$; e padrões sinópticos no instante alvo, $\mathbf{P}(s)$:

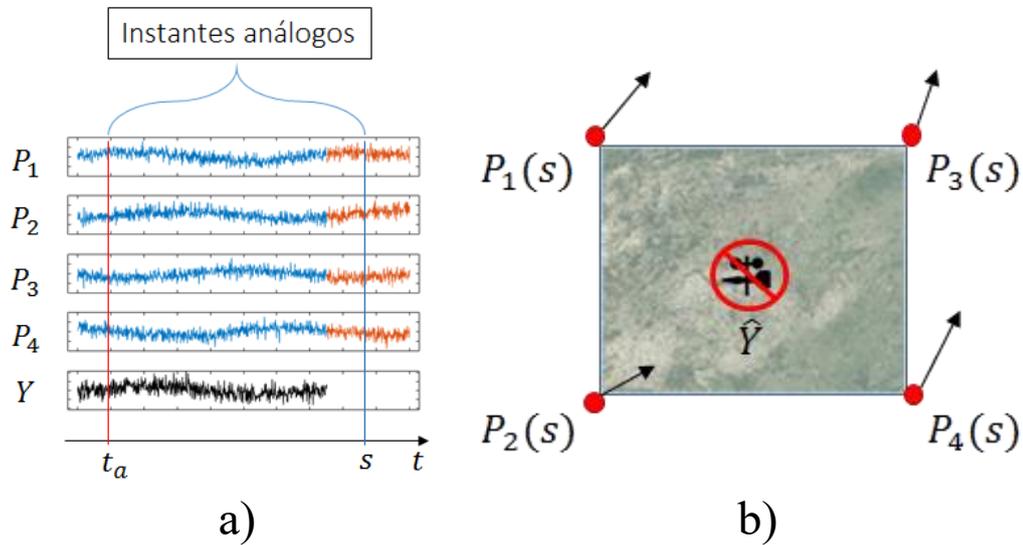
$$\hat{Y}(t^*) = f(Y(t_a), \mathbf{P}(t_a), \mathbf{P}(s))$$

Figura 13 – AN+MLR



Modelo de *downscaling* estatístico baseado em composição de análogos e regressão. Fonte: autor.

Figura 14 – Instantes análogos e alvo.



Fonte: autor.

4.3.5 ENR – “*Elastic Net Regression*”

Os modelos baseados no algoritmo “EN” se baseiam em método de regularização “*Elastic Net*”, que consiste em modificação no processo de inferência de parâmetros de modelos lineares com vistas à prevenção de vieses (“sobreajustes”), o que é particularmente importante em conjuntos com alta redundância (comum em saídas modelos de previsão numéricas do tempo). Tal modificação ocorre diretamente na função de custo do método de mínimos quadrados tradicional (Eq. 7; Seção 4.3.2), introduzindo-se um termo de penalização da variância da matriz de coeficientes β , tal como indicado por P_α nas Eq. 10. O parâmetro α no termo de regularização pondera entre as normas do tipo L1 (soma de módulos dos termos em β) e L2 (raiz da soma dos quadrados dos termos em β ; Eq. 11). No caso particular em que $\alpha = 0$ (regressão linear tipo “*ridge*”), os coeficientes tendem a zero com o aumento do termo de regularização λ , sendo indicado no caso de conjuntos com poucas variáveis, a partir dos quais se deseja obter relação robusta. No caso oposto, em que $\alpha = 1$ (regressão linear tipo “*lasso*”), os coeficientes atingem o valor zero com o aumento do termo de regularização, sendo indicados para aplicação em conjuntos extensos de preditores como método de seleção de variáveis (Eq. 12). No presente trabalho, os modelos

“ENRs”, de *downscaling* estatístico, e “CBLEN”, de combinação, adotam método de seleção baseado em Lasso, em que são determinados conjuntos de variáveis e coeficientes do modelo a partir da minimização do MSE obtido a partir da função de custo L (Eq. 10).

$$L = \min_{\beta_0, \beta} \left(\frac{1}{2T} \sum_{t=1}^T (y_t - \beta_0 - x_t^T \beta)^2 - \lambda P_\alpha(\beta) \right) \quad (10)$$

$$P_\alpha = \sum_{j=1}^K \left(\frac{(1-\alpha)}{2} \beta_j^2 + \alpha |\beta_j| \right) \quad (11)$$

$$P_\alpha = \lambda \sum_{j=1}^K |\beta_j|, \quad \alpha = 1 \quad (12)$$

O método Lasso também é aplicado como método de seleção de variáveis regressoras para os modelos baseados em complementaridade espectral, servindo como modelo de referência contra o qual serão intercomparados os resultados dos novos modelos propostos (Seção 4.4). Tal escolha se deve à possibilidade de se comparem possíveis ganhos com os métodos propostos nesta tese.

4.3.6 SWR, SWM – “*Stepwise Regression*”

O algoritmo *stepwise regression* consiste na aplicação de testes sucessivos, que incluem a inclusão de potenciais preditores ao modelo, e, a cada passo, avalia-se a remoção de um dos termos presentes no modelo. Tipicamente, a avaliação de pertinência de preditores ao modelo é feita, por meio de testes de hipótese, e.g., teste F ou p-valor sobre os coeficientes do modelo. Como hipótese nula, contra a qual se avalia a significância dos coeficientes do modelo, considera-se que associado coeficiente de um preditor k provém de distribuição com média nula e variância estimada tal como se descreve na Seção 4.3.2. Se não houver distinção suficiente entre as distribuições ao nível de confiança estabelecido, retira-se a variável do modelo. Após avaliar possíveis adições/remoções, o método retorna ao passo anterior, adicionando a variável (ainda não contemplada) com maior significância estatística.

Os modelos de *downscaling* estatístico “SWRs” se baseiam no algoritmo supracitado. O modelo “SWMs”, proposto neste trabalho, consiste em um algoritmo modificado para a seleção de variáveis com base em índices de habilidade distintos. Em outras palavras, os passos definidos no parágrafo anterior são mantidos, no entanto, o teste de aceitação de variáveis se dá por meio da maximização/minimização índices de habilidade definidos na etapa de Configuração Geral, calculados sobre o período de validação dos modelos (CG; Seção 4.2). Ao final do ajuste, aplica-se algoritmo “MLR” para a obtenção da previsão.

4.3.7 MR – Máxima Relevância Mínima Redundância

O método MRMR (Máxima Relevância Mínima Redundância) avalia a complementaridade entre preditores por meio da ponderação entre termos de relevância e redundância inferidos a partir da Informação Mútua entre os sinais (Eq. 13; DARBELLAY; VAJDA, 1999; DING; PENG, H., 2005). Os termos de relevância (V_K) e redundância (W_K) podem ser calculados sobre conjunto de preditores K , a partir da informação mútua entre predidando e preditor e dos preditores entre si, respectivamente (Eqs. 14 e 15).

$$I(X, Y) = \sum_{i,j} P_{XZ}(X = x_i, Z = z_j) \log \left(\frac{P_{XZ}(X = x_i, Z = z_j)}{P_X(X = x_i)P_Z(Z = z_j)} \right) \quad (13)$$

$$V_K = \frac{1}{|K|} \sum_{x \in K} I(x, y) \quad (14)$$

$$W_K = \frac{1}{|K|^2} \sum_{x,z \in K} I(x, z) \quad (15)$$

Em que: I – informação mútua; P_{XZ} – função de distribuição de probabilidade (PDF) conjunta dos preditores x e z ; P_X, P_Y – funções de distribuição de probabilidade independentes (de x e y); K – conjunto total de preditores;

Para avaliar todos os conjuntos possíveis, faz-se necessário aplicar o método um número da ordem de 2^K de vezes, o que se torna impraticável em conjuntos com um número elevado de preditores. Para contornar este problema, aplica-se o método heurístico, em que se avalia a importância dos preditores a partir da maximização do quociente de informação mútua (QIM_x) razão a seguir (Eq. 16):

$$\max QIM_x = \max \frac{I(x_k, y)}{\frac{1}{|S|} \sum_{i,j \in S} I(x_i, x_j)} \quad (16)$$

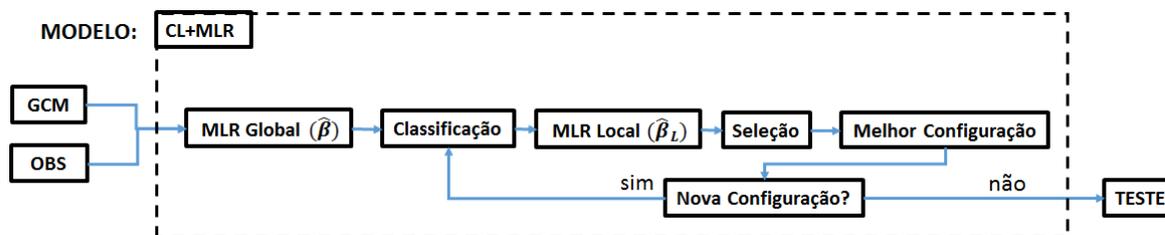
Em que: S – subconjunto de preditores;

No método aplicado, no entanto, modifica-se o procedimento baseado em *stepwise regression* – descrito na seção anterior (Seção 4.3.6) – a partir da substituição dos testes de significância por otimização baseada na Eq. 16. Mais detalhadamente, adiciona-se a cada passo uma variável segundo ordenamento decrescente de índices de QIM_x (*forward search*). Como critério de parada, avaliam-se os índices de qualidade sobre o conjunto de validação (definidos no bloco de Configuração Geral; Seção 4.2). Na metodologia proposta, o método MRMR se aplica à combinação de previsões de modelos de *downscaling* estatístico (modelo “CBLMR”).

4.3.8 CL – Classificação e Agrupamento de Padrões Sinópticos

A técnica de agrupamento de padrões sinópticos, proposta neste trabalho, é aplicada para fins de ajuste de modelos com arquitetura de parâmetros locais (Seção 2.3). Além do número e tipo de técnica aplicada ao agrupamento de padrões, o algoritmo desenvolvido realiza a configuração automática em relação a outras configurações, como métricas de distância entre padrões, tipo de vínculo entre *clusters* (Figura 15). A configuração inicial do modelo se baseia em parâmetros globais ajustados com o algoritmo apresentado na Figura 11, i.e., sem a aplicação de *clustering* (bloco “MLR global”; Seção 4.3.2). Em seguida, regressões locais (“MLR local”) serão realizadas para as classes ou grupos identificados no âmbito do bloco anterior (“Classificação”). Após o ajuste de regressões locais, faz-se a seleção de *clusters* sobre os quais se obtém aprimoramento na acurácia em relação às previsões obtidas com regressão global, armazenando informações sobre os melhores resultados obtidos até que haja convergência nos resultados, ou um critério de parada seja atingido (e.g., 10 iterações seguidas sem melhoria ou exaustão de todas as configurações disponíveis para o modelo).

Figura 15 – CL+MLR



Modelos de *downscaling* baseado em classificação de padrões sinópticos e regressão multilinear. Um processo de otimização de configuração para algoritmos de classificação também é ilustrado. Fonte: autor.

Para o bloco “Classificação”, ou de *clustering*, indicado na Figura 15, são consideradas duas abordagens para o agrupamento de variáveis predictoras: (a) agrupamento hierárquico aglomerativo (AHC, *Agglomerative Hierarchical Clustering*; PRYOR; SCHOOF; BARTHELMIE, 2005; VRAC; STEIN, M.; HAYHOE, 2007); (b) procedimento não hierárquico (NHC, *Non-Hierarchical Clustering*) baseado no algoritmo k-means (e.g., GUTIÉRREZ et al., 2004). Em relação às técnicas de agrupamento, há duas abordagens principais: No AHC, o número de clusters L é estabelecido à posteriori, em função de um critério de parada de fusões. No NHC, por sua vez, L é decidido à priori, em função das classes adotadas na inicialização¹⁴. No método implementado, as classes adotadas para a inicialização do NHC são obtidas dos grupos estabelecidos pelo AHC. A principal aplicação do AHC se dá com vistas à definição de número ótimo de grupos ou classes (L). Para isso, em abordagem não convencional voltada para a performance de regressões locais, considera-se um método de otimização baseado na avaliação de performance dos modelos (wrapper) que estabelece parâmetros ótimos para as configurações do AHC e NHC. Além do número de grupos L , o qual é aplicado como critério de parada para o AHC e número inicial de grupos no NHC, são selecionadas também configurações relativas às métricas de distância para o AHC e NHC, bem como ao algoritmo de agrupamento aplicado no AHC (Tabela 2).

¹⁴ É possível que o algoritmo de clustering não hierárquico resulte em um número de classes menor que o valor inicial. Isso se deve ao procedimento de estabilização de grupos formados, que reclassifica sucessivamente os elementos em função de um critério de parada. Portanto, a configuração inicial de classes e o critério de parada podem influenciar significativamente os resultados.

Três métricas de distância (d) são consideradas para as técnicas de AHC e NHC: i) euclidiana; ii) *cityblock* (soma de catetos); iii) cosseno do ângulo entre os vetores. Além disso, para a classificação baseada em AHC, consideram-se os seguintes cálculos de dissimilaridade (D), ou algoritmos de agrupamento: i) distâncias médias; ii) agrupamento pela mínima variância agregada (WARD, 1963); iii) vizinhos mais próximos; iv) vizinhos mais distantes.

O algoritmo de agrupamento está relacionado com o cálculo de “vínculo”, ou, com a minimização de dissimilaridades entre grupos. O cálculo de dissimilaridade estabelece um valor de distância representativo entre dois grupos. O valor representativo é estabelecido pelo algoritmo de agrupamento, o qual calcula distâncias para todas as combinações de pares formadas por elementos de dois grupos. No caso do agrupamento hierárquico aglomerativo, grupos com mais de um elemento ocorrem a partir da segunda fusão. Em seguida, o vínculo ou fusão de grupos é realizada considerando o par de grupos associado à menor distância representativa.

Tabela 2. Configuração geral de técnicas de agrupamento.

a) Distâncias entre padrões (P_t e P_s)	Equações
<i>Cityblock</i> (<i>Minkowski</i> , Ordem 1)	$d = \sum_{j=1}^m P_j(t) - P_j(s) $
Euclideana (<i>Minkowski</i> , Ordem 2)	$d = \left[\sum_{j=1}^m (P_{t,j} - P_{s,j})^2 \right]^{1/2} = \ P_t - P_s\ $
Cosseno	$d = 1 - \frac{P_t P_s'}{\sqrt{(P_t P_t') (P_s P_s')}}$
b) Dissimilaridades (D) entre G e H	Algoritmos
Vizinhos mais Próximos	$D(G, H) = \min (d(G_i, H_j)), i \in \{1, \dots, n_G\}, j \in \{1, \dots, n_H\}$
Vizinhos mais Distantes	$D(G, H) = \max (d(G_i, H_j)), i \in \{1, \dots, n_G\}, j \in \{1, \dots, n_H\}$
Dissimilaridade Média	$D(G, H) = \frac{1}{n_G n_H} \sum_{i=1}^{n_G} \sum_{j=1}^{n_H} d(G_i, H_j)$
Dissimilaridade por mínima variância agregada (Método de Ward)	$D(G, H) = \sqrt{\frac{2n_G n_H}{(n_G + n_H)}} \ \bar{G} - \bar{H}\ $

(a) Métricas adotadas para AHC e NHC; (b) Algoritmos de Fusão de AHC. Os padrões sinópticos P_t e P_s são vetores m-dimensionais contendo as variáveis da malha do GCM para os instantes t e s , respectivamente. G e H representam grupos arbitrários, as quais contêm n_G e n_H elementos associados, respectivamente. Note que G_i e H_j correspondem, respectivamente, ao i -ésimo e j -ésimo elemento de G e H . Com respeito ao método de Ward, \bar{G} e \bar{H} correspondem aos centroides dos padrões (P) pertencentes a uma classe G e H . O método de Ward apenas se aplica a distâncias euclidianas. Fonte: adaptado de Perruci (2018).

Após a classificação (considerando configurações iniciais para o AHC e NHC), a seleção de clusters é realizada de forma individual por meio do ajuste dos coeficientes para os conjuntos de instantâneos separados em cada grupo. A técnica de ajuste de parâmetros locais também segue estrutura descrita em Seção 4.3.2. Como mencionado no início da seção, a especificação inicial do modelo é dada por regressão multilinear ajustada sobre todo o período de dados de treinamento (arquitetura de parâmetros globais, Eq. 17a).

Estimativas para os instantes designados ao q -ésimo grupo são realizadas por meio de parâmetros locais (Eq. 17b), e superpostas à estimativa produzida pela regressão global. No bloco de “seleção”, o método de otimização implementado mantém grupos para os quais há melhora em relação ao índice de qualidade, quando comparadas a série inicial (produzida apenas com os parâmetros na Eq. 17a) e a série composta (estimada simultaneamente pelos parâmetros globais e locais, nas Eqs. 17a e 17b, respectivamente). A forma final do modelo é expressa pela Equação 17:

$$\hat{y}(t) = \begin{cases} \beta_0 + \sum_{j=1}^m \beta_j P_j(t) & , \forall n_q > \alpha \cdot n_T \\ \beta_{0,q} + \sum_{j=1}^m \beta_{j,q} P_j(t) & , \forall n_q \leq \alpha \cdot n_T \end{cases} \quad \begin{matrix} (17a) \\ (17b) \end{matrix}$$

em que: n_T – número total de ocorrências; n_q – Número de ocorrências no grupo de índice q ($= 1, 2, \dots, L$); $\alpha \in [0, 1]$ – fração de corte. Note que, demonstrado na Eq. 17, são empregados apenas os parâmetros locais derivados de grupos que possuem um número de ocorrências maior que o valor crítico $\alpha \cdot n_T$. Caso, essa condição não seja satisfeita, são adotados parâmetros globais para a simulação (17a).

Os parâmetros salvos da configuração selecionada são: i) número de grupos; ii) modo, AHC ou NHC; iii) métricas de distância para o AHC e NHC; iv) algoritmo de agrupamento (cálculo de dissimilaridade).

Os modelos baseados no algoritmo de agrupamento de padrões sinópticos recebem o termo “CL”.

4.4 Modelos baseados em complementaridade espectral

A quantificação da complementaridade espectral, tal como se define nesta seção, se fundamenta na inferência sobre relações de relevância e redundância entre variáveis preditoras (e.g., saídas de modelos de larga escala ou de *downscaling* estatístico) a partir de informações extraídas de domínio da frequência. Para a projeção de sinais para o domínio da frequência, aplica-se a transformada wavelet contínua, tendo em vista ser possível avaliar, a partir de tal técnica, aspectos globais e locais (em relação ao tempo) do comportamento espectral do sinal (Seção 2.6). Ressalta-se, no entanto, que os métodos propostos nesta seção se aplicam apenas ao treinamento de modelos estatísticos, não sendo necessário, por essa razão, sua aplicação durante operação em tempo real, sendo essa uma importante contribuição do trabalho.

Os modelos baseados em complementaridade espectral se dividem em dois grupos principais: i) modelos “SCW”, modelos baseados em regressão global, porém, com modificações locais às amostras durante o processo de ajuste de coeficientes (“pesos espectrais”); ii) modelos “e”, modelos baseados em agrupamentos de padrões modificados para a classificação da entropia cruzada entre os preditores e o alvo. A Tabela 3 resume as informações sobre os modelos propostos. A descrição das técnicas empregadas aos modelos baseados em complementaridade espectral global e local se encontram ao longo das Seções 4.4.1 e 4.4.3. Nas Seções 4.4.2 e 4.4.4 são mostrados exemplos ilustrativos a partir de sinais sintéticos com propriedades conhecidas, comportamentos estacionários e não estacionários, respectivamente.

Tabela 3 – Informações sobre modelos e variantes baseados em complementaridade espectral global e local.

Modelos	Métodos	Variante	Modelo mais próximos (sem quantificação de complementaridade espectral)

SCWs-V1 CBLSCW-V1	<u>Seleção de variáveis:</u> EN	V1: <u>Espaço de busca por coeficientes (A e B):</u> [-3,3]; <u>Otimização:</u> Levenberg-Maquardt	ENRs CBLEN
SCWs-V2 CBLSCW-V2	<u>Regressão global com modificação local:</u> Pesos espectrais (w_k ; Seção 4.4.3)	V2: <u>Espaço de busca por coeficientes (A e B):</u> [-1,1] <u>Otimização:</u> Grid Search	ENRs CBLSW CBLEN
CLpe/ CBCLe/ CBCLpe	<u>Seleção de variáveis:</u> (EN+ICE; Seção 4.4.1) <u>Agrupamento de padrões:</u> Entropia espectral cruzada (w_k ; Seção 4.4.3)	-	CLp / CBCL / CBCLp

Fonte: autor.

4.4.1 Complementaridade espectral global

Para a identificação de conjuntos que com maior complementaridade, propõe-se a aplicação de técnica voltada para a quantificação de relevância e redundância em um conjunto a partir da magnitude espectro estimado por meio da transformada “SSWT” (mais detalhes sobre a técnica SSWT estão disponíveis na Seção 2.6). Tal índice, de caráter global, deve ser capaz de identificar a “originalidade” nas contribuições de cada preditor, com penalização às informações (relevantes ou não) já contempladas.

Tal como definido nas Eqs. 18 e 19, considere-se a magnitude das transformada para o domínio de tempo-frequência de $y(t)$ (variável cujo comportamento se deseja prever) e de conjunto (sincronizado com y) contendo um número K de séries temporais de previsões $x_p(t)$ (saídas de modelos de larga escala ou de modelos de *downscaling* estatístico). Doravante, comunica-se a partir do termo “espectro” a magnitude dos números complexos obtidos por meio de aplicação de transformadas para o domínio da frequência. A partir do (auto)espectro de Y e do conjunto X_m , calcula-se, em pares, o espectro cruzado entre os preditores (X_k) e o preditando (Y), “ E_k ” (Eq. 20), bem como o espectro cruzado dos preditores entre si, “ $E_{i,j}$ ” (Eq. 21).

$$Y(\omega, t) = |SSWT(y(t))|, \quad \omega \in [1, 2, \dots, \Omega]; t \in [1, 2, \dots, T] \quad (18)$$

$$X_k(\omega, t) = |SSWT(x_k(t))|, \quad k \in [1, 2, \dots, K] \quad (19)$$

$$E_k(\omega, t) = (X_k(\omega, t) \cdot Y(\omega, t))^{0,5} \quad (20)$$

$$E_{i,j}(\omega, t) = (X_i(\omega, t) \cdot X_j(\omega, t))^{0,5}, \quad i \neq j, \quad i, j \in [1, \dots, K] \quad (21)$$

Propõe-se, em seguida, que seja calculado atributos “globais” (em relação ao tempo) a partir dos conjuntos de espectros (E), do qual resultam os periodogramas modificados, “ REL ” e “ RED ”, os quais transportam informações sobre a relevância e a redundância espectral no conjunto de preditores, respectivamente (Eqs. 22 e 23):

$$REL(\omega) = \frac{1}{K} \sum_{k=1}^K \left[Prc_{\forall t}(E_k(\omega, t), \alpha_0) \right], \quad \alpha_1 \in [0,1] \quad (22)$$

$$RED(\omega) = \sum_{\substack{i>j \\ i,j \in K}} Prc_{\forall t}(E_{i,j}(\omega, t), \alpha_1), \quad \alpha_2 \in [0,1] \quad (23)$$

Em que: $prc(E, \alpha)$ – percentil α calculado sobre os instantes de tempo do espectro E (em cada frequência); α_0, α_1 – coeficientes de sensibilidade aplicados à relevância e redundância, respectivamente. Para cada frequência ω , os valores armazenados no vetor REL consistem em valores de referência obtidos a partir dos espectros cruzados ao serem “colapsadas” informações temporais e sobre os diferentes modelos. Entretanto, ressalta-se a distinção no cálculo da redundância, que em lugar de valor médio sobre os diferentes modelos, calcula-se a soma. Dessa forma, em função do aumento do número de preditores no conjunto avaliado, na medida em que informações sobre o preditendo são contempladas, os valores associados à relevância atingem nível de “saturação” (assíntota). Os valores de α nas Eqs. 22 e 23 apenas representam generalização do método, permitindo maior controle sobre o valor a ser extraído de cada distribuição (e.g., para $\alpha = 0,5$ aplica-se a mediana). Os valores associados à redundância, por sua vez, crescem continuamente em razão da adição de informações já contempladas anteriormente. Note-se que o parâmetro α referido nesta seção é específico da abordagem baseada em complementaridade espectral, sendo distinto daquele apresentado na Seção 4.3.5.

Para normalizar os vetores REL e RED , são sugeridas as expressões a seguir (Eqs. 24 e 25). Note-se que, para a normalização da redundância, considera-se a união com os valores de relevância (mantendo a inferência da relevância independente do conteúdo redundante no conjunto). Para evitar a penalização de

valores de relevância nos casos em que $RED > REL$, valores de redundância são desconsiderados no denominador da Eq. 24. Para a análise realizada neste trabalho, os valores de relevância espectral são considerados independentes da redundância no conjunto.

$$REL_n(\omega) = \frac{REL(\omega)}{\max_{\forall \omega}(REL(\omega))} \quad (24)$$

$$RED_n(\omega) = \frac{RED(\omega)}{\max_{\forall \omega}(REL \cup RED)} \quad (25)$$

Os valores normalizados REL_n e RED_n são, por fim, combinados tal como se expressa na Eq. 26, sobre o cálculo do “Índice de Complementaridade Espectral” (ICE). No numerador, calcula-se a soma do vetor REL_n para a quantificação da relevância, enquanto – no denominador – somam-se os termos compostos pela soma da média (temporal) do autoespectro normalizado da série de observações (\bar{Y}_n). O parâmetro α_2 é aplicado para regular a intensidade da penalização à redundância sobre o valor realizado para a complementaridade espectral, sendo complementar aos parâmetros α_0 e α_1 nas Eqs. 22 e 23 (os quais controlam a sensibilidade geral à relevância/redundância). No caso ideal, a relevância espectral coincide com o autoespectro do sinal observado, enquanto a redundância é nula, do qual resulta que $ICE = 1$. Por outro lado, quando há maior redundância no conjunto, ao passo que há pouca relevância, o índice ICE tende a zero.

$$ICE = \frac{\sum_{\omega} REL_n(\omega)}{\sum_{\omega} (\bar{Y}_n(\omega)) + \alpha_2 \cdot (\sum_{\omega} RED_n(\omega))} \quad (26)$$

No presente trabalho, aplica-se o método acima com vistas à seleção de variáveis para fins de classificação e regressão em modelos de *downscaling* e de combinação. Isto é realizado a partir da avaliação de conjuntos pré-definidos (por modelo de seleção baseado em “Lasso”), tal como no caso descrito na Seção 4.3.5, no entanto, para os modelos baseados em complementaridade espectral, aplica-se o Índice de Complementaridade Espectral (ICE) como critério de otimização.

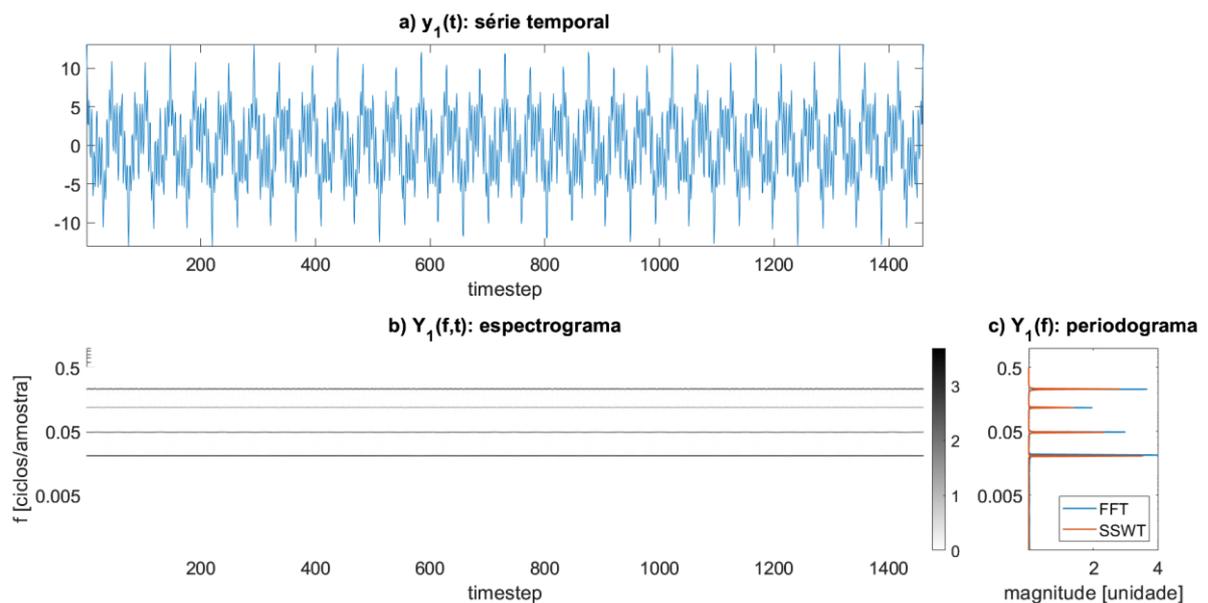
4.4.2 Exemplo: complementaridade espectral global

Para exemplificar os atributos baseados em complementaridade espectral apresentado na Seção 4.4.1, nesta seção, empregam-se sinais sintéticos com propriedades conhecidas a fim de demonstrar a avaliação de complementaridade espectral global.

A Figura 16 exemplifica um sinal estacionário $y_1(t)$, formado a partir da combinação linear de harmônicos cujas frequências características equivalem a 30, 70, 170 e 330 ciclos contidos em um período composto por 1400 amostras, com amplitudes de 4,3,2 e 4, respectivamente. No exemplo ilustrado, o sinal $y_1(t)$ possui estrutura de frequências constante em relação ao tempo, como mostra o espectrograma na Figura 16b e a Eq. 27. Na representação global, observa-se que as amplitudes A_k são proporcionais à magnitude do espectro nas frequências ω_k , havendo boa correspondência entre os espectros inferidos pelas técnicas FFT e SSWT.

$$y_1(t) = \sum_{k=1}^{k=4} A_k \cos(\omega_k t), \quad k = 1,2,3,4 \quad (27)$$

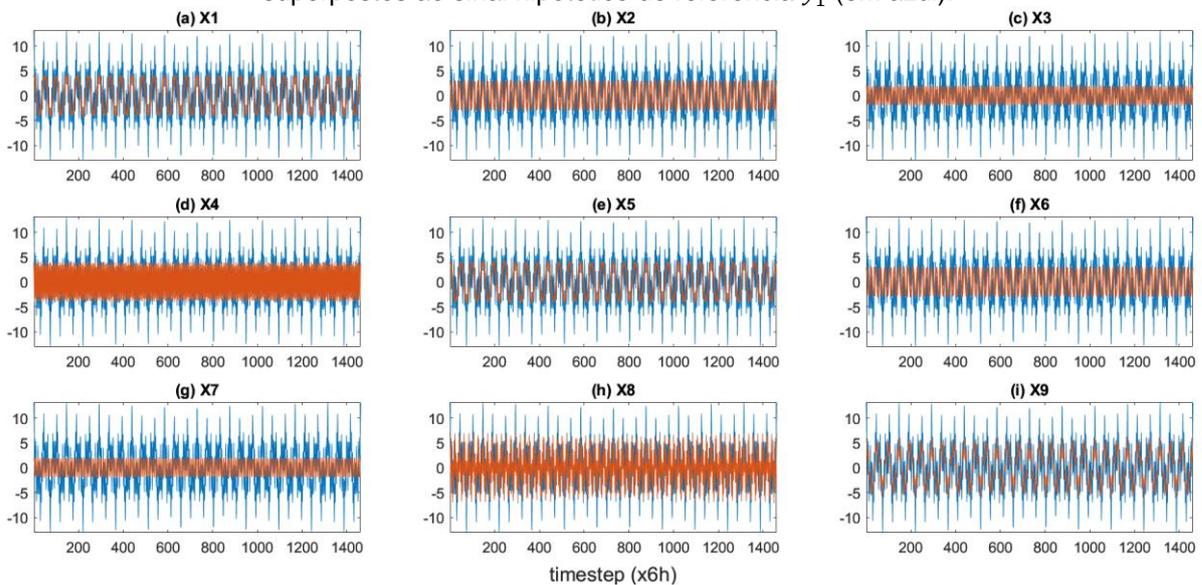
Figura 16 – Sinal sintético estacionário: combinação de harmônicos puros.



(a) série temporal; (b) espectrograma: espectro local (*synchrosqueezed wavelet transform*; SSWT); (c) periodograma: espectro global (*Fast Fourier Transform* – FFT; em azul; em vermelho: média temporal do espectrograma obtido com a SSWT). Fonte: autor.

Tal como demonstra a Figura 17, para avaliar sinais de previsões candidatos à descrição de comportamento tal qual se representa por $y_1(t)$ (em azul), considerem-se os sinais descritos a partir de nove sinais $x_k(t)$ (em vermelho). Os quatro primeiros sinais em (a), (b), (c) e (d) contêm um dos modos de frequência característicos de $y_1(t)$. Os sinais seguintes nos gráficos de (e) até (i) são redundantes em relação aos quatro primeiros, sendo os dois últimos a combinação linear, entre x_2 e x_4 , e entre x_1 e x_3 , respectivamente.

Figura 17 – Complementaridade espectral global: sinais hipotéticos de preditores (em vermelho) superpostos ao sinal hipotético de referência y_1 (em azul).

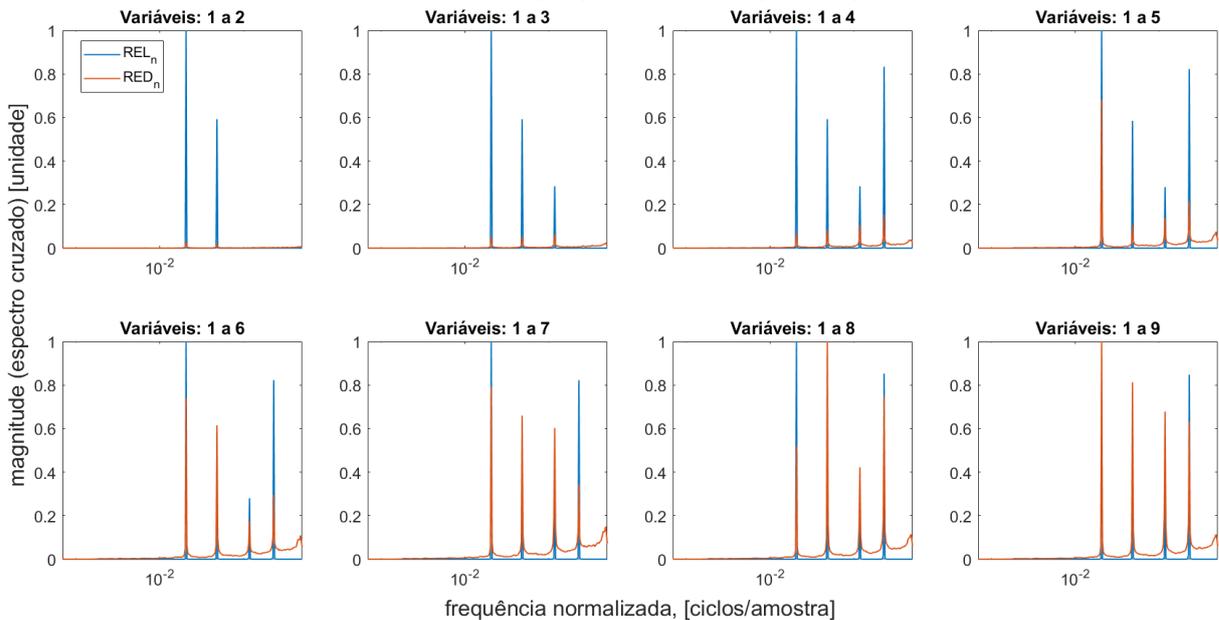


Fonte: autor

Os gráficos apresentados na Figura 18 demonstram os valores calculados, em função da frequência, para a relevância e a redundância espectral normalizada, REL_n e RED_n (em azul e vermelho, respectivamente), quando considerados conjuntos formados a partir do acúmulo dos sinais 1 a k , para $k > 1$. No exemplo ilustrado, para o cálculo da redundância, foram considerados os valores para α_1 e α_2 de 0,5 e 1, respectivamente (Eqs. 22 e 23). Observa-se, a partir da inclusão das quatro primeiras variáveis, a contabilização da relevância espectral em cada uma das frequências compartilhadas entre x_k e y . Após a adição de novas variáveis ao conjunto, as quais possuem apenas informações redundantes (havendo já sido contempladas as frequências estruturais do sinal de referência y), os valores de redundância espectral aumentam, o que reduz o ICE (definido na Eq. 26).

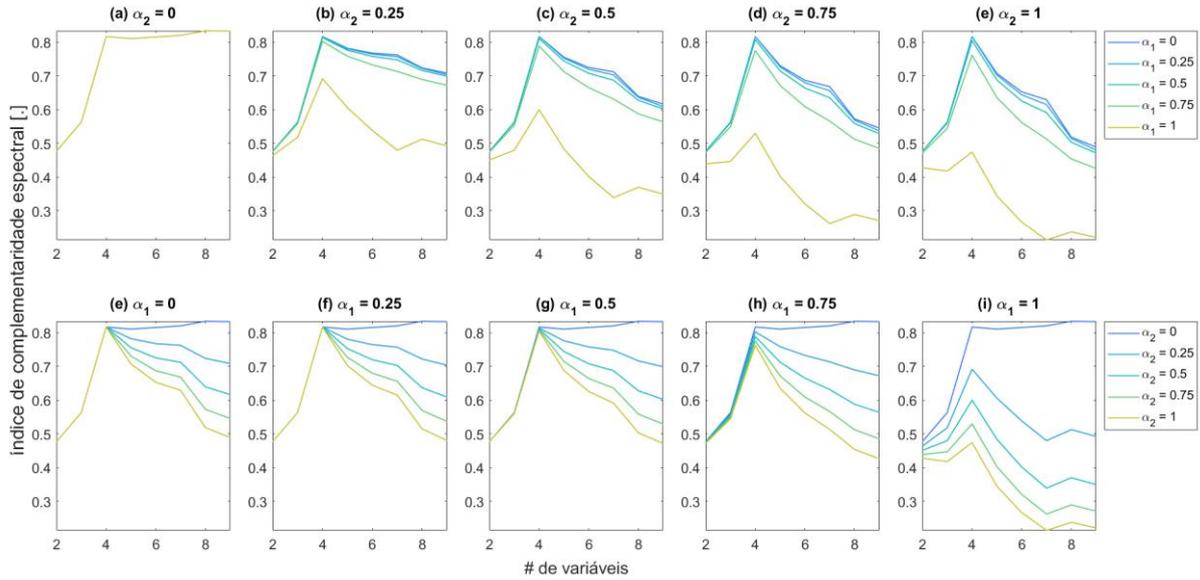
Na Figura 19 são apresentados, para cada conjunto de variáveis formado (a partir do acúmulo de variáveis), os valores calculados para o *ICE* considerando diferentes pares de parâmetros α_1 e α_2 . Como se demonstra, Para $\alpha_2 = 0$, desconsidera-se a redundância entre as variáveis. Como consequência, obtêm-se um aumento rápido até o acúmulo da 4ª variável, seguido de um crescimento lento em função do aumento do conjunto de analisado. Para outros valores de α_2 , observa-se um ponto máximo de complementaridade espectral, com redução da complementaridade a partir deste ponto. De forma geral, os valores de α_1 regulam a sensibilidade geral em relação à redundância, enquanto os valores de α_2 regulam a “intensidade” com que se penaliza o índice em função da redundância percebida. Em cada aplicação, a definição de valores ótimos para α_1 e α_2 é realizada por meio de métodos baseados em otimização por validação cruzada (*wrappers*).

Figura 18 – Valores de relevância e redundância espectral considerados 8 conjuntos formados a partir do acúmulo de sinais.



Fonte: autor.

Figura 19 – Análise de sensibilidade de índices de complementaridade espectral em função dos parâmetros α_1 e α_2 .



4.4.3 Complementaridade espectral local

Para a avaliação de sinais não estacionários, propõe-se a aplicação da informação temporal de espectrogramas com vistas ao agrupamento de padrões para regressões locais. No entanto, para resumir a informação associada com diferentes valores de frequência, propõe-se o cálculo de entropia espectral a partir da conversão dos espectrogramas (Eqs. 18 e 19) em “PMFs” (*Probability Mass Function*; Eqs. 28 e 29). O cálculo da entropia é um indicativo do aspecto geral da PMF aplicada, sendo igual a 0 se toda a probabilidade se concentra em um único “evento” (dentro um espaço de eventos). Ao se aproximar de aspecto uniforme (eventos equiprováveis), o cálculo da entropia resulta em 1, indicando baixa previsibilidade no sinal em questão (comportamento “aleatório”). Ao ser avaliada sobre o espectro de um sinal, por outro lado, a entropia informa sobre a concentração da energia espectral em torno de uma única frequência, sendo igual a um para um harmônico puro qualquer, enquanto tende a zero no caso de um sinal aleatório (i.e., ruído), em que todas as frequências possuem magnitudes próximas. A partir das PMFs, calcula-se a entropia espectral normalizada do alvo y , H_Y (Eq. 30), bem como a entropia cruzada entre x e y (Eq. 31).

$$\text{PMF}_Y(\omega, t) = \frac{Y(\omega, t)}{\sum_{\omega} Y(\omega, t)} \quad (28)$$

$$\text{PMF}_X(k, \omega, t) = \frac{X_k(\omega, t)}{\sum_{\omega} Y(\omega, t)}, \quad k \in [1, 2, \dots, K] \quad (29)$$

$$H_Y(t) = - \sum_{\omega=1}^{\Omega} \text{PMF}_Y(\omega, t) \cdot \log(\text{PMF}_Y(\omega, t)) / \log(\Omega) \quad (30)$$

$$H_{XY}(k, t) = - \sum_{\omega=1}^{\Omega} \text{PMF}_Y(\omega, t) \cdot \log(\text{PMF}_X(k, \omega, t)) / \log(\Omega) \quad (31)$$

Em que: k – índice do preditor; ω – índice de frequência; t – índice de tempo; Ω – número de frequências descritas;

O cálculo da entropia cruzada permite inferir a similaridade global entre espectros dos sinais x e y para cada instante de tempo t , que se apresenta quando $\text{PMF}_X(k, \omega, t) \rightarrow \text{PMF}_Y(\omega, t)$, que é a condição para que $H_{XY} \rightarrow H_Y$ (Eqs. 30 e 31). O método proposto nesta seção fundamenta novos modelos baseados em classificação de padrões sinópticos, os quais agrupam padrões de H_{XY} em lugar de padrões de estados atmosféricos. Tal mudança se alinha com a hipótese levantada no presente trabalho, que se direciona à extração de informações úteis presentes na estrutura de covariâncias entre os sinais, supondo ser possível, dessa forma, realizar ajustes mais acurados aos modelos.

Para que seja considerado o caráter probabilístico da ocorrência de eventos atmosféricos, neste trabalho, propõe-se modificação nos valores das PMFs (Eqs. 28 e 29) para que eventuais casos em que $\text{PMF}(\omega, t) \rightarrow 0$ sejam substituídos por valor mínimo de probabilidade, evitando singularidades nas Eqs. 30 e 31. Dessa forma, em um dado instante de tempo t , considerem-se os valores de probabilidade inferiores¹⁵ a $1/\Omega$, os quais estão associados ao conjunto de frequências $\zeta(t)$ (Eq. 32). Sobre tal conjunto se realiza modificação tal como indica a Eq. 33:

$$\zeta(t) = \{ \omega \mid \text{PMF}(\omega, t) < 1/\Omega \} \quad (32)$$

$$\text{PMF}(\zeta, t) = \frac{1}{\Omega} \sum_{\omega \in \zeta(t)} \text{PMF}(\omega, t) \quad (33)$$

Como mencionado na Seção 2.3, modelos (com arquitetura de parâmetros) locais são ajustados a subconjuntos das amostras. Entretanto, propõe-se, a partir do cálculo entropia cruzada na Eq. 31, método voltado para ponderar amostras do conjunto treinamento em relação à similaridade/dissimilaridade espectral do preditor

¹⁵ Note-se que o valor de corte $1/\Omega$ está associado com um sinal qualquer puramente aleatório, com eventos equiprováveis (i.e., descritos por distribuição de probabilidades uniforme).

x sobre o preditando y , representadas por PMF_X e PMF_Y , respectivamente. Nesse estudo, tal método de ponderação é aplicado para fins de ajuste de modelos globais, no intuito de avaliar possíveis contribuições em relação a outros modelos similares. Isso é realizado por meio de vetor de pesos espectrais w que altera artificialmente o sinal de resíduos sobre o qual se ajustam os coeficientes de regressão e conseqüentemente a estrutura de covariâncias percebida pelo modelo, forçando-o a dar maior ênfase ao erro de amostras específicas.

A ponderação realizada por meio de w_k leva em consideração, por um lado, a diferença entre o ângulo de fase θ entre os sinais, por outro lado, diferenças na magnitude do espectro, sendo estes atributos diretamente relacionados com a estrutura de covariâncias compartilhada entre dois sinais (Eqs. 35 a 38). Por fim, considera-se a operação na Eq. 39 para que $w_k(t)$ assumam propriedades de vetor de ponderação (ou de “pesos”), cuja soma é igual a 1.

$$w_k(t) = \left(\frac{F_k(t)}{\max(F_k(t))} \right)^A \left(\frac{M_k(t)}{\max(M_k(t))} \right)^B, \quad A, B \in \mathbb{R} \quad (34)$$

$$M_k(t) = \frac{2 \cdot (H_{XY}(k, t) - H_Y(t))}{H_{XY}(k, t) + H_Y(t)} \quad (35)$$

$$F_k(t) = 1 - \frac{1}{\Omega\pi} \sum_{\omega}^{\Omega} \left| |\theta_Y(\omega, t)| - |\theta_X(k, \omega, t)| \right|, \quad \theta \in [-\pi, \pi] \quad (36)$$

$$\theta_Y(\omega, t) = tg^{-1} \left\{ \frac{Im[SSWT(y)]}{Re[SSWT(y)]} \right\} \quad (37)$$

$$\theta_X(k, \omega, t) = tg^{-1} \left\{ \frac{Im[SSWT(x_k)]}{Re[SSWT(x_k)]} \right\} \quad (38)$$

$$w_{N,k}(t) = \frac{w_k(t)}{\sum_{t=1}^T w_k(t)} \quad (39)$$

Os valores A e B permitem que os pesos espectrais w_k enfatizem amostras em diferentes circunstâncias, e.g., ao assumirem valores negativos, serão mais importantes as amostras de instantes cujo espectro está bem representado em relação à fase (A) e/ou amplitude de variações (B). Se assumem valores positivos, por outro lado, destacam-se as amostras de instantes com menor similaridade espectral entre os sinais comparados. Caso $A = B = 0$, não haverá alteração em

relação ao ajuste dos coeficientes. A modificação da importância das amostras se dá a partir de método minimização ponderado de erros baseado em mínimos quadrados. Destaca-se, no entanto, que os pesos espectrais w_k se aplicam a qualquer método regressivo (e.g., *Generalized Least Squares*, *Support Vector Machines*, redes neurais etc.). Para encontrar os valores ótimos de A e B , consideram-se dois métodos distintos de otimização baseada em validação cruzada: i) Levenberg-Maquardt; ii) Grid Search. No primeiro caso, considera-se intervalo $[-3,3]$ para ambos os parâmetros. No segundo, por sua vez, realiza-se busca considerando o intervalo $[-1,1]$.

Para a modificação dos coeficientes de modelos lineares, propõe-se aplicar o modelo regressivo baseado no algoritmo “MLR” (Seção 4.3.2), entretanto, modificando a Eq. 7 para que sejam introduzidos os pesos por meio de matriz W , como se expressa nas Eqs. 40 e 41.

$$\hat{\beta} = (X^T W X)^{-1} X^T W y \quad (40)$$

$$W_k = \begin{bmatrix} w_k(t) \\ w_k(t+1) \\ \vdots \\ w_k(T) \end{bmatrix} \quad (41)$$

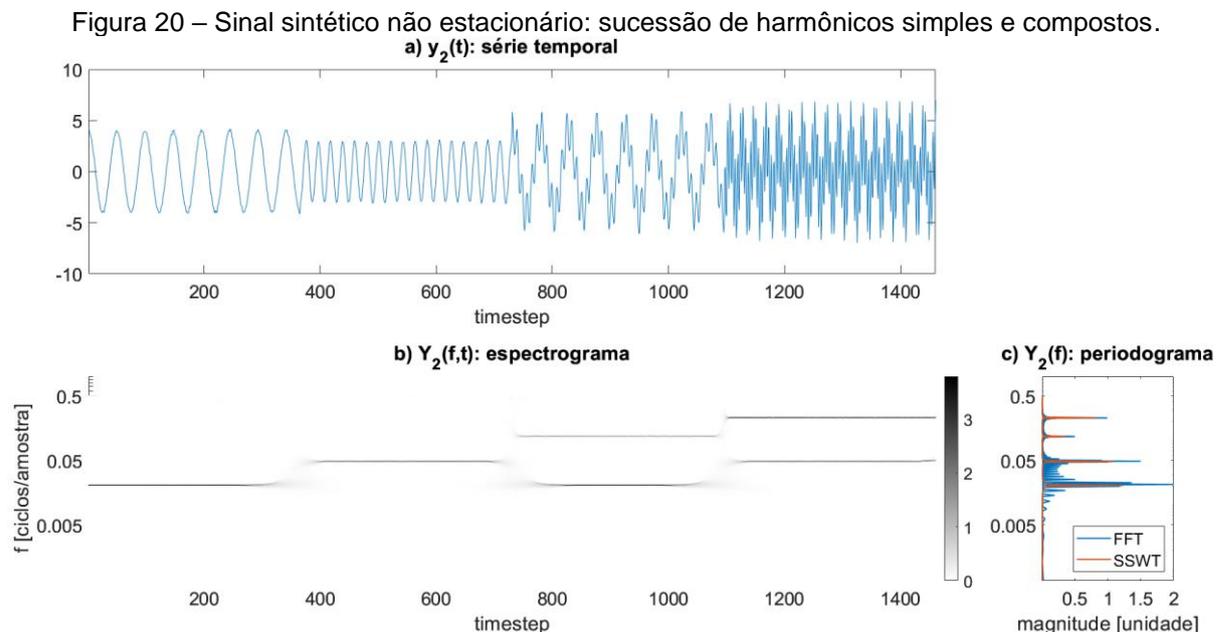
O vetor de pesos w também pode ser explorado para fins de classificação, entretanto, neste trabalho, aplica-se a entropia cruzada para este fim (Eq. 31). Nas modificações propostas aos modelos com base nos pesos espectrais, como abordagem inicial, propõe-se a aplicação da saída do modelo estatístico (com $A, B = 0$) como sinal a partir do qual são extraídos os pesos espectrais. Em seguida, os resultados da regressão inicial são modificados em função de diferentes valores para A e B . Além disso, como se indica na Eq. 34, propõe-se o produto entre os termos associados com a magnitude e fase, fazendo com que haja impacto sobre a importância de uma determinada amostra caso a fase ou a magnitude no entorno não estejam bem descritas. Vale ressaltar que as projeções para o domínio da frequência baseadas em transformada *wavelet* permitem a identificação de relações não lineares, podendo ser subaproveitadas em estritamente lineares.

4.4.4 Exemplo: Complementaridade espectral local

Para exemplificar os métodos baseados em complementaridade espectral local, descritos na Seção 4.4.3, adota-se o sinal sintético não estacionário $y_2(t)$ como referência (preditando), tal como se define na Eq. 42. O sinal é composto pela sucessão entre 4 harmônicos sinais, cujas características estatísticas estão definidas na Eq. 42.

$$y_2(t) = \begin{cases} A_1 \cos(\omega_1 t), & t = 1, \dots, T/4 \\ A_2 \cos(\omega_2 t), & t = T/4 + 1, \dots, T/2 \\ A_1 \cos(\omega_1 t) + A_3 \cos(\omega_3 t), & t = T/2 + 1, \dots, 3T/4 \\ A_2 \cos(\omega_2 t) + A_4 \cos(\omega_4 t), & t = 3T/4 + 1, \dots, T \end{cases} \quad (42)$$

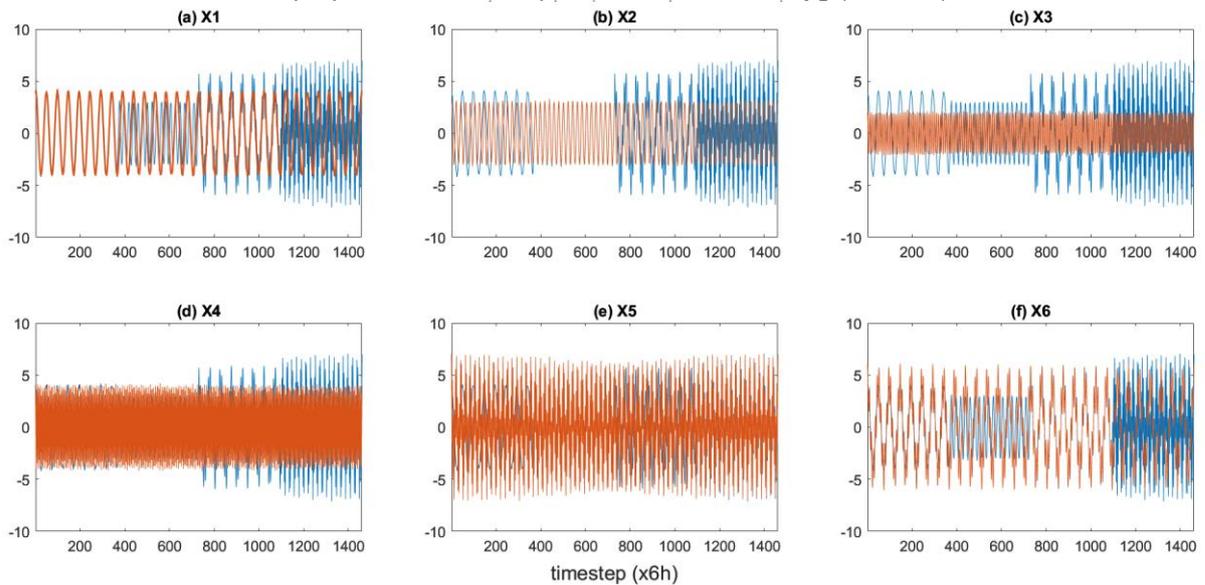
Os gráficos mostrados na Figura 20 apresentam o sinal $y_2(t)$ sob as perspectivas do domínio do tempo (a) e da frequência, em (b) e (c). Observa-se, a partir do espectrograma, a estrutura local de frequências do sinal $y(t)$, enquanto, por meio do periodograma ao lado, faz-se a inferência da estrutura global de frequências. O gráfico em (c) compara os espectros globais inferidos por meio de transformada rápida de Fourier (FFT) e por meio da transformada *Synchrosqueezed Wavelet Transform* (SSWT). Observa-se maior precisão em relação aos harmônicos introduzidos quando aplicada a SSWT, em comparação com espectro de Fourier (que possui limitações para a descrição de comportamentos não estacionários). Ressalta-se com tal comparação a importância de aplicação de técnica apropriada para a projeção ao domínio da frequência para casos em que há comportamentos não estacionários nos sinais dedicados ao treinamento de modelos estatísticos.



(a) série temporal; (b) espectrograma: espectro local (SSWT); (c) periodograma: espectro global (*Fast Fourier Transform* – FFT; em azul; em vermelho: SSWT). Fonte: autor.

Nos gráficos da Figura 21, consideram-se os sinais dos preditores hipotéticos mostrados (em vermelho) superpostos ao sinal de referência y_2 (em azul). Observa-se que, cada preditor x , em determinado período, possui características semelhantes às do sinal y_2 . Mais especificamente, o sinal representado por x_1 coincide com o primeiro período em y_2 , enquanto x_2 se assemelha ao segundo. No terceiro e quarto período, por outro lado, o sinal y_2 apresenta estrutura de frequência composta pela superposição de dois harmônicos simples. Dentre os sinais adotados, aqueles que possuem estrutura composta, similar àquela que se apresenta em y_2 , são os sinais x_5 e x_6 . Nos dois últimos períodos, os sinais x_3 e x_4 possuem informação apenas sobre um dos harmônicos componentes de y_2 .

Figura 21 – Complementaridade espectral local: sinais hipotéticos de preditores (em vermelho) superpostos ao sinal hipotético de referência y_2 (em azul).

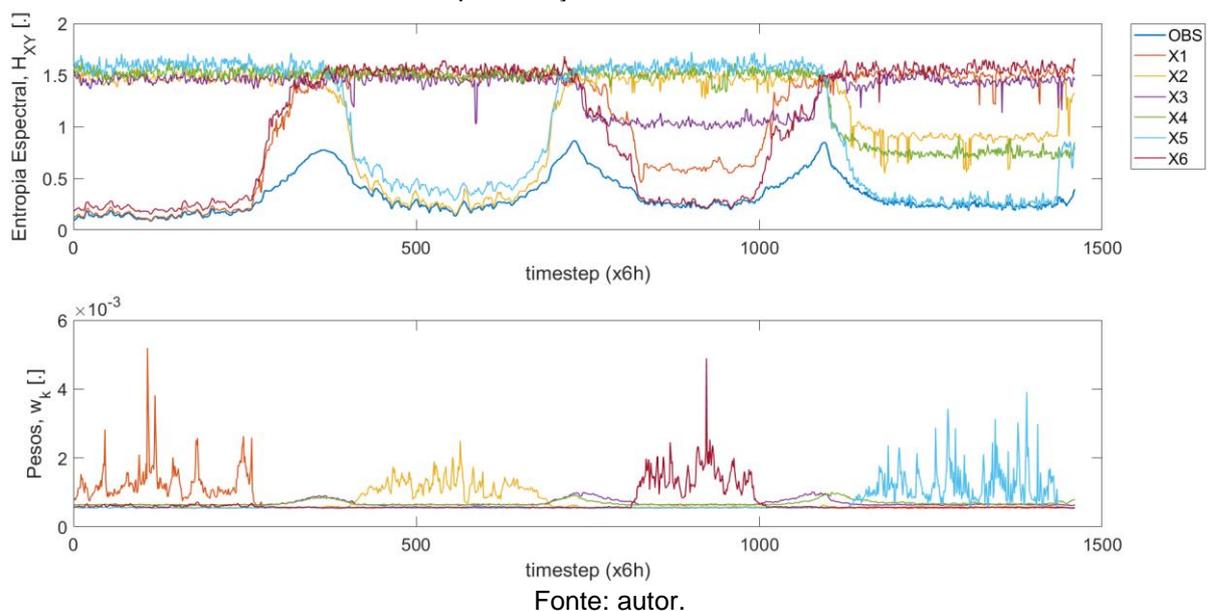


Fonte: autor.

Na Figura 22, as informações supracitadas podem ser visualizadas no primeiro gráfico, a partir do cálculo da entropia espectral cruzada H_{XY} e de pesos w_k (Eqs. 31 e 34). Quando há similaridade entre os espectros comparados, a entropia cruzada tende à entropia da observação hipotética (“OBS”), uma vez que $PMF_X(k, \omega, t) \rightarrow PMF_Y(\omega, t)$ no cálculo da Eq. 31. No segundo gráfico, por sua vez, são mostrados os pesos w_k calculados a partir da Eq. 34, considerando os coeficientes $A = -0,1$ e $B = -0,5$, os quais, ao assumirem valores negativos, aumentam a importância de amostras pertencentes a períodos bem representados no domínio da frequência. A partir da entropia cruzada, é possível verificar maior aptidão dos sinais x_1 e x_6 no

primeiro período, e dos sinais x_2 e x_5 no período seguinte. No terceiro e quarto período, observa-se que os sinais x_6 e x_5 demonstram maior similaridade espectral em relação à magnitude espectral do sinal de referência y_2 , que expressa a amplitude de variações dos harmônicos do sinal. Os picos apresentados pela entropia espectral do sinal observado (“OBS”) resultam do espalhamento da energia espectral em torno dos pontos de transição de “regime”, o que torna o aspecto das PMFs inferidas mais homogêneo (efeito mais acentuado em baixas frequências, como pode ser visto no espectrograma da Figura 20b). No segundo gráfico, por sua vez, é possível observar que, em cada trecho (não considerando as faixas de transição), apenas os sinais x cujos espectros mais se assemelham à referência em cada trecho se destacam em relação aos pesos, demonstrando o impacto da contabilização conjunta da similaridade nas estruturas de fase e frequência do sinal.

Figura 22 – Complementaridade espectral local: entropia espectral cruzada e pesos para a ponderação de amostras.



4.5 Parametrização de Modelos

De maneira geral, no diagrama de blocos ilustrado na Figura 23, três camadas principais de modelos são consideradas. Na primeira, são formados diferentes conjuntos de previsões do modelo de larga escala (GFS), incluindo domínios “fixos” (pré-concebidos por um número de pontos em torno do local de interesse) e domínios formados a partir de seleção de pontos na malha mais ampla em torno do ponto de interesse (24 x 24 ou 576 pontos), são aplicados na camada seguinte com vistas ao

ajuste dos coeficientes das funções de transferência dos modelos de *downscaling* estatístico, na camada seguinte. Na segunda camada, os diferentes modelos (IBL, M1, M2, M3 etc.) são parametrizados assumindo como entrada as previsões, no nível superficial da irradiância e temperatura ambiente fornecidas pelo modelo de larga escala. Por fim, uma camada final de modelos realiza previsões combinando as saídas da camada anterior (*downscaling* estatístico ou modelos “individuais”).

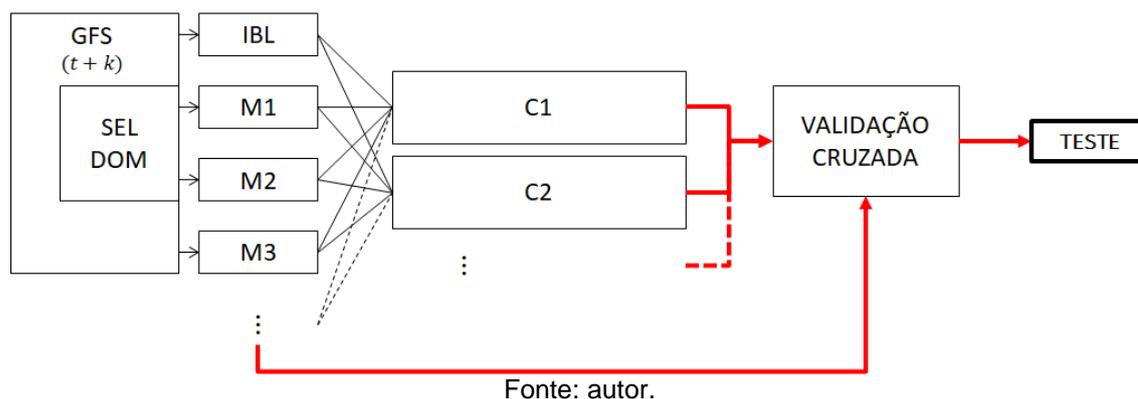
O método de seleção e validação de modelos foi realizado por meio de estratégia de validação cruzada baseada em “*hold out*”; (CURRY; KAMP, VAN DER; MONAHAN, A. H., 2012). Em relação à acurácia dos modelos, a técnica de *hold out* é escolhida, uma vez que oferece meios de avaliação a partir de um único ajuste, o que reduz o custo computacional consideravelmente em caso de ajustes de elevado esforço computacional (e.g., redes neurais). Para evitar vieses nos parâmetros devido à escolha do período de treinamento, serão consideradas configurações de validação cruzada distintas, as quais serão submetidas a um mesmo período adotado para a simulação da operação em tempo real (“teste”, Seção 4.6). A avaliação é realizada em relação ao índice de qualidade estabelecido no bloco de configurações gerais (Seção 4.2) e tem por finalidade a seleção objetiva de configurações para os modelos. Mais detalhes sobre os índices de qualidade adotados estão disponíveis na Seção 4.6 (sobre a avaliação de resultados).

A variedade de modelos de *downscaling* estatístico visa conferir ao resultado combinado maior acurácia e robustez, tendo em vista que as diferentes abordagens individuais apresentam diferentes aptidões. Vale ressaltar que, para a avaliação da performance dos modelos “M”, a IBL é adotada como a referência contra a qual são comparados todos os demais modelos (individuais ou combinados), uma vez que a IBL representa a saída do GCM diretamente sobre o ponto de interesse. Isso é realizado por meio de interpolação bilinear aplicada ao quadrículo em cujos vértices estão dispostos os quatro pontos da malha do GCM mais próximos ao local de interesse (Seção 4.3.1). Dessa forma, o modelo IBL não assimila a amplitude de variações e a média do sinal observado, embora transporte informações dos quatro pontos mais próximos ao local de interesse por meio de ponderação pelas distâncias geométricas das previsões do GFS. Ressalta-se, no entanto, que devido à ausência de procedimentos de ajuste baseados na minimização de gradientes de erro, faz-se necessário a modelagem física do sistema de geração para capacitar o modelo IBL à conversão de velocidade do vento ou irradiância solar em potência produzida.

Contudo, tais modelos estão fora do escopo do presente trabalho. Por essa razão, para ater-se à avaliação do desempenho de modelos estatísticos de curva de potência, nesta seção, adota-se a regressão multilinear dos quatro pontos mais próximos (“MLR1”; Seção 4.3.2) como modelo de referência alternativamente à interpolação bilinear (“IBL”).

Os modelos de combinação (C1, C2 etc.) são aplicados com o objetivo de identificar e explorar relações de complementaridade entre os modelos individuais, buscando uma acurácia mais elevada que a dos modelos individuais. Por outro lado, na hipótese de não haver tal relação de complementaridade entre os modelos individuais, o propósito da combinação passa a ser o de reproduzir o melhor resultado individual com suficiente fidelidade.

Figura 23 – Diagrama de blocos da metodologia baseada em *downscaling* estatístico e combinação de modelos com vistas à previsão do comportamento do vento local.



Como base nos algoritmos descritos na Seção 4.3, foram considerados 5 grupos de modelos de *downscaling* estatístico (camada de blocos “M” na Figura 23) e de combinação de modelos (camada de blocos “C”), são eles: Regressão Linear Múltipla, Análogos, *Clustering*, Complementaridade Espectral e seleção de atributos. Uma breve descrição sobre tais algoritmos é dada a seguir:

A1 – Regressão Linear Múltipla: Algoritmo composto por regressão multilinear acoplada à seleção iterativa de variáveis regressoras. A seleção de variáveis regressoras pode se dar através dos procedimentos de “poda de coeficientes” (GAO, L.; SCHULZ, Karsten; BERNHARDT, Matthias, 2014; HAMMAMI *et al.*, 2012; TIBSHIRANI, 1996), ou eliminação iterativa – *backward elimination* – (HAMMAMI *et al.*, 2012; PERRUCCI, 2018) associada à otimização por validação cruzada (*wrapper-based optimization*) (GUYON, ISABELLE; ELISSEFF, 2003). Os modelos

desenvolvidos com ênfase neste algoritmo de regressão linear múltipla identificam-se pelas siglas contendo “MLR” (Seção 4.3.2).

A2 – Análogos: Algoritmo baseado na classificação de padrões sinópticos que pressupõe que a similaridade entre 2 estados sinópticos implica em similaridade entre observações (i.e., similaridade na escala local). Mais detalhadamente, o modelo permite a utilização de observação do vento local (dados históricos) como preditor por meio da associação entre os instantes de tempo “atual” e um determinado instante anterior, sendo o último obtido a partir de minimização da métrica de distância que quantifica a similaridade entre estados atmosféricos (WETTERHALL; HALLDIN; XU, 2005). Considerando o conjunto de preditores formado pelas previsões geradas pelo modelo numérico associadas a dados observacionais históricos (selecionados a partir da similaridade entre estados atmosféricos), adota-se o procedimento descrito no algoritmo base “MLR” com vistas à seleção de variáveis e parametrização da função de transferência. Os modelos desenvolvidos com ênfase neste algoritmo de análogos identificam-se pela sigla “AN” (Seção 4.3.4);

A3 – Clustering: Algoritmo baseado em modelo com arquitetura de parâmetros locais, i.e., especializados para grupos de dados (ou subconjuntos) formados por meio de técnicas de agrupamento de padrões (ou estados de operação) de dois tipos: i. agrupamento hierárquico aglomerativo (GUTIÉRREZ *et al.*, 2004); ii. agrupamento não hierárquico (ROUX, LE *et al.*, 2018). No caso do agrupamento hierárquico, são consideradas 4 métricas de agregação de classes: i. média aritmética; ii. método de Ward (incremento à variância interna do cluster após fusão; considera a distância euclidiana); iii. vizinhos mais próximos; iv. vizinhos mais distantes. Para ambos os tipos de agrupamento, foram consideradas 3 métricas de distância: i. euclidiana (hipotenusa); ii. *cityblock* (soma dos catetos); iii. cosseno do ângulo entre os dois vetores que definem os padrões. Na metodologia de classificação, primeiro é aplicada a classificação hierárquica, com vistas a definir o melhor número de clusters. Em seguida, é aplicada a classificação não hierárquica. É importante ressaltar que os clusters formados podem ser adotados ou não para a regressão local, estando tal decisão condicionada à existência de um número mínimo de elementos no subconjunto, o qual é estabelecido subjetivamente. A saída final do modelo é formada a partir do resultado do modelo baseado em regressão linear múltipla (descrito no item anterior), aplicada sobre cada um dos clusters. Os modelos desenvolvidos com ênfase neste algoritmo de *clustering* identificam-se pela sigla “CL”;

A4 – Seleção de atributos aplicada à regressão: Conjunto de algoritmos aplicados para a seleção de modelos de *downscaling* estatístico (camada de modelos “M” e “IBL”; Figura 23) individuais e para fins de combinação. Para este conjunto de algoritmos base (“A4”), foram adotadas funções de transferência multilíneas com arquitetura de parâmetros global (i.e., ajustados para um único conjunto de treinamento). Tais algoritmos são baseados em procedimentos do tipo *wrapper* fundamentados no acoplamento entre o algoritmo de eliminação iterativa (descrito no âmbito do algoritmo base “A1”) e métodos de seleção iterativa (e.g., *forward search* e *stepwise selection*; HARPHAM; WILBY, Robert L., 2005; HESSAMI *et al.*, 2008; YANG, C.; WANG, Ninglian; WANG, Shijin, 2017) e regularização (e.g., *elastic net selection*; Li, 2019). Os modelos individuais associados a tais algoritmos identificam-se pelas seguintes siglas em suas respectivas nomenclaturas: 1) “ENR” – *elastic net regression*; “SWR” – *stepwise regression*; “SWM” – *stepwise* modificado: *forward selection + backward elimination search* (Seções 4.3.5 e 4.3.6). Os modelos de combinação linear associados a tais algoritmos são identificados pela sigla “CBL”, que inclui os modelos “CBLEN”, “CBLSW” e “CBLSWM”, os quais se apropriam dos mesmos algoritmos aplicados aos modelos individuais “ENR”, “SWR” e “SWM”, respectivamente. Particularmente para os modelos de combinação, avalia-se ainda modelos baseado em Máxima Relevância Mínima Redundância (mRmR), que tem por finalidade a avaliação de conjuntos preditores complementares a partir do cálculo da informação mútua (Seção 4.3.7).

A5 – Análise de complementaridade espectral: Modelos baseados em métodos de quantificação de complementaridade espectral. São aplicadas três modificações principais: i) introdução de pesos para a indução do processo de ajuste de coeficientes com base na representatividade espectral entre os sinais (modelos “SCW”); ii) seleção de variáveis com base em maximização de relevância e minimização de redundância espectral; iii) extração de atributo voltado para o agrupamento de padrões sinóticos com base em relação espectral entre as variáveis preditoras e o alvo. As duas últimas modificações se aplicam aos modelos “CL” (os quais recebem a letra “e” ao final de suas siglas) (Seção 4.4).

Ainda em relação às variantes de modelo, como resultados da primeira camada da metodologia proposta, diversas configurações de domínio foram avaliadas para o treinamento dos modelos de *downscaling* estatístico e de combinação. Como mencionado anteriormente, os domínios considerados são compostos, em essência,

por simulações numéricas providas por GCMs (neste relatório, o GFS). É importante ressaltar que a realização de *downscaling* estatístico a partir de simulações de modelos mesoescalares tende a permitir melhor compreensão sobre diversos fenômenos atmosféricos atuantes na região de interesse. Entretanto, a utilização de modelos de grande escala reduz significativamente o esforço computacional associado. Sobre a composição das malhas de previsões numéricas do tempo, foram consideradas a magnitude e as componentes zonal (U) e meridional (V) da velocidade do vento horizontal dispostas em um número pré-estabelecido de níveis verticais (níveis de pressão) mais próximos à superfície. Para previsões de irradiância global horizontal, aplicou-se o nível superficial de DSR (*Downward Shortwave Radiation*). Ao todo, quatro grupos principais de configuração de domínio foram considerados, os quais produziram diferentes variantes de modelos em associação com os algoritmos supracitados. Tais configurações são apresentadas em mais detalhes a seguir:

D1 – Domínio Fixo de Preditores: Domínio formado por número fixo de preditores. Para modelos de *downscaling* estatístico, a configuração em questão é formada por malhas horizontais de previsões numéricas do tempo de formato quadrangular em torno do local de interesse (i.e., centradas em torno das coordenadas geográficas da estação anemométrica avaliada). Essas malhas horizontais estão dispostas em níveis verticais (*pressure levels*). Nessa configuração de domínio, os modelos são treinados separadamente para cada nível vertical, havendo, portanto, diferentes subvariantes dos modelos associados. Os modelos de *downscaling* estatístico que adotam tal configuração de domínio são identificados por meio dos números 1, 2 e 3, relativos à aplicação de malhas fixas com 4 (2 x 2), 16 (4 x 4) e 36 (6 x 6) pontos, respectivamente, as quais estão contidas em malha ampla com 24 x 24 pontos. Em relação aos domínios dos modelos de combinação, aplicam-se as saídas dos diferentes modelos de *downscaling* estatístico, sendo tal conjunto também representado pelo indicador de domínio “D1”. Note-se que o nível vertical que apresentou o melhor resultado para cada um dos modelos individuais é selecionado no âmbito do teste de generalização (i.e., utilizando-se os dados contidos no conjunto de validação);

D2 – Componentes Principais: Domínio formado a partir da Análise de Componentes Principais aplicada sobre um conjunto extenso de variáveis regressoras. No caso dos modelos individuais de *downscaling*, esse conjunto extenso de variáveis compreende todos os níveis verticais relacionados a uma mesma malha

horizontal de 36 pontos (i.e., definida no âmbito da configuração “D1”). Já no caso dos modelos de combinação, são consideradas as saídas dos modelos individuais referentes a todos os níveis verticais estudados. O critério de redução de dimensionalidade consiste em desconsiderar as componentes de menor contribuição, de forma que o novo conjunto de variáveis regressoras esteja associado a um erro de reconstrução de, no máximo, 5%. Neste relatório, os modelos que empregam este algoritmo recebem o sufixo “p” à sua respectiva nomenclatura;

D3 – Seleção de Domínio: Domínio consideravelmente mais amplo que os descritos anteriormente, composto por malhas horizontais com 576 pontos em torno do ponto de interesse em formato quadricular dispostas em nove níveis verticais de pressão (ou seja, 24x24 pontos). A configuração de domínio adotada neste relatório tem o intuito de avaliar possíveis ganhos com a introdução de informação sobre o comportamento de variáveis atmosféricas em áreas mais distantes do local de interesse. Destaca-se, ainda, que para eliminar a redundância da informação contida no amplo domínio descrito, aplica-se algoritmo voltado para a seleção de pontos de malha no domínio baseado em acoplamento de métodos de seleção e eliminação iterativa (sendo o último tal como descrito anteriormente no âmbito dos algoritmos “A4” e “A1”, respectivamente);

D4 – Classificação de Domínio: Configuração de domínio voltada para modelos de ajuste local composta por subconjuntos formados a partir do agrupamento dos pontos de malha pertinentes à configuração “D3” (descrita anteriormente). De forma análoga àquela descrita para o algoritmo base “A3”, o agrupamento em questão é realizado sobre dois conjuntos principais, os quais são compostos por: a) previsões dos modelos em questão; b) atributos das séries no domínio da frequência (e.g., (GRINSTED; MOORE; JEVREJEVA, 2004; PERRUCI, 2018). Em relação à segunda, os modelos associados são identificados por meio da letra “e” na sua respectiva nomenclatura.

A Tabela 4 apresenta, de forma resumida, as características gerais das variantes de modelo desenvolvidas a partir dos algoritmos base e configurações de domínio supracitados.

Tabela 4 - Informações gerais sobre os modelos de *downscaling* estatístico e de combinação aplicados à previsão do comportamento do vento local.

Grupo (Algoritmos)	Configurações de Domínio	Algoritmos base	Modelos (nomenclatura)	Descrição/informação
--------------------	--------------------------	-----------------	------------------------	----------------------

MLR Seção 4.3.2	D1, D2, D3	A1	MLR1, MLR2, MLR3, MLRp, MLRs	Modelos de <i>downscaling</i> estatístico com ênfase em regressão linear múltipla e “poda” de coeficientes (<i>backward elimination search</i>)
AN Seção 4.3.4	D1, D2, D3	A1, A2	AN1, AN2, AN3, ANp, ANs	Modelos de <i>downscaling</i> estatístico baseados no método de análogos
ENR, SWR, SWM Seções 4.3.5 e 4.3.6	D3	A1, A4	ENRs, SWRs, SWMs	Modelos de <i>downscaling</i> estatístico baseados em algoritmos de seleção de variáveis (<i>feature selection models</i>)
SCW Seções 4.3.5, 4.4.3	D3	A1, A4, A5	SCWs-V1, SCWs- V2	Modelos de <i>downscaling</i> estatístico com ênfase em complementaridade espectral
CL Seção 4.3.8	D1, D2, D4	A1, A3, A5	CL1, CL2, CL3, CLp, CLpe	Modelos de <i>downscaling</i> estatístico com arquitetura de parâmetros locais (agrupamento de padrões semi supervisionado)
CBL Seções 4.3.5, 4.3.6 e 4.3.7	D1, D2	A1, A4	CBLp, CBLEN, CBLSWM, CBLSW,	Modelos de combinação com ênfase em regressão linear múltipla e complementaridade entre preditores
CBLSCW Seções 4.3.5 e 4.4.3	D1	A1, A4, A5	CBLSCW-V1, CBLSCW-V2	Modelos de combinação com ênfase em complementaridade espectral
CBCL Seções 4.3.5, 4.4.1 e 4.4.3	D1, D4	A1, A3	CBCL, CBCLe, CBCLp, CBCLpe	Modelos de combinação com arquitetura de parâmetros locais (agrupamento de padrões)

Fonte: autor.

4.6 Teste: Avaliação de resultados

O bloco de Testes da metodologia proposta na Figura 7 consiste na avaliação dos modelos em simulação de operação em tempo real, i.e., neste período, os dados observacionais apenas são adotados para a avaliação dos resultados, uma vez que, em se tratando de modelos com parâmetros fixos no tempo, toda decisão objetiva e subjetiva já estão estabelecidas a partir de amostras anteriores, uma vez que, no

contexto do teste da metodologia, as amostras mês de operação dizem respeito a estados futuros, sobre os quais ainda não se têm dados observacionais.

Considerando o número de resultados produzidos em função dos espaços de busca definidos no bloco de Configuração Geral (Seção 4.2) em cada local, mês de operação, horizonte de previsão e configurações de treinamento de modelos, a avaliação dos resultados se deu em função de três espaços amostrais distintos:

Espaços Amostrais 1 (E1) – conjuntos que reúnem os resultados do k -ésimo modelo para todos os casos incluídos no escopo da análise, e.g., agrupam-se resultados obtidos para todos os horizontes de previsão e demais graus de liberdade considerados, como indica a Eq. 43; em que: q – índice de qualidade das estimativas; K – número de modelos intercomparados (32, no total); f – índice de horizonte de previsão (16, no total); F – número total de horizontes avaliados; γ – casos avaliados (e.g., tipo de variável, configurações de hiperparâmetros, locais, meses de operação etc.); Γ – número de total de casos γ . Dessa forma, a análise das distribuições dos estatísticos nos espaços E1 permite a intercomparação entre os modelos em relação à robustez apresentada, sendo os melhores resultados aqueles que apresentam menores dispersões em torno da mediana associada a cada distribuição, que deve estar próxima ao valor ideal (i.e., zero para métricas de erro e um para correlação e razão entre os desvios padrão). Ainda sobre os espaços amostrais E1, calcula-se o aprimoramento dos estatísticos de todos os modelos em relação aos resultados obtidos pelo modelo de referência a partir da Eq. 44. O valor de aprimoramento calculado indica diferenças percentuais em relação a um determinado estatístico representado pelo índice de qualidade q , os quais são modificados para índices de habilidade cujos valores mínimos se relacionam aos melhores resultados (Tabela 5). Dessa forma, valores positivos de aprimoramento indicam melhoria em determinado estatístico (i.e., aumento da correlação ou redução do erro);

Espaços Amostrais 2 (E2) – consiste em distribuições parciais obtidas a partir dos espaços E1 ao serem discriminados os horizontes de previsão, e.g., para cada modelo, formam-se 16 conjuntos que agrupam as 2 configurações de hiperparâmetros associadas a cada horizonte de previsão (Eq. 45). Tendo em vista a tendência de perda de previsibilidade com o aumento do intervalo entre os horários de lançamento e da previsão (devido à natureza caótica da dinâmica atmosférica), a mediana calculada para os espaços E2 permite intercomparar os modelos em relação à

homogeneidade nos resultados obtidos para os diferentes horizontes de previsão. Os espaços E2 são avaliados por meio da mediana das amostras em cada conjunto;

Espaços Amostrais 3 (E3) – conjuntos que agrupam os resultados de todos os modelos para cada um dos 84 casos isoladamente (Eq. 46). Os conjuntos $E3(f, \gamma)$ são adotados para o cálculo do percentil $u\%$ ($p_{u\%}$), valor de referência adotado nas desigualdades apresentadas na Eq. 47. A matriz de pontuação “R” possui estrutura tridimensional (i.e., distribui-se em função de k, f e γ) contendo valores binários que quantificam os casos em que, frente aos demais modelos, resultado do k -ésimo modelo, para o horizonte f e configuração de hiperparâmetros γ , é o melhor ou está entre os melhores resultados. Em função do número de modelos considerados (18), o percentil 5% (para índices cujo ótimo é o mínimo) foi estabelecido para que, em conjunto $E3(f, \gamma)$, seja pontuado o modelo associado ao menor índice de qualidade obtido, bem como outros que por ocasião estejam empatados com o primeiro. Note, entretanto, que ao ser considerado um estatístico cujo valor ótimo deve ser maximizado (e.g., correlação), inverte-se o sinal do estatístico (alternativamente, altera-se a desigualdade na Eq. 47 para $q(k, f, \gamma) \geq p_{100-u\%}(E3(f, \gamma))$). Por fim, a análise realizada a partir dos E3 se baseia na avaliação da frequência (relativa), ou pontuação global dos modelos em função do número percentual de casos em que o resultado do modelo k apresentou os melhores, ao ser comparado com os resultados dos demais modelos para o mesmo horizonte de previsão (f) e configuração de hiperparâmetros (γ), tal como indica a Eq. 48.

$$E1(k) = \{ q(k | \forall f, \gamma) \}, \quad k \in [1, 2, \dots, K] \quad (43)$$

$$I_{REF}(q | q_{REF}) = 100\% \times \frac{(q_{REF} - q_m)}{|q_{REF}|} \quad (44)$$

$$E2(k, f) = \{ q(k, f | \forall \gamma) \}; \quad k \in [1, 2, \dots, K], f \in [1, 2, \dots, F] \quad (45)$$

$$E3(f, \gamma) = \{ q(f, \gamma | \forall k) \}; \quad f \in [1, 2, \dots, F]; \gamma \in [1, 2, \dots, \Gamma] \quad (46)$$

$$\begin{cases} R(k, f, \gamma) = 1, & \text{se } q(k, f, \gamma) \leq p_{u\%}(E3(f, \gamma)); \\ R(k, f, \gamma) = 0, & \text{caso contrário;} \end{cases} \quad (47)$$

$$FR(k) = 100\% \times \sum_f \sum_\gamma \frac{R(k, f, \gamma)}{F \cdot \Gamma} \quad (48)$$

Tabela 5 – Estatísticos e índices de qualidade correspondentes adotados para a avaliação de resultados.

Estatístico	Equação	Índice de qualidade (q)
Correlação (COR)	$\rho = \frac{cov(x, y)}{\sigma_x \sigma_y}$	inverte-se o sinal do estatístico ou a desigualdade na Eq. 47.

Razão entre os desvios padrão (STDr)	σ_x/σ_y	$ \sigma_x/\sigma_y - 1 $
Raiz do erro médio quadrático das anomalias (RMSd)	$\sqrt{\frac{1}{N} \sum_{n=1}^N [(x_n - \bar{x}) - (y_n - \bar{y})]^2}$	Sem alteração
Índice de habilidade de Taylor ("SS4"; Eq. 3)	$\frac{(1 + \rho)^4}{4(\hat{\sigma} + 1/\hat{\sigma})^2}$	inverte-se o sinal do estatístico ou a desigualdade na Eq. 47.
bias	$\bar{x} - \bar{y}$	$ \bar{x} - \bar{y} $
Raiz do erro médio quadrático (RMSe)	$\sqrt{\frac{1}{N} \sum_{n=1}^N (x_n - y_n)^2}$	Sem alteração

Símbolos: x – saídas do modelo; y – observações; σ – desvio padrão; cov – covariância; n –ésimo elemento de conjunto contendo N amostras; \bar{x}, \bar{y} – médias dos sinais modelados e observados, respectivamente.

5 RESULTADOS

Os resultados apresentados nesta subseção dizem respeito aos estudos de casos considerados no bloco de testes da metodologia apresentada, que consiste em simulação da previsão operacional variáveis pertinentes aos setores de geração de energia eólica e solar em meses de alta e baixa produção de energia. Ao todo, foram considerados dados de 6 estações anemométricas, 1 estação solarimétrica, 1 aerogerador e 1 sistema fotovoltaico (*stringbox*). Informações gerais sobre as bases de dados observacionais adotadas estão disponíveis na Tabela 6. A coluna “ID” diz respeito ao índice dos locais estudados, cujas localizações também se encontram indicadas na Figura 24. Há um total de 4 estações anemométricas localizadas na costa e 2 duas localizadas no interior da Região NE. A estação solarimétrica se encontra localizada no interior do estado de Pernambuco. As centrais de geração eólica e solar estudadas estão localizadas no interior dos estados de Pernambuco e Bahia, na mesma localização de estações eólica e solar também contempladas. O período estudado, em todas as centrais se estende desde 01/08/2019 até 31/08/2020, com simulações de operação em tempo real nos meses de Março, Junho, Agosto, Setembro e Outubro, a depender de cada local (como indicado na Tabela 6).

Tabela 6 – Informações gerais sobre os dados observacionais adotados nos estudos de caso.

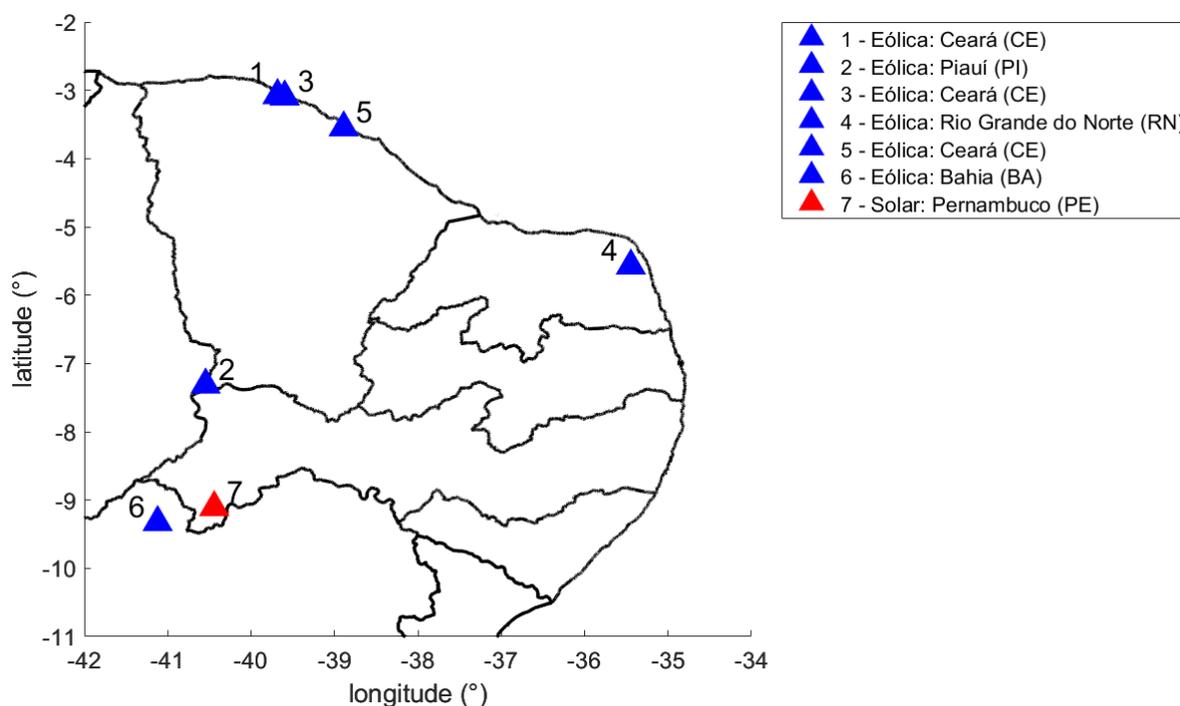
ID (#)	Tipo de dado	Costa / Interior	Informações do sensor/sistema	Testes: operação (2020)	Intervalo de integração
1	Torre anemométrica	Costa	Nível (anemômetro): 87 m	Set/Out	30 minutos
2	Torre anemométrica	Interior	Nível (anemômetro): 87 m	Set/Out	30 minutos
3	Torre anemométrica	Costa	Nível (anemômetro): 99,7 m	Set/Out	30 minutos
4	Torre anemométrica	Costa	Nível (anemômetro): 87 m	Set/Out	30 minutos
5	Torre anemométrica	Costa	Nível (anemômetro): 99,5 m	Set/Out	30 minutos
6	Torre anemométrica	Interior	Nível (anemômetro): 98,4 m	Mar/Ago/Out	10 minutos
7	Estação Solarimétrica	Interior	solarímetro	Mar/Jun/Ago/Out	1 minuto

					(integrado em intervalos de 3 ou 6 horas)
6	SCADA (Aerogerador)	Interior	Potência nominal: 2350 kW	Mar/Ago/Out	10 minutos
7	SCADA (Stringbox)	Interior	Potência nominal: 50,16 kW	Mar/Jun/Ago/Out	1 minuto (integrado em intervalos de 3 ou 6 horas)

Fonte: autor.

Todas as campanhas de dados observacionais foram submetidas a testes globais e locais garantia de qualidade de dados observacionais. Tais testes tem como objetivo a qualificação das séries temporais empregadas por meio de rotulação de amostras que atendem a critérios aceitáveis do ponto de vista físico e estatístico, bem como em relação à ausência ou problemas diversos associados com o processo de aquisição de dados em geral, como a repetição de (MORAES, 2015; PEDROSA, 2016; PETRIBÚ; SOUZA, 2018).

Figura 24 – Locais das estações anemométricas/solarimétricas e centrais eólicas/solares analisadas.

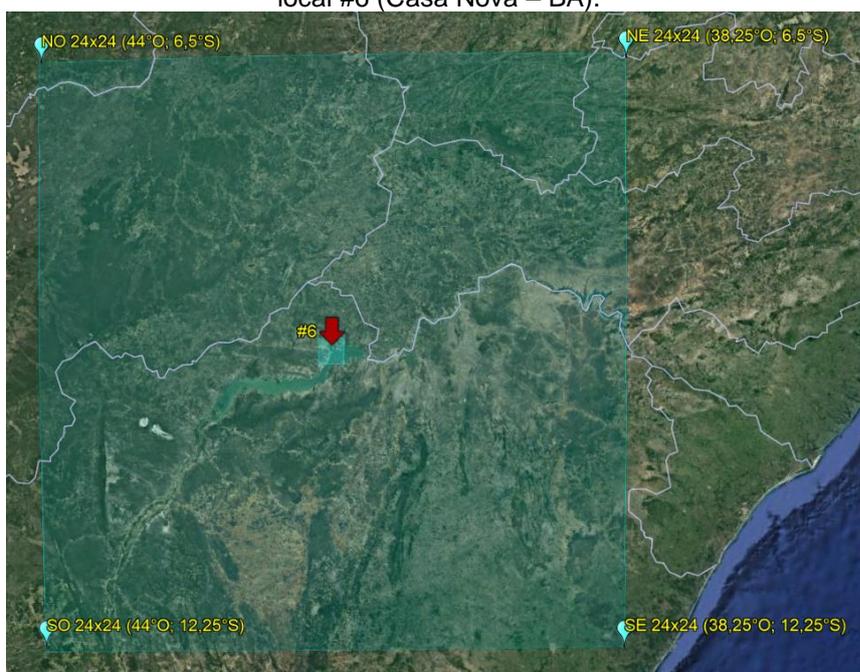


Fonte: autor.

Como entrada para os modelos de *downscaling* estatístico, adotaram-se saídas do GFS 0.25, com resolução horizontal de 0,25 grau de latitude/longitude (aproximadamente 27 km entre pontos adjacentes). O domínio total empregado para os estudos de caso apresentados é composto por malhas quadriculares de 24 x 24

pontos em torno do ponto de interesse, tal como ilustra a Figura 25. Sobre os horizontes de previsão, foram consideradas as previsões de +3 a +48 horas com resolução de 3 horas, totalizando 16 horizontes por ciclo. Note-se que, em relação aos dados anemométricos e de potência eólica, as séries estão disponíveis em intervalos de 10 ou 30 minutos na escala temporal referida, neste trabalho, como o “curto prazo”. Portanto, foram selecionados para esse estudo apenas os *timesteps* da série observacional que coincide com os do horário alvo do modelo numérico. Devido à natureza operacional do GFS, as atualizações tornam as parametrizações do GFS heterogêneas em períodos suficientemente longos. Portanto, em virtude de uma importante atualização realizada no ano de 2019, os períodos considerados neste relatório têm início em 01 de agosto do mesmo ano.

Figura 25 – Exemplo de área contemplada por domínio contendo 576 pontos (24x24) centrado no local #6 (Casa Nova – BA).



A área menor no interior do domínio em destaque exemplifica domínio contendo 4 (2x2) pontos em torno do local de interesse. Fonte: Google Earth Pro.

Sobre os níveis verticais e lançamento de previsão do modelo de larga escala, adotou-se o nível de pressão de 1000 mb do GFS (o mais próximo à superfície) e os 4 ciclos diários do modelo (i.e., os lançamentos correspondentes a 00 06 12 18 UTC). No entanto, para modelos baseados em PCA ou seleção de domínio, os 6 níveis de pressão mais próximos da superfície foram adotados. Destaca-se, no entanto, que – em relação aos níveis verticais – a escolha se deve ao escopo deste trabalho, cuja ênfase se dá em relação à capacidade de reproduzir informações locais a partir de

associação entre a larga e a microescala. Contudo, como evolução natural da metodologia proposta, faz-se possível também a avaliação de sensibilidade dos modelos em relação a diferentes do comportamento atmosférico no âmbito de decisões tomadas no bloco de Configurações Gerais (Seção 4.2).

No tocante às previsões de Irradiância Global Horizontal (GHI) realizadas a partir de saídas do GFS, foram adotadas as saídas de (“*Downward Shortwave Radiation*” - DSR), que diferem das previsões das componentes U e V da velocidade do vento ao descreverem o valor médio das 3 horas anteriores ao horário da previsão nos horizontes de previsão “ímpares” (i.e., +3, +9, +15 etc.) ou das últimas 6 horas no caso horizontes de previsão “pares” (+6, +12, +18 etc.). Apesar de tal distinção, ressalta-se que os GCMs apenas se propõem a descrever a dinâmica de larga escala do comportamento atmosférico. Nesse sentido, para as previsões de irradiância, deve-se atentar à comparação com valores observacionais “equivalentes” em cada caso (i.e., para o treinamento e validação dos modelos, calcula-se a média em períodos de 3 ou 6 horas, de acordo com o horizonte de previsão em questão). Destaca-se, ainda, que os horários do dia contemplados pelas previsões em cada horizonte de previsão ficam determinados em função da posição geográfica do local estudado. Por exemplo, para o local #7 (localizado no semiárido brasileiro; GMT-3), em horário local, os horários das previsões em horizontes “pares” dizem respeito a 03:00, 09:00, 15:00 e 21:00, enquanto os horários das previsões nos horizontes “ímpares” estão associados a 00:00, 06:00, 12:00, 18:00.

A metodologia proposta permite avaliar diversos graus de liberdade em relação à parametrização dos modelos, o que produz um número pré-determinado de parametrizações. Como mostra a Tabela 7, tal número se dá em função dos diferentes graus de liberdade considerados no espaço de busca por hiperparâmetros, bem como os diferentes horizontes de previsão considerados em cada caso (ou “alvo”). No âmbito dos resultados discutidos nesta seção (derivados diretamente das previsões de DSR providas pelo GFS), os modelos foram aplicados à previsão de irradiância global horizontal em curto prazo (até +48 horas à frente; médias de 3 ou 6 horas). Adicionalmente, para os resultados avaliados nesta tese, os graus de liberdade incluem duas opções para conjunto de treinamento dos modelos. Em relação aos índices de qualidade empregados para fins de otimização de modelos, para modelos de *downscaling* estatístico, adota-se a correlação de Pearson, enquanto que,

para os modelos de combinação, aplica-se o RMSe como critério tal como mostra a Tabela 7.

Tabela 7 – Casos e variantes considerados para cada local.

Variantes	Mudança	Opções	# de Variantes
Horizontes de previsão	Alvo	$f = +3, +6, +9, +12, \dots, +48$ horas à frente Variáveis: Velocidade do vento (m/s) Irradiância Global Horizontal (W/m^2)* Potência gerada (kW) *(horizontes “ímpares”: médias das últimas 3 horas; horizontes “pares”: médias das últimas 6 horas)	16
Treinamento dos modelos	Hiperparâmetros	Casos: Para os modelos individuais: 1ª parte aplicada à calibração dos modelos; 2ª parte aplicada à validação; 2ª parte aplicada à calibração dos modelos; 1ª parte aplicada à validação. Índices de qualidade (otimização): Modelos de <i>downscaling</i> estatístico: Correlação (Pearson) Modelos de combinação: RMSe	2

Fonte: autor.

Considerado os locais mostrados na Tabela 6, bem como os casos descritos na Tabela 7, os resultados dos modelos são organizados em cinco seções principais: Previsões para velocidade do vento (5.1); Previsões para GHI (5.2); Previsões para geração de potência eólica (5.3); Previsões para geração de potência solar fotovoltaica (5.4); Resultados gerais (5.5). Nas primeiras quatro seções, serão apresentados resultados com respeito ao índice de habilidade de Taylor ($SS4$; Eq. 3), que visa demonstrar a capacidade geral dos modelos em relação à estrutura de fase e frequência, bem como à amplitude de variações reproduzidas. Em outras palavras, quanto maior o $SS4$, melhor tendem a ser a correlação e a variância reproduzidas pelo modelo de forma conjunta. Além do $SS4$, também são apresentados resultados em

relação a outros estatísticos (e.g., STDr, bias e RMSe), com vistas à verificação de diferentes aspectos do aprimoramento conferido pela metodologia proposta. Na Seção 5.5, os resultados são avaliados em conjunto considerando todos os testes realizados (incluindo-se todas as previsões avaliadas nas seções anteriores). Além disso, são apresentadas discussões em relação às principais contribuições dos modelos propostos no âmbito da metodologia adotada.

5.1 Previsões para velocidade do vento

Nesta subseção são avaliados os resultados da aplicação da metodologia proposta considerando a previsão da velocidade do vento a partir de saídas do GFS para os horizontes +3 até +48 horas à frente, em passos de 3 horas (+3, +6, +9 etc.). Para tal, foram utilizados dados anemométricos aferidos nos locais de índices #1 ao #6, considerando (individualmente) os meses de operação indicados na Tabela 6. Para cada local/mês de operação, avaliam-se resultados de testes realizados considerando os diferentes horizontes de previsão e configurações de hiperparâmetros, como indicado na Tabela 7. Ao todo, foram realizados 416 testes para previsão de velocidade do vento.

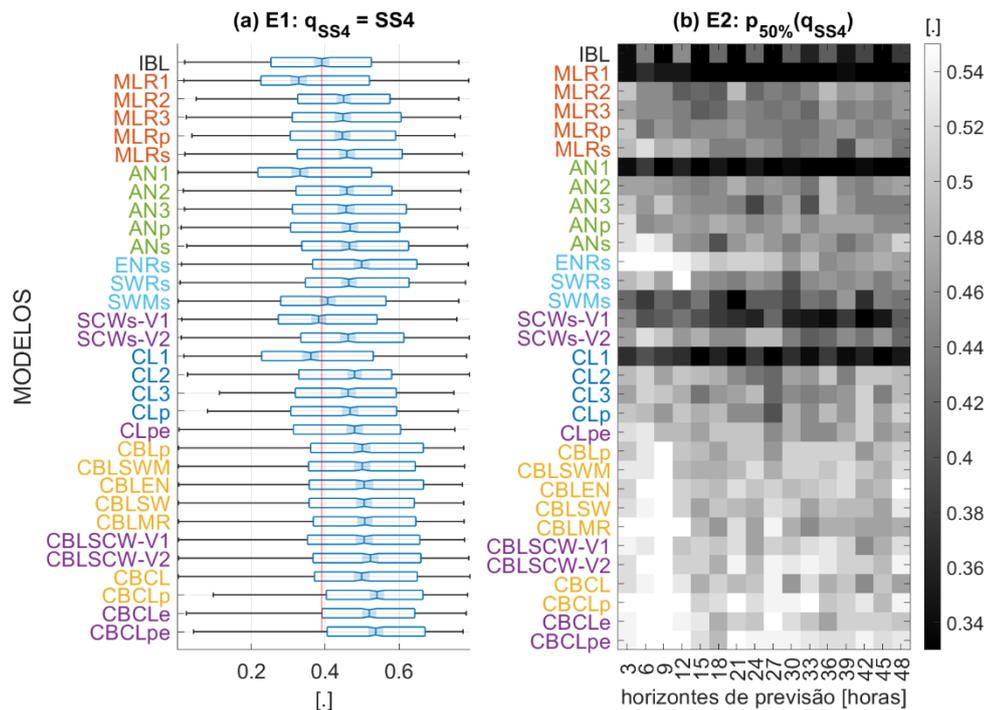
Considerando o número de resultados produzidos em função dos casos definidos na Tabela 7, os quais se aplicam a cada local e para cada mês de operação indicados na Tabela 6, a avaliação dos resultados se deu em função de três agrupamentos de resultados distintos (Seção 4.6): E1, agrupamento dos resultados para todos os horizontes de previsão e configurações de hiperparâmetros, sobre os quais também se avalia o aprimoramento geral obtido pelos modelos estatísticos em relação ao modelo de referência IBL (Eq. 44); E2, divisão dos conjuntos anteriores em função de cada horizonte de previsão, sendo de cada subconjunto avaliada a mediana); E3, agrupamento dos resultados de todos os testes, i.e., resultados de todos os modelos em cada horizonte/conjunto de hiperparâmetros (análise particular a cada caso). Em relação aos conjuntos E3, para cada subconjunto, pontua-se o modelo que apresentar o melhor resultado (pontuando também modelos em empate técnico). Dessa forma, partindo dos conjuntos E3, avalia-se a pontuação global de cada modelo (percentual de vezes em que o modelo apresentou os melhores resultados em um caso individual). Destaca-se, ainda, em relação aos conjuntos E1, os resultados são mostrados em gráficos “*boxplot*”, cuja barra central indica o intervalo

interquartil ($IQR = Q_3 - Q_1 = p_{75\%} - p_{25\%}$). O intervalo de confiança de 95% da mediana ($p_{50\%}$) é indicado por meio de constrictões e intervalos escurecidos sobre barra de IQR , enquanto que as linhas delimitam distância de $1,5 \cdot IQR$ em relação aos extremos da barra central. Valores de distâncias acima ou abaixo são considerados *outliers* (não mostrados no gráfico).

A Figura 26 apresenta os índices de habilidade $SS4$ das previsões considerando os modelos “individuais” (*downscaling* estatístico) e de combinação. As cores indicadas na abcissa separam as diferentes variantes de modelos em grupos de acordo com a proximidade entre os algoritmos adotados (Seção 4.3). Em roxo, representa-se o grupo de modelos de *downscaling* estatístico (SCWs-V1; SCWs-V2; CLpe) e de combinação (CBLSCW-V1; CBLSCW-V2; CBCLe; CBCLpe) baseados em complementaridade espectral. De maneira geral, há uma tendência de aumento na complexidade dos modelos na medida em que se distancia da posição do modelo IBL (no topo). Os modelos baseados em complementaridade espectral, portanto, se situam abaixo do grupo de modelos mais similares em termos de complexidade.

Os gráficos mostrados na Figura 26a apresentam as distribuições gerais dos índices de habilidade $SS4$ para cada modelo (416 testes), enquanto que, na Figura 26b, são mostradas as medianas para modelo em cada horizonte de previsão (26 testes por horizonte de previsão). Em geral, é possível observar valores entre 0,3 e 0,7, com melhorias perceptíveis dos modelos de combinação sobre os modelos individuais (de *downscaling* estatístico). As melhorias proporcionadas pelos modelos de combinação, por um lado, se refletem na tendência de aumento na mediana da distribuição global, por outro lado, na homogeneização dos resultados em relação aos diferentes horizontes de previsão, mantendo maior estabilidade na acurácia das previsões. Observa-se, no entanto, que alguns modelos com menor número domínio apresentam deficiência na reprodução de características gerais do sinal (e.g., MLR1, AN1, SCWs-V1 e CL1). Porém, tal instabilidade é devidamente corrigida pelos modelos de combinação, os quais consolidam os melhores resultados conferindo maior acurácia e robustez às previsões finais.

Figura 26 – Previsões de curto prazo de velocidade do vento: Índice de habilidade “SS4”.



(a) Espaços E1: distribuições para os 416 testes; (b) Espaços E2: 26 testes por horizonte de previsão.
Fonte: autor.

Os resultados apresentados na Figura 27 dizem respeito ao aprimoramento proporcionado pelos modelos frente ao modelos de referência IBL (aplicados à previsão de velocidade do vento). Na Figura 27a, são apresentadas as distribuições dos estatísticos, para cada modelo, considerando todos os 416 testes realizados. Observa-se, na maior parte dos casos, medianas positivas aprimoramentos em relação ao modelo IBL. Entretanto, em relação aos modelos individuais (de *downscaling* estatístico), observam-se distribuições com o primeiro quartil negativo, o que indica menor probabilidade de aprimoramento. Tal instabilidade por parte dos modelos individuais possivelmente se deve ao fato de que, diferentemente dos modelos de combinação, há disponibilidade de informações sobre o comportamento atmosférico em locais próximos ou distantes do local de interesse, o que aumenta o risco em relação à quebra de pressuposto de estacionariedade nas relações estabelecidas através de métodos regressivos mais simples (e.g., regressão linear múltipla). Entretanto, tais resultados compõem conjunto de entradas (preditores) heterogêneos que se aplicam aos modelos de combinação, de forma que tais resultados – apesar de apresentarem menor relevância – são úteis para a obtenção de resultados mais acurados.

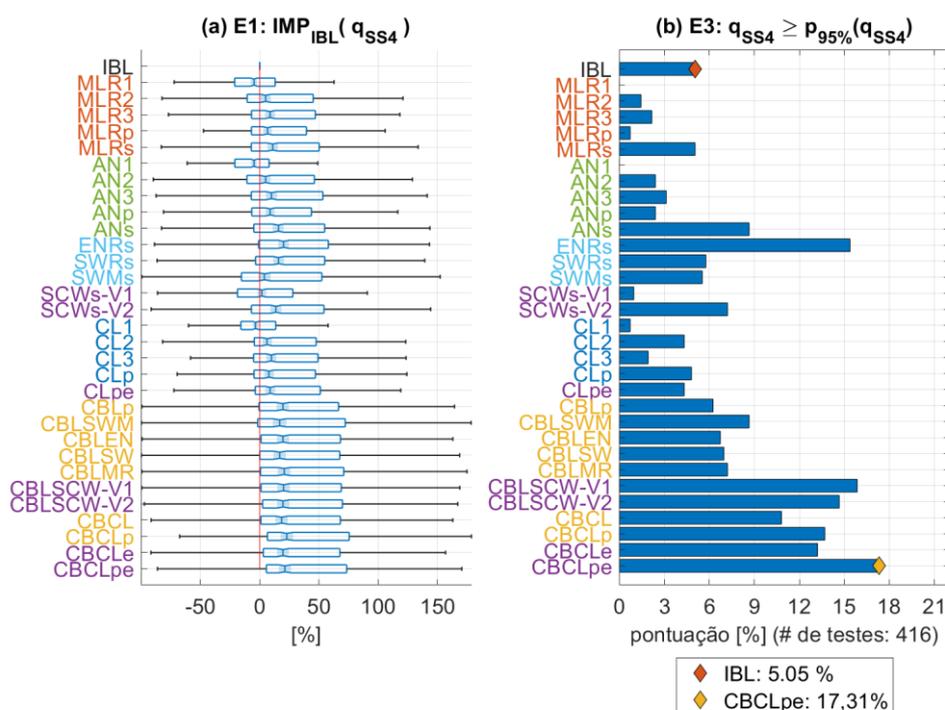
Na Figura 27b são apresentadas as pontuações gerais dos diferentes modelos ao serem intercomparados caso a caso (i.e., número percentual de vezes em um ou mais modelos apresentaram os melhores resultados em um determinado caso). De maneira geral, os modelos que se aplicam ao domínio pré-selecionado (com a letra “s” ao final de suas respectivas siglas) apresentaram melhores resultados. Dentre todos os modelos individuais, observam-se maiores aprimoramentos para o modelo ENRs (baseado em regularização tipo LASSO), com os melhores resultados em 15% dos testes, conferindo aprimoramentos de, em média, 20% sobre a interpolação bilinear. A aptidão de modelos baseados em técnicas de regularização do tipo LASSO aplicadas a problemas similares ao de *downscaling* estatístico é amplamente documentada na literatura (i.e., conjunto de preditores com alta colinearidade), tendo sido reportados resultados com acurácia e robustez comparáveis àqueles produzidos por modelos de estado-da-arte (HE *et al.*, 2019; RACHMAWATI; SUNGKAWA; RAHAYU, 2019; RANSTAM; COOK, 2018). Dentre os modelos individuais (sem modificações baseadas em complementaridade espectral), destaca-se também o modelo ANs (baseado em aplicação de método de “análogos” aplicado ao conjunto pré-selecionado), com pontuação global de 9%, com aprimoramento médio sobre o modelo IBL em torno de 15%.

Ainda sobre os modelos de *downscaling* estatístico, a Figura 27b mostra que o modelo SCWs-V2 (baseado em pesos espectrais) apresenta aprimoramentos similares ao dos modelos ENRs, SWRs e SWMs (grupo azul claro), produzindo os melhores resultados em aproximadamente 7,2% dos 416 testes analisados nesta seção, com aprimoramento médio em torno de 14%. O modelo CLpe, por outro lado, apresentou resultados similares aos do modelo CLp, o que sugere “equivalência” na performance entre os métodos baseados em agrupamento de padrões sinópticos e de atributos baseados em complementaridade espectral (no caso de previsões de velocidade do vento).

Sobre os modelos de combinação que não empregam complementaridade espectral (modelos “CB” cujas siglas estão indicadas pela cor laranja na abcissa dos gráficos mostrados na Figura 27b). Os modelos baseados em análise de componentes principais (CBLp) e seleção de atributos (CBLsVM, CBLEN, CBLSW e CBLMR) apresentaram performance similar, com pontuações para os melhores resultados entre (aproximadamente) 6,7% e 8,6% do total 416 dos testes, produzindo aprimoramento, em média, entre 16% e 20% sobre o modelo IBL. Ainda sobre o

mesmo grupo de modelos, constata-se melhor performance por parte dos modelos de combinação baseados em agrupamento de padrões “CBCL” (baseados em aplicação de *clustering* às saídas de modelos de *downscaling* estatístico). Os modelos CBCL e CBCLp produziram os melhores resultados em aproximadamente 10,8% e 13,7% dos testes realizados, sobre os quais se obtém distribuição de aprimoramentos (sobre o modelo IBL) com medianas de 18,5% e 22,9%, respectivamente.

Figura 27 – Previsões de curto prazo de velocidade do vento: aprimoramento sobre índice de habilidade “SS4”.



(a) Espaços E1: distribuições de valores de aprimoramento percentual sobre o modelo de referência (IBL) para os 416 testes realizados; (b) Espaços E3: pontuam-se os melhores dentre os 32 modelos em cada caso. As barras indicam a pontuação total em função do número de testes realizados. Fonte: autor.

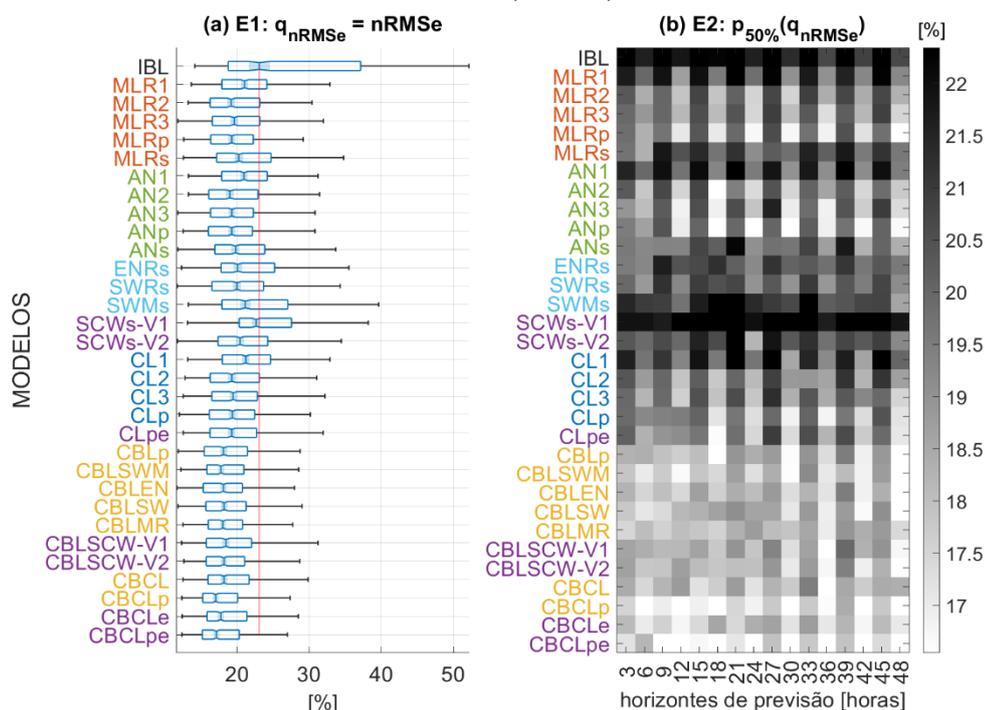
Seguindo divisão de modelos de combinação estabelecida no parágrafo anterior, sobre os modelos de combinação baseados em complementaridade espectral, separam-se duas classes principais: i) CBLSCW-V1 e CBLSCW-V2: modelos com arquitetura de parâmetros global; ii) CBCLe e CBCLpe: modelos com arquitetura de parâmetros local., a Figura 27b mostra que, em geral, os modelos produziram aprimoramentos significativos, com medianas de aproximadamente 20%. Os modelos CBLSCW-V1 e CBLSCW-V2 apresentaram os melhores resultados em 16% e 14,7% dos testes realizados, respectivamente. Tais valores representam aumentos expressivos em comparação com aqueles obtidos a partir de modelos semelhantes (e.g., CBLsVM, CBLen, CBLsW e CBLMR). De maneira similar,

observa-se que os modelos CBCLe e CBCLpe apresentam pontuações de 13,2% e 17,3%, respectivamente, sendo tais valores superiores àquelas apresentadas pelos modelos CBCL e CBCLp (para previsões para velocidade do vento). Dentre todos os modelos, o CBCLpe foi aquele que apresentou a mais alta pontuação (i.e., com melhores resultados para o maior número de testes analisados nesta seção), com mediana de distribuição de aprimoramentos em torno de 22%.

A Figura 28 apresenta os resultados das previsões de velocidade do vento em curto prazo em termos de valores de RMSe. Note-se, no entanto, para que diferentes tipos de variáveis possam ser comparados (nas seções seguintes), os valores de RMSe apresentados são normalizados pelo valor médio dos sinais observados em cada teste realizado (nRMSe). A interpretação dos resultados sob a perspectiva do nRMSe é similar àquela apresentada para o SS4, na Figura 26. Destaca-se, no entanto, que, em geral, os modelos individuais e de combinação reproduzem o sinal observado com erros abaixo de 20% do valor médio. No que concerne previsões de velocidade do vento, os modelos CBCLp e CBCLpe apresentaram os valores de nRMSe mais baixos para previsões de velocidade do vento, com medianas em torno de 17%.

Tal qual o SS4, os valores de RMSe são indicativos da qualidade geral das previsões. No entanto, tal estatístico difere em relação à sensibilidade a erros sistemáticos, como no caso da amplitude de variações (razão entre desvios padrão, “STDr”) e valor médio reproduzido (“bias”) (Tabela 5). Nesse sentido, destaca-se – ainda na Figura 26 – que tal diferença de sensibilidade evidencia padrões distintos na acurácia dos modelos entre os horizontes de previsão, com valores de nRMSe abaixo de 20% em horizontes pares (+06, +12, +18 etc.), e acima desse mesmo valor em horizontes ímpares (+03,+09,+15 etc.). Tal padrão é constatado para as diversas localidades estudadas, o que sugere que essa limitação resulta da própria capacidade preditiva do modelo GFS em relação ao comportamento observado (em escala local). Tal heterogeneidade impacta na forma e dispersão das distribuições dos estatísticos ao serem agrupados resultados para diferentes horizontes de previsão.

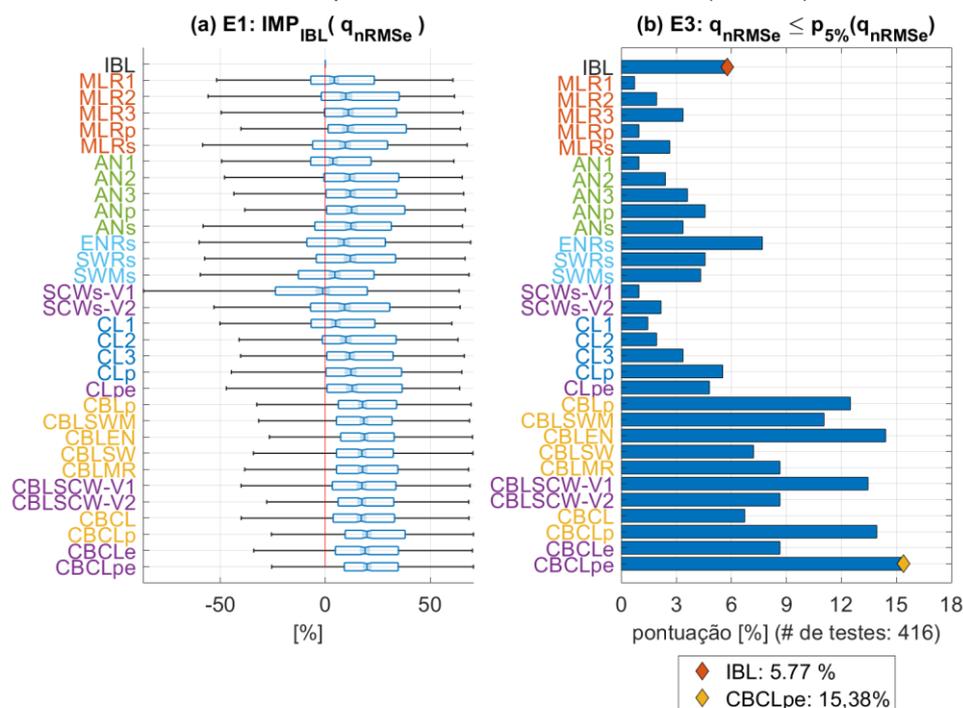
Figura 28 – Previsões de curto prazo de velocidade do vento: RMSe normalizado pela média do sinal observado (nRMSe).



(a) Espaços E1: distribuições para os 416 testes; (b) Espaços E2: 26 testes por horizonte de previsão. Fonte: autor.

Sobre os aprimoramentos percentuais em relação aos valores de nRMSe (Figura 29), observa-se um perfil similar àquele apresentado na Figura 27 (aprimoramentos sobre o SS4 do modelo IBL). Dentre os modelos avaliados, destacam-se o CBLp, CBLLEN, CBLSCW-V1, CBCLp e CBCLpe, os quais apresentaram os melhores resultados em mais de 12% dos testes, com aprimoramentos cujas medianas variam entre 17% e 20%. O maior aprimoramento global é obtido por meio do modelo CBCLpe, que produziu os melhores resultados em 15,4% dos testes. Destaca-se, também o aprimoramento obtido com o modelo CBLSCW-V1, de 18%. Tais resultados demonstram que, ao serem contemplados aspectos relacionados à estrutura de fase e frequência, amplitude de variações e valor médio do sinal observado, aprimoramentos significativos foram obtidos por meio de inferência sobre a complementaridade espectral entre preditores e sinais observados de velocidade do vento.

Figura 29 – Previsões de curto prazo de velocidade do vento: aprimoramento sobre RMSe normalizado pela média do sinal observado ($nRMSe$).



(a) Espaços E1: distribuições de valores de aprimoramento percentual sobre o modelo de referência (IBL) para os 416 testes realizados; (b) Espaços E3: pontuam-se os melhores dentre os 32 modelos em cada caso. As barras indicam a pontuação total em função do número de testes realizados. Fonte: autor.

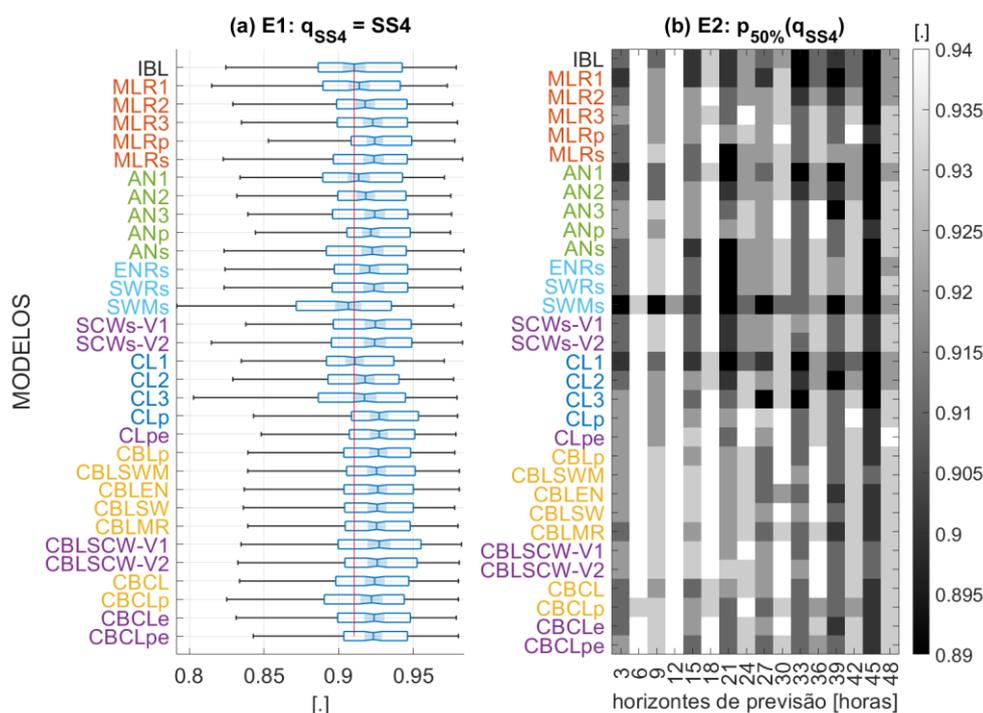
5.2 Previsões para Irradiância Global Horizontal (GHI)

Nesta subseção são avaliados resultados para previsão de GHI considerando os horizontes de +3 a +48 horas à frente, em passos de três horas, bem como duas configurações de hiperparâmetros (Tabela 7). O local avaliado é indicado pelo índice #7, considerando os meses de operação indicados na Tabela 6 (março, junho, agosto e outubro de 2020). Ao todo, 128 testes foram realizados no âmbito desta subseção.

Tal como na Seção 5.1, a Figura 30 apresenta os valores de SS4 relativos às previsões de curto prazo de GHI para todos os testes realizados (a), bem como a distribuição dos estatísticos em função dos horizontes de previsão (b). Em geral, os modelos produziram resultados similares, com a maior parte dos valores de SS4 entre 0,9 e 0,95. É possível observar, no entanto, tendências de aprimoramentos maiores por parte dos modelos de *downscaling* estatístico “3” e “p” (com domínios fixos de 36 pontos em torno da estação solarimétrica e domínios compostos por componentes principais, respectivamente; Seção 4.5): MLR3, MLRp, AN3, ANp, CLp e CLpe.

Os modelos de combinação, por outro lado, consolidam os melhores resultados obtidos no nível anterior (*downscaling* estatístico), com a maior parte dos aprimoramentos em até 2%, podendo chegar a chegar a 4% em situações específicas. Em relação aos horizontes de previsão, observa-se que a acurácia das previsões é maior no caso de horizontes de previsão “pares” (+6, +12, +18 etc.), cujos valores de SS4 se distribuem com mediana de aproximadamente 0,94. Nos horizontes de previsão ímpares (+3, +9, +15 etc.), observa-se redução nesses valores, que se concentram em torno de 0,9. Avalia-se, ainda, degradação na acurácia das previsões obtidas por modelos individuais (sem o indicador “CB”) a partir do segundo dia de previsão, o que é atenuado por modelos de combinação (com o indicador “CB”).

Figura 30 – Previsões de curto prazo de Irradiância Global Horizontal (GHI): Índice de habilidade “SS4”.

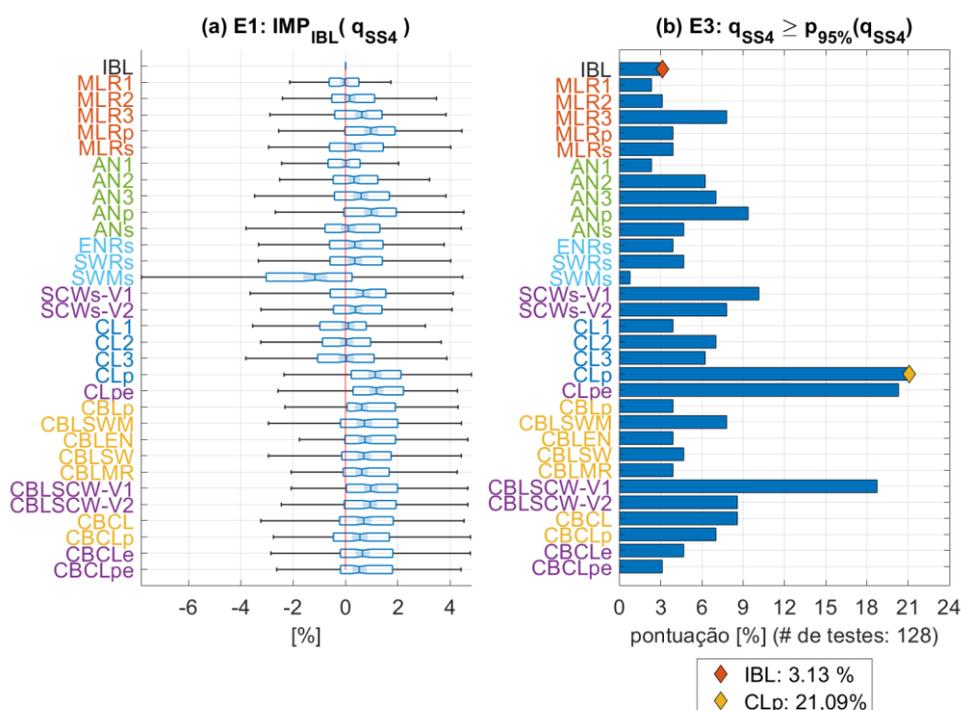


(a) Espaços E1: distribuições para os 128 testes; (b) Espaços E2: 8 testes por horizonte de previsão. Fonte: autor.

Em comparação com os modelos similares, os modelos baseados em complementaridade espectral apresentam boa performance, com aprimoramentos significativos, como se evidencia a partir da Figura 31. Em (a), apresentam-se resultados em termos de aprimoramento (aumento percentual) de valores de SS4 em comparação com o modelo IBL. Como mencionado no parágrafo anterior, os modelos individuais que se beneficiam de informações de malhas mais amplas (modelos “3”, “p” e “s”) produzem aprimoramentos significativos em relação à interpolação bilinear.

Ao serem intercomparados em cada caso, destacam-se os modelos ANp, SCWs-V1, CLp, CLpe, CBLSCW-V1, os quais apresentam os melhores resultados em mais de 9% dos testes realizados. Os modelos que apresentaram as mais altas pontuações foram o CLp (21,1%), CLpe (20,3%) e CBLSCW-V1 (18,8%), com aprimoramentos que se distribuem em torno de mediana de 1% (com intervalo de confiança a 95% positivo).

Figura 31 – Previsões de curto prazo de Irradiância Global Horizontal (GHI): aprimoramento sobre índice de habilidade “SS4”.

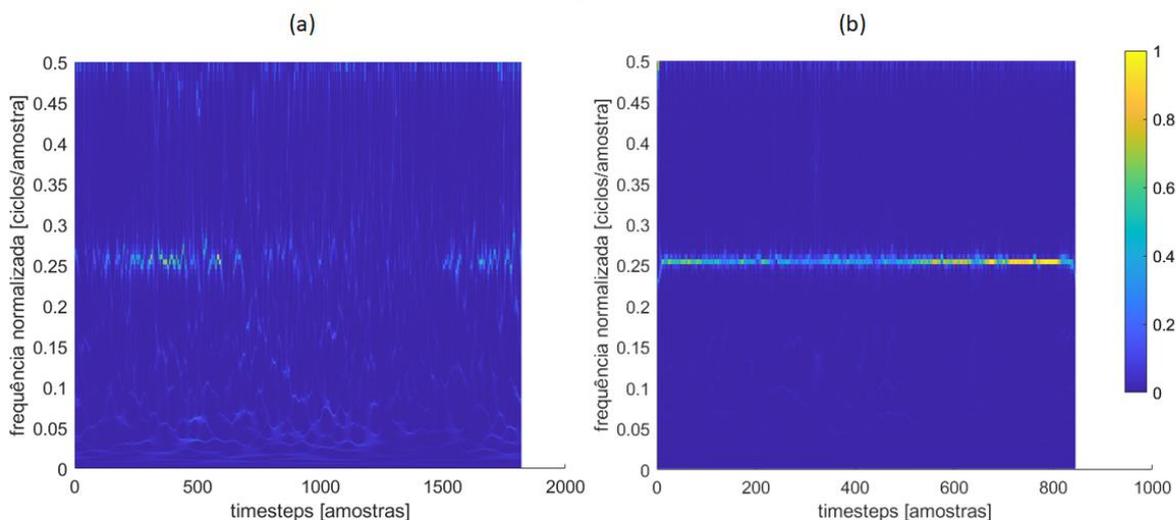


(a) Espaços E1: distribuições de valores de aprimoramento percentual sobre o modelo de referência (IBL) para os 128 testes realizados; (b) Espaços E3: pontuam-se os melhores dentre os 32 modelos em cada caso. As barras indicam a pontuação total em função do número de testes realizados. Fonte: autor.

Diferentemente do caso de sinais de velocidade do vento (apresentados na Seção anterior), em relação às previsões de irradiância solar, a maior parte da variância do sinal se encontra concentrada em harmônicos cujos períodos são de ordem diária ($f \cong \frac{1 \text{ ciclo}}{4 \text{ amostras}} = 0,25$), como mostram os espectrogramas normalizados na Figura 32. Por essa razão, em comparação com séries temporais de velocidade do vento, as séries associadas a GHI (integradas em intervalos de 3 ou 6 horas) aparentam ter estrutura de fase e frequências “simplificada”, da qual resultam maiores valores para estatísticos como o $SS4$ (que é proporcional à correlação). Nesse caso, é natural que os aprimoramentos alcançados por parte dos modelos estatísticos sejam

“atenuados”, uma vez que – em relação ao SS4 – há maior aproximação entre a habilidade preditiva dos modelos e a previsibilidade inerente ao sinal previsto (GHI)(KUMAR; CHEN, M., 2015).

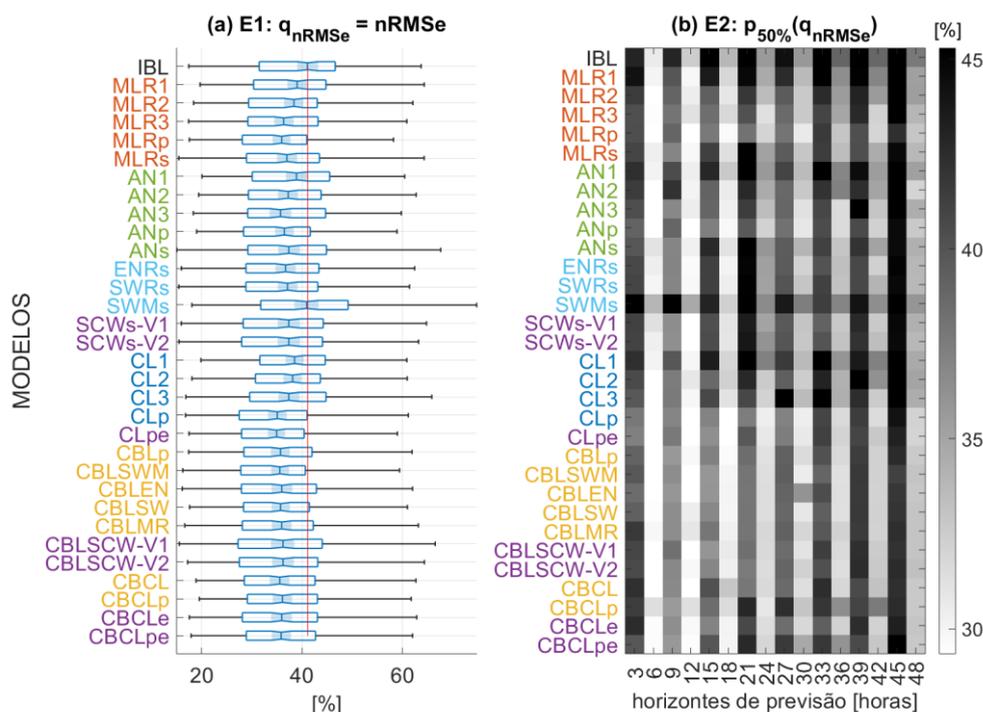
Figura 32 – Comparação entre espectrogramas obtidos para sinais de (a) velocidade do vento; (b) irradiância global horizontal.



Espectrogramas obtidos a partir da aplicação de SSWT (Seção 2.6) sobre séries temporais com taxa de amostragem de 6 horas (4 lançamentos por dia). Em (a): dados anemométricos de local #6, considerando o período entre 08/2019 e 08/2020; Em (b): dados solarimétricos para local #7, considerando o período entre 04/2020 e 08/2020. Fonte: autor.

A Figura 33 apresenta os resultados das previsões de curto prazo de GHI considerando o estatístico nRMSe. É importante ressaltar que para o cálculo de normalização do RMSe foram considerados amostras referentes a todos os lançamentos, o que inclui horários à noite. Dessa forma, devido à natureza da irradiância solar, os valores médios tendem a ser reduzidos ao serem incluídos de forma sistemática valores de irradiância nula. Tal característica tende a inflacionar os valores de nRMSe em comparação com aqueles tipicamente obtidos para a previsão de velocidade do vento a partir do GFS (Seção 5.1). Observa-se, no entanto, que os modelos estatísticos reduzem a mediana das distribuições de nRMSe para aproximadamente 35%, contra a mediana 41% do modelo IBL. Observa-se alternância no patamar de acurácia dos modelos entre horizontes pares e ímpares, com a maior parte dos resultados entre 30% e 35%, e entre 40% e 45%, respectivamente.

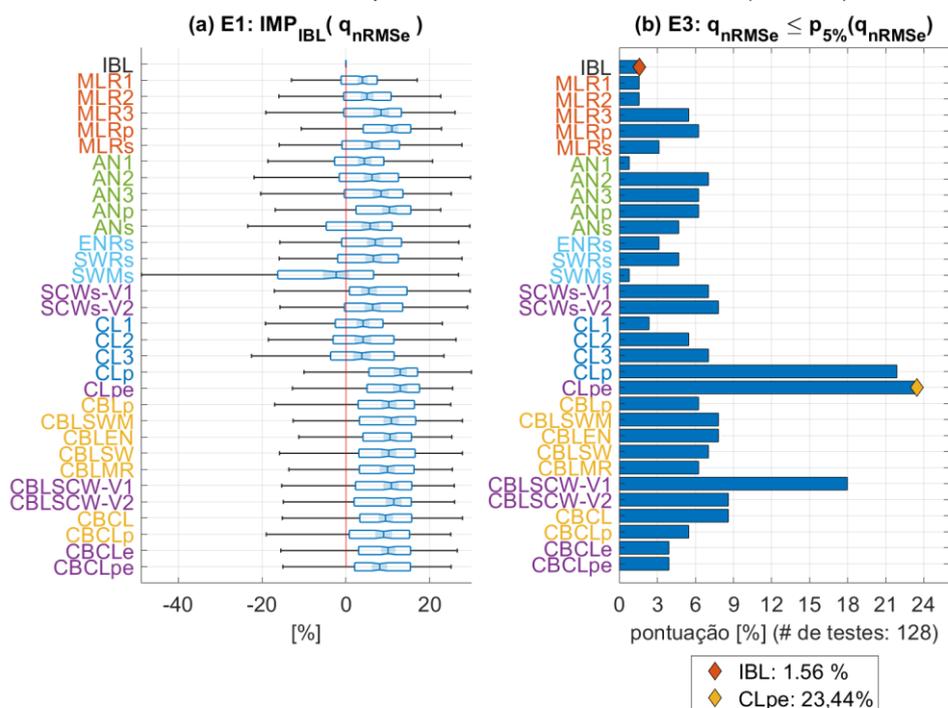
Figura 33 – Previsões de curto prazo de Irradiância Global Horizontal (GHI): RMSe normalizado pela média do sinal observado (nRMSe).



(a) Espaços E1: distribuições para os 128 testes; (b) Espaços E2: 8 testes por horizonte de previsão. Fonte: autor.

Na Figura 34, são mostrados os aprimoramentos sob perspectiva geral (a) e por meio de comparações realizadas sobre cada teste individualmente (b). Verifica-se perfil similar àquele apresentado Figura 31. Destaca-se, aprimoramento geral promovido pelo modelo CLpe, com mediana de 13%, sendo o maior dentre todos os demais. Ao serem comparados caso a caso, o modelo CLpe apresenta os melhores resultados em 23,44% dos testes, seguido do modelo CLp. Dentre os modelos de combinação, destaca-se o modelo CBLSCW-V1, tal como se observou na seção anterior (previsões para velocidade do vento). Dessa forma, avalia-se que os modelos baseados em complementaridade espectral propostos, também no caso de GHI, apresentam aprimoramentos importantes sobre aspectos diversos do sinal observado.

Figura 34 – Previsões de curto prazo de Irradiância Global Horizontal (GHI): aprimoramento sobre RMSe normalizado pela média do sinal observado ($nRMSe$).



(a) Espaços E1: distribuições de valores de aprimoramento percentual sobre o modelo de referência (IBL) para os 128 testes realizados; (b) Espaços E3: pontuam-se os melhores dentre os 32 modelos em cada caso. As barras indicam a pontuação total em função do número de testes realizados. Fonte: autor.

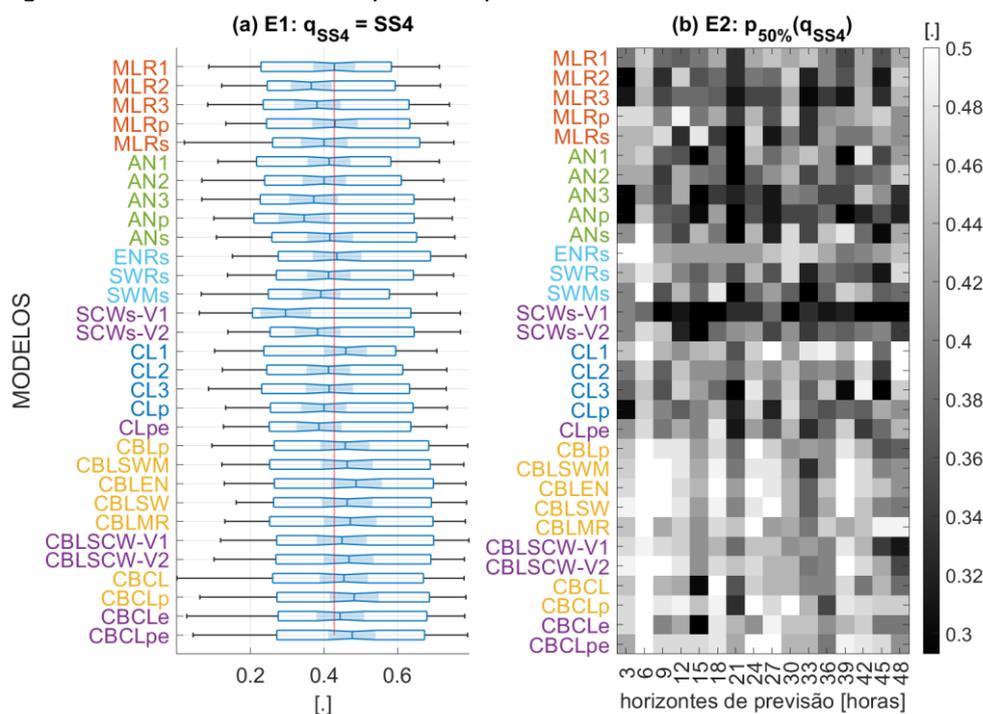
5.3 Previsões para geração de potência eólica

Nesta subseção são apresentados os resultados para a potência eólica gerada no local #6, referente a aerogerador de potência nominal 2350 kW. Como nos casos anteriores, serão avaliados diferentes meses de operação, horizontes de previsão, bem como configurações de hiperparâmetros (Tabela 6 e Tabela 7). Ao todo, foram considerados 96 testes no âmbito desta subseção.

A Figura 35 apresenta os resultados referentes ao índice de habilidade SS4, sobre os espaços E1 e E2, i.e., sob visão geral (a) e por horizonte de previsão (b), respectivamente. Para previsões de potência, adota-se o modelo MLR1 como referência contra qual se comparam os demais resultados, uma vez que o modelo IBL não permite a conversão da unidade da variável de entrada, não sendo, portanto, aplicável aos casos de previsão de potência gerada. Em relação aos valores de SS4 obtidos para velocidade do vento (Seção 5.1), observa-se valores similares na dispersão em relação aos valores de SS4 obtidos, entretanto, com redução sensível redução nos valores em geral (com medianas em torno de 0,4). Por outra parte, em

relação aos modelos individuais (de *downscaling* estatístico) observam-se padrões heterogêneos entre os diferentes horizontes de previsão, quando comparados ao caso de velocidade do vento exceto pelos modelos ENRs e CL1, os quais produzem resultados estáveis em diferentes horizontes de previsão. É possível observar, no entanto, melhoria da performance por parte dos modelos de combinação, produzindo resultados mais homogêneos em função dos diferentes horizontes de previsão.

Figura 35 – Previsões de curto prazo de potência eólica: Índice de habilidade “SS4”.



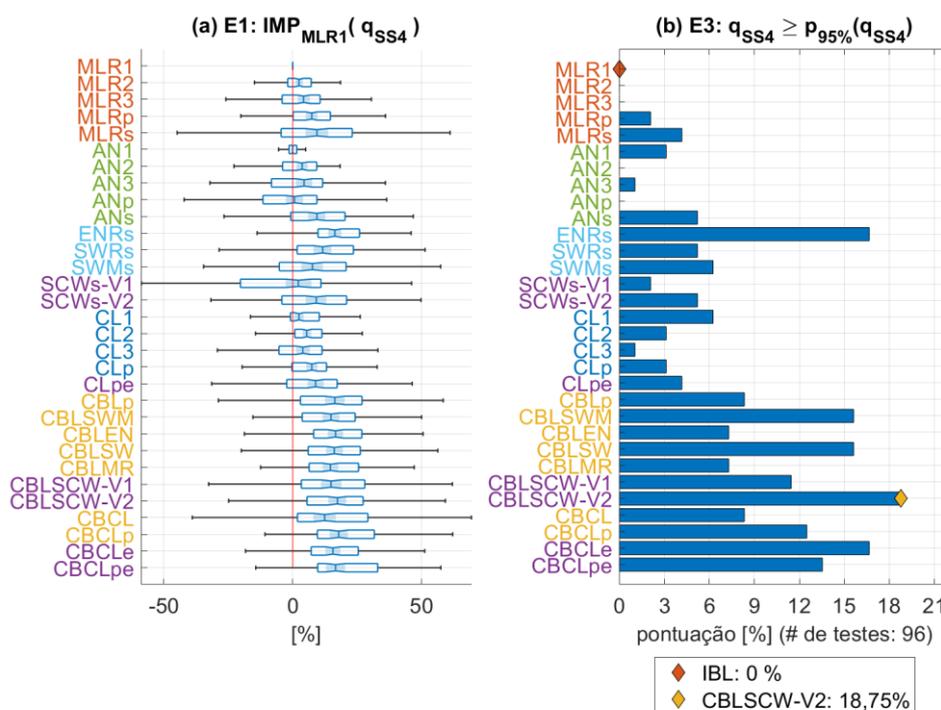
(a) Espaços E1: distribuições para os 96 testes; (b) Espaços E2: 6 testes por horizonte de previsão.
Fonte: autor.

A Figura 36 apresenta os resultados em termos de aprimoramento sobre os valores de SS4 resultantes do modelo de referência (MLR1), sendo avaliados tanto sob perspectiva global (a), bem como através de pontuação dada em função de comparações realizadas em cada teste separadamente (b).

A partir de progressão do tamanho do domínio aplicado aos modelos (e.g., modelos “1”, “2”, “3”, “p” e “s”), é possível observar tendência de crescimento nos valores positivos e negativos calculados para o aprimoramento, indicando que, por um lado, há potenciais relações que podem ser obtidas a partir dos modelos adotados, por outro lado, possivelmente há um aumento acentuado no risco associado aos pressupostos inerentes aos modelos. Tal afirmação se embasa no fato de que a velocidade do vento e a potência eólica gerada estão relacionados por polinômio de terceiro grau, enquanto que os modelos aplicados exploram relações conjuntas de

primeira ordem entre os preditores (i.e., regressões lineares múltiplas). Entretanto, de maneira geral, aprimoramentos de aproximadamente 20% foram obtidos em relação ao modelo de referência MLR1, especialmente no caso do modelo individual ENRs e os modelos de combinação em geral (modelos “CB”). Os modelos ENRs, CBLSWM, CBLSW, CBLSCW-V2 e CBCLe apresentaram os melhores resultados dentre todos os modelos avaliados, produzindo os melhores resultados em mais de 15% dos testes realizados. O modelo CBLSCW-V2 apresentou os melhores resultados, em geral, considerando a pontuação na Figura 36b (18,75%), com aprimoramento (relativo à mediana) de 17,2% sobre o modelo MLR1.

Figura 36 – Previsões de curto prazo de potência eólica: aprimoramento sobre índice de habilidade “SS4”.



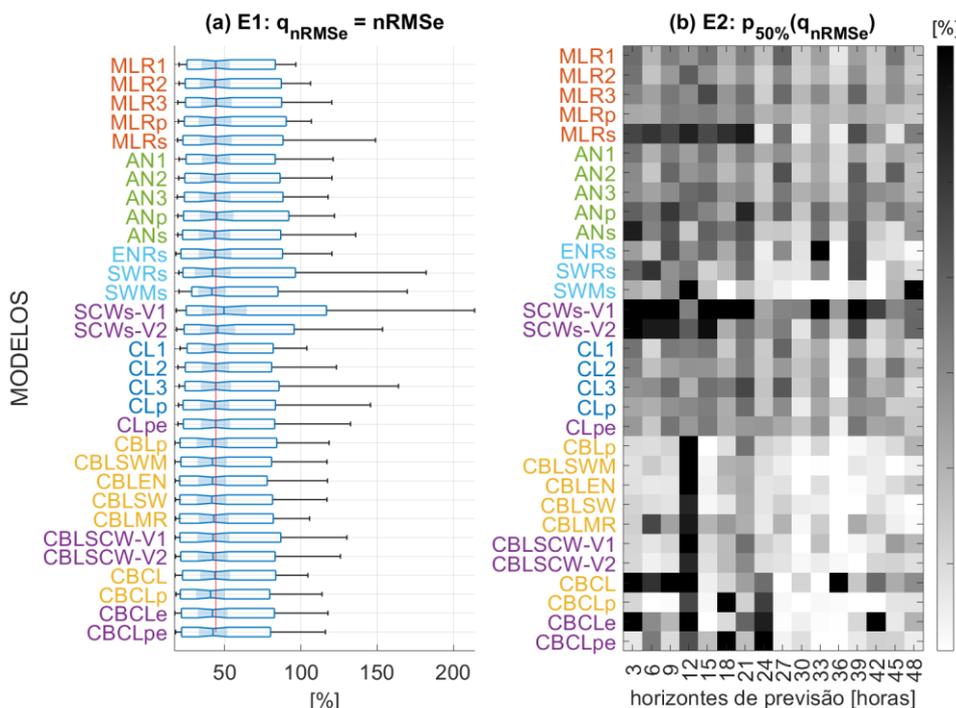
(a) Espaços E1: distribuições de valores de aprimoramento percentual sobre o modelo de referência (IBL) para os 96 testes realizados; (b) Espaços E3: pontuam-se os melhores dentre os 32 modelos em cada caso. As barras indicam a pontuação total em função do número de testes realizados. Fonte: autor.

A Figura 37 apresenta os resultados de potência eólica em termos de nRMSe. Observa-se que, de maneira geral, todos os modelos apresentam performance similar em relação aos erros apresentados, não havendo diferenças significativas em relação à mediana do modelo de referência MLR1. Observa-se, no entanto, maior instabilidade em relação aos modelos SCWs-V1. Em relação à mediana das distribuições, observam-se erros da ordem de 50% do valor médio das observações

em cada caso, o que consiste em aumento considerável se comparados aos erros obtidos para previsão de velocidade do vento. Exceto pelo horizonte de previsão de +12 horas, em comparação com modelos individuais, observa-se melhor performance por parte de modelos de combinação. Tal como no caso das previsões para velocidade do vento, observa-se maior acurácia nos horizontes de previsão pares. Para redução dos erros obtidos, é possível que, por um lado, modelos não lineares sejam capazes de melhor reproduzir a potência (particularmente em relação a valores extremos). Por outro lado, outras variáveis (e.g., temperatura ambiente) podem ser levadas em consideração, uma vez que podem impactar no funcionamento do sistema de geração. Ressalta-se, porém, que a validade dos métodos propostos é analisada a partir da comparação entre diferentes modelos (i.e., análise relativa), os quais são submetidos às mesmas condições de ajuste.

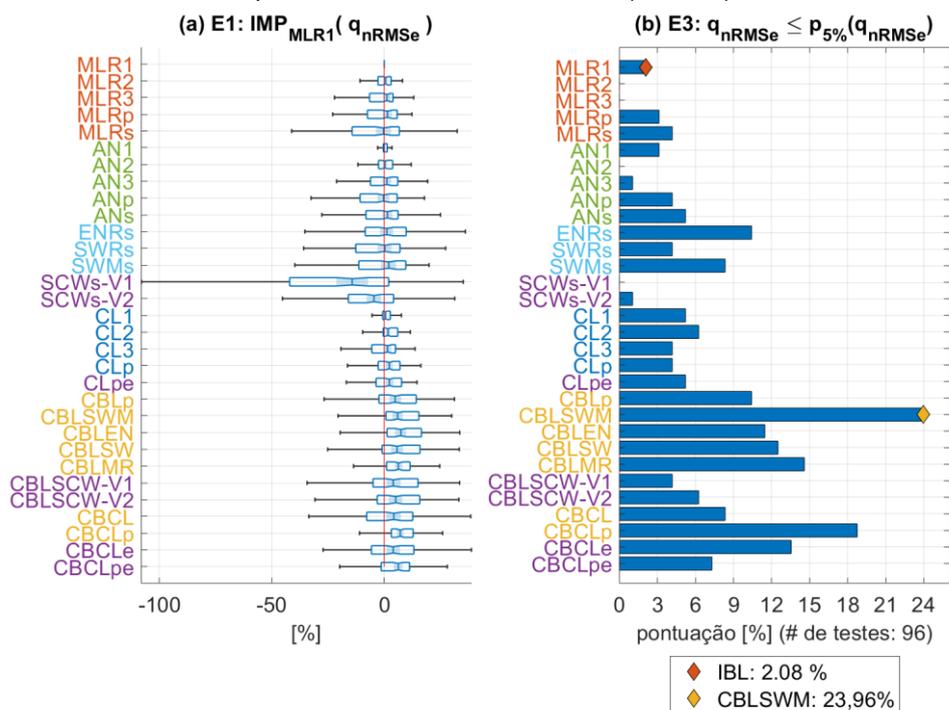
A Figura 38 apresenta os resultados para a previsão de potência eólica em termos de aprimoramentos sobre o nRMSe. Ao serem feitas comparações entre resultados dos modelos, evidenciam-se tendências de aprimoramento apenas em relação aos modelos de combinação (com medianas em torno de 7%), tal como avaliado no parágrafo anterior. Dentre os modelos analisados, o modelo CBLSWM apresenta os melhores resultados em relação ao nRMSe, sendo o mais acurado em aproximadamente 24% dos testes analisados nesta subseção. Destacam-se também os resultados obtidos pelos modelos CBLSW, CBLMR, CBCLp e CBCLe, os quais apresentam os melhores resultados em mais de 12% dos testes.

Figura 37 – Previsões de curto prazo de potência eólica: RMSe normalizado pela média do sinal observado ($nRMSe$).



(a) Espaços E1: distribuições para os 96 testes; (b) Espaços E2: 6 testes por horizonte de previsão. Fonte: autor.

Figura 38 – Previsões de curto prazo de potência eólica: aprimoramento sobre RMSe normalizado pela média do sinal observado ($nRMSe$).



(a) Espaços E1: distribuições de valores de aprimoramento percentual sobre o modelo de referência (IBL) para os 96 testes realizados; (b) Espaços E3: pontuam-se os melhores dentre os 32 modelos em cada caso. As barras indicam a pontuação total em função do número de testes realizados. Fonte: autor.

5.4 Previsões para geração de potência solar fotovoltaica

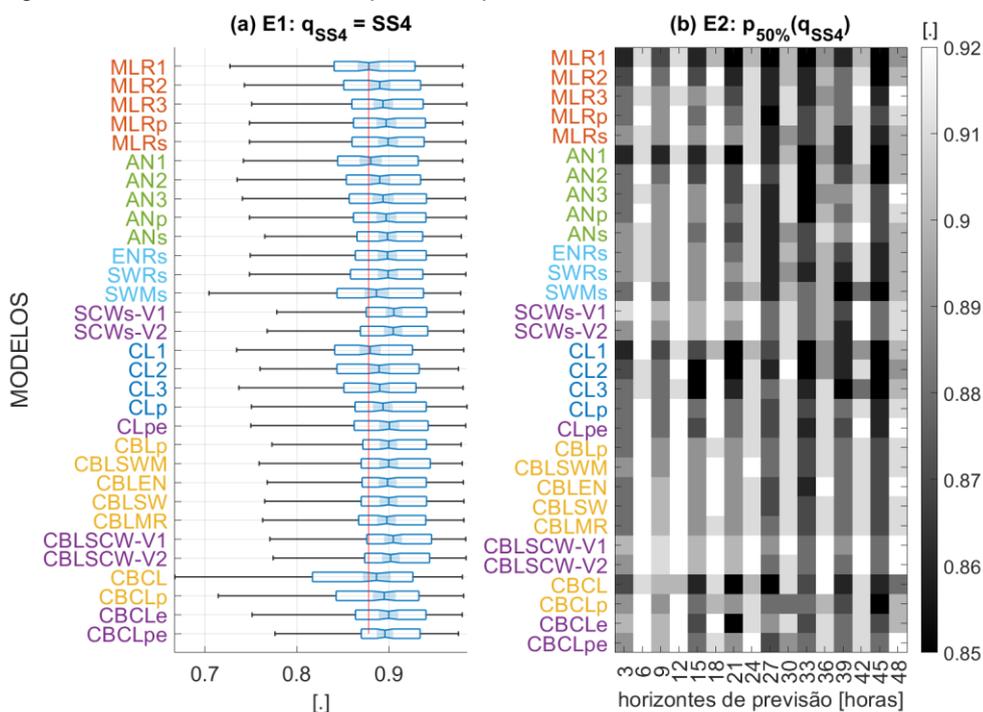
Nesta subseção são apresentados os resultados para a potência solar fotovoltaica gerada no local #7, referente a *stringbox* de potência nominal 50,16 kW. Ao todo, foram considerados 128 testes no âmbito desta subseção, incluindo os meses de operação (março, junho, agosto e outubro de 2020), bem como 16 horizontes de previsão e 2 configurações de hiperparâmetros (Tabela 6 e Tabela 7).

A Figura 39 mostra os resultados obtidos em relação ao índice de habilidade SS4 sob perspectiva geral (a) com discriminação entre horizontes de previsão (b). Tal como no caso referente às previsões de GHI (Seção 5.2), foram obtidos valores mais elevados, devido à intermitência natural da irradiância solar e, conseqüentemente, da potência fotovoltaica. Também de forma similar a casos anteriores, sobre os diferentes horizontes de previsão, observam-se dois patamares de valores de SS4, os quais se situam (na maior parte dos testes) abaixo de 0,9 nos horizontes ímpares, e acima desse mesmo valor em horizontes pares. Dentre os modelos de *downscaling* estatístico (individuais) e de combinação, destacam-se os modelos SCWs-V1, SCW-V2, CBLSCW-V1 e CBLSCW-V2 (os quais se baseiam na aplicação de pesos espectrais; Seção 4.4.3), os quais produziram resultados mais acurados e estáveis para os diferentes horizontes de previsão. Os resultados obtidos para os modelos de combinação locais baseados em complementaridade espectral CBCLe e CBCLpe demonstram redução na dispersão dos valores de SS4 que constam abaixo da linha de referência (que indica a mediana relativa ao modelo de referência MLR1), se comparados às dispersões dos modelos CBCL e CBCLp, o que indica aumento na robustez por parte dos novos modelos propostos.

Na Figura 40, em relação ao aprimoramento proporcionado pelos diferentes modelos (em relação ao modelos de referência MLR1), verificam-se maiores aprimoramentos para os modelos individuais com indicadores “3”, “p”, “s”, os quais representam diferentes configurações de domínio amplo (com maior cobertura espacial na entrada), com medianas em torno de 1,5%. De maneira geral, os resultados produzidos por tais modelos são consolidados pelos modelos de combinação (com exceção dos modelos CBCL e CBCLp), com sensível melhora por parte dos modelos CBLSCW-V1 e CBLSCW-V2, os quais produzem aprimoramentos

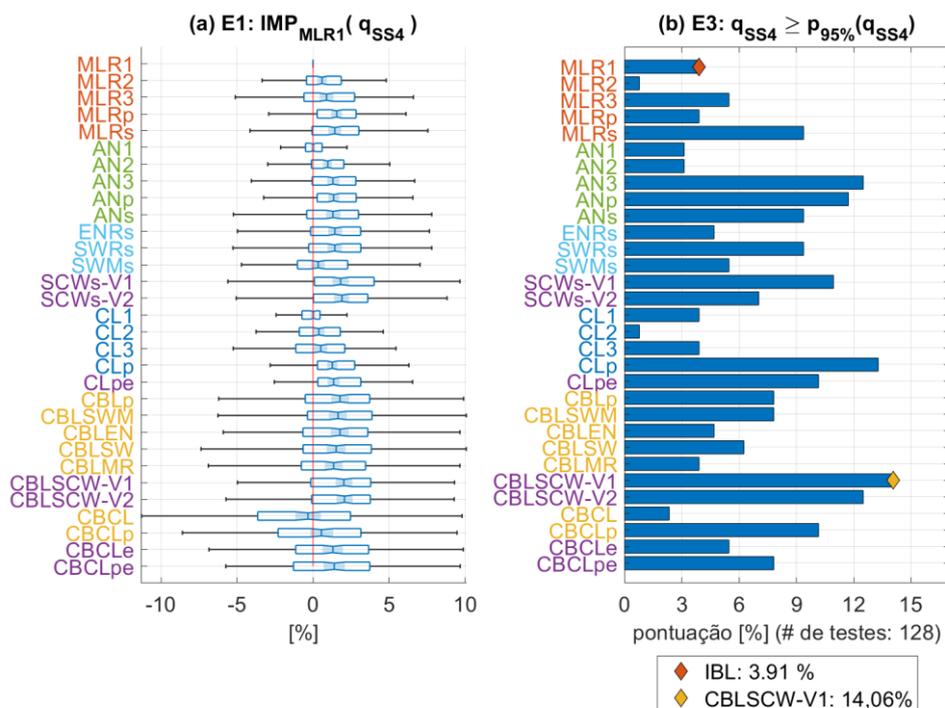
com medianas de 2%, além de produzirem os melhores resultados em 14% e 12,5% dos testes realizados.

Figura 39 – Previsões de curto prazo de potência solar: índices de habilidade “SS4”.



(a) Espaços E1: distribuições para os 128 testes; (b) Espaços E2: 8 testes por horizonte de previsão. Fonte: autor.

Figura 40 – Previsões de curto prazo de potência solar: aprimoramento sobre índices de habilidade “SS4”.

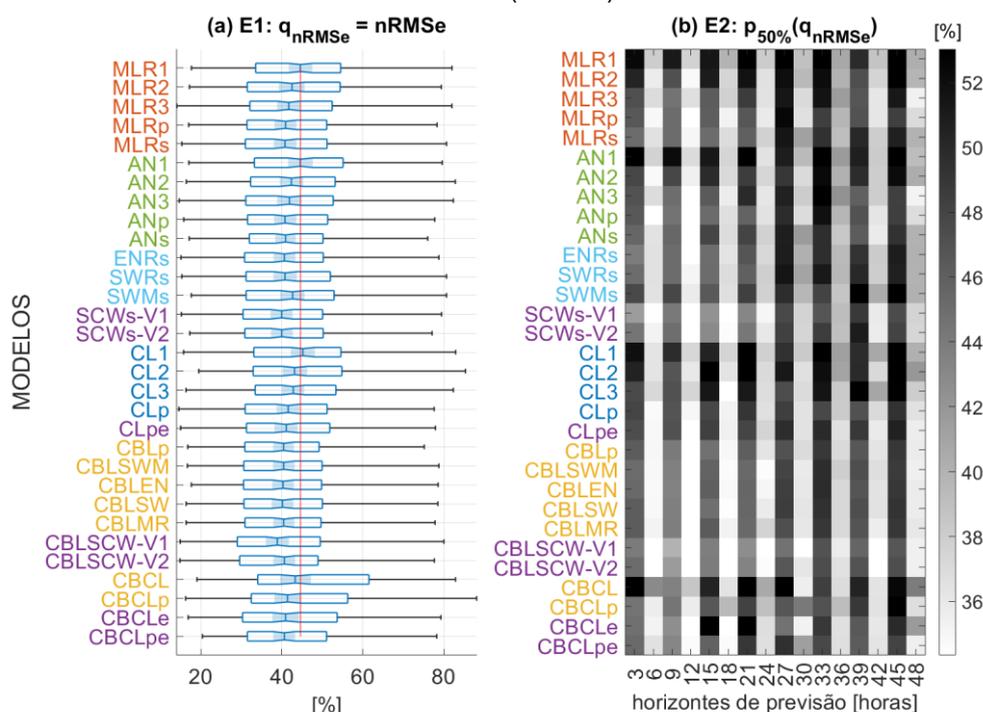


(a) Espaços E1: distribuições de valores de aprimoramento percentual sobre o modelo de referência (IBL) para os 128 testes realizados; (b) Espaços E3: pontuam-se os melhores dentre os 32 modelos em cada caso. As barras indicam a pontuação total em função do número de testes realizados. Fonte: autor.

Na Figura 41 são mostrados os resultados em termos de nRMSe para as previsões de potência solar fotovoltaica. Na maior parte dos casos, observa-se um aumento em relação aos erros normalizados obtidos para previsões de GHI (Seção 5.2). Como em casos anteriores, observa-se alternância na acurácia das previsões em horizontes pares e ímpares, com patamares de erro em torno de 36% e 50%, respectivamente. Em relação ao modelo de referência MLR1, observa-se tendência de redução nas medianas das distribuições, indicando melhoria da acurácia para modelos de *downscaling* estatístico e de combinação.

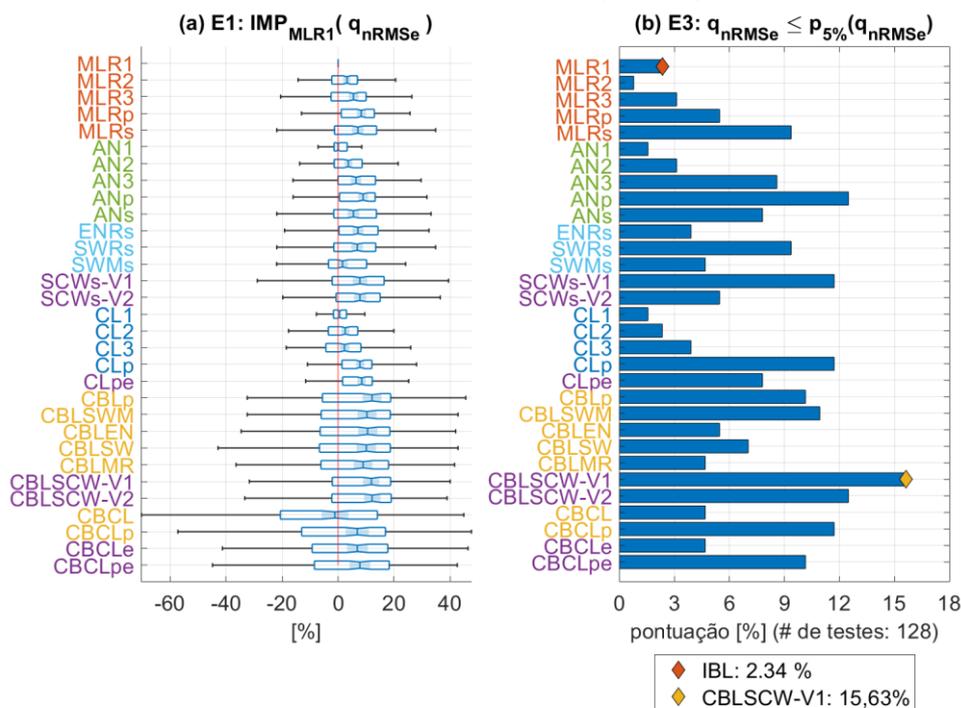
A Figura 42 apresenta os resultados de aprimoramento relativo ao estatístico nRMSe, tal como realizado em seções anteriores. Observa-se, também para o caso de previsões de potência solar fotovoltaica, que os modelos baseados em complementaridade espectral produzem aprimoramentos importantes (da ordem de 10%), sendo o modelo CBLSCW-V1 aquele com que se obtém a maior pontuação, apresentando os melhores resultados em 15,63% dos testes realizados.

Figura 41 – Previsões de curto prazo de potência solar: RMSe normalizado pela média do sinal observado (nRMSe).



(a) Espaços E1: distribuições para os 128 testes; (b) Espaços E2: 8 testes por horizonte de previsão. Fonte: autor.

Figura 42 – Previsões de curto prazo de potência solar: aprimoramento sobre RMSe normalizado pela média do sinal observado ($nRMSe$).



(a) Espaços E1: distribuições de valores de aprimoramento percentual sobre o modelo de referência (IBL) para os 128 testes realizados; (b) Espaços E3: pontuam-se os melhores dentre os 32 modelos em cada caso. As barras indicam a pontuação total em função do número de testes realizados. Fonte: autor.

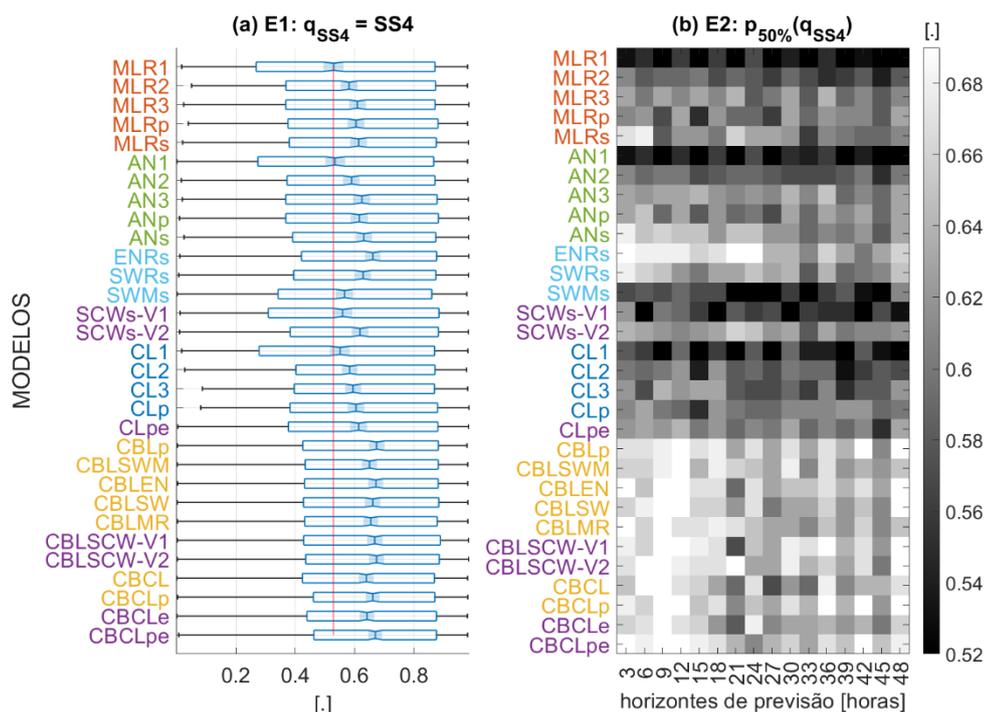
5.5 Resultados Gerais

Esta subseção se dedica à avaliação dos resultados considerando as previsões realizadas para todas as variáveis (velocidade do vento, GHI, potência eólica e solar fotovoltaica), em todos os locais, meses de operação, horizontes de previsão e configuração de hiperparâmetros. Tal como nos casos de potência eólica e solar, adota-se o modelo MLR1 como referência para a comparação dos resultados.

A Figura 43 apresenta os resultados relativos aos índices de habilidade SS4 considerando todos os 768 testes realizados no presente estudo. Os testes mostraram valores de SS4 mais baixos para previsões de velocidade do vento e potência eólica e mais altos para previsões de GHI e potência solar fotovoltaica, resultando nas dispersões demonstradas em (a). Em relação à mediana das distribuições totais de resultados para cada modelo, observa-se que, com respeito ao SS4, o modelo ENRs apresentou melhor performance dentre os modelos de *downscaling* estatístico (modelos individuais), com degradação sensível na acurácia dos resultados a partir

do segundo dia de previsão. Por outro lado, o modelo SCWs-V2 apresentou resultados relativamente homogêneos em função dos diferentes horizontes de previsão. Também se confirma, em relação aos modelos de combinação, maior acurácia quando se comparam os resultados àqueles obtidos por modelos individuais (com aprimoramentos significativos em todos os horizontes de previsão).

Figura 43 – Previsões de curto prazo para velocidade do vento, GHI, potência eólica e solar: índices de habilidade “SS4”.



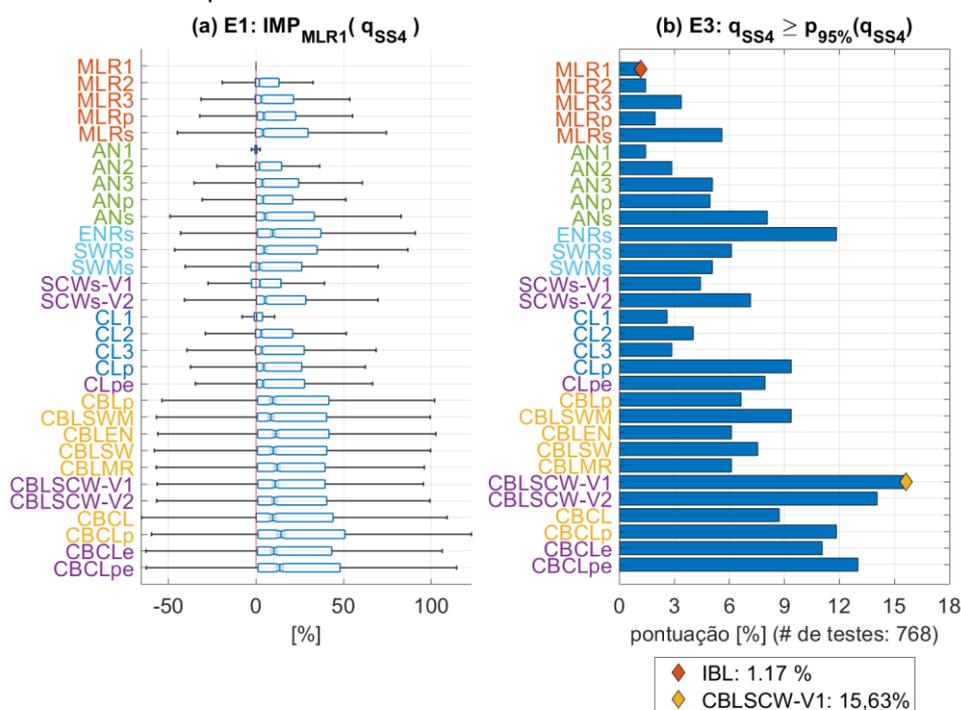
(a) Espaços E1: distribuições para os 768 testes; (b) Espaços E2: 48 testes por horizonte de previsão. Fonte: autor.

Os resultados apresentados na Figura 44 dizem respeito ao aprimoramento global obtido em relação ao SS4 sobre modelo de referência MLR1. Em geral, os modelos apresentam medianas positivas (indicando aprimoramento significativos sobre modelos mais simples). Em relação ao primeiro nível de modelos (modelos individuais ou de *downscaling* estatístico), verificam-se maiores aprimoramentos por parte de modelos de *downscaling* estatístico com maior cobertura espacial em seus respectivos domínios. Por outro lado, modelos de complementaridade espectral tendem a consolidar os melhores resultados ou aprimorar a acurácia em relação aos modelos no primeiro nível.

Em relação aos modelos baseados em complementaridade espectral propostos, ainda na Figura 44, observa-se que, para um número de testes entre aproximadamente 6% e 8% dos testes, os modelos SCWs-V1, SCWs-V2 e CLpe

apresentaram os melhores resultados. Por outro lado, em relação às combinações, os modelos propostos CBLSCW-V1, CBLSCW-V2, CBCLe e CBCLpe demonstraram maiores ganhos em relação aos demais – incluindo-se os mais próximos, como os modelos CBLen, CBCLe e CBCLpe. As maiores pontuações globais foram obtidas pelos modelos CBLSCW-V1 e CBLSCW-V2, que reproduzem as variáveis de interesse com maior acurácia em 15,63% e 14,1% do total de testes realizados. Em relação à mediana dos valores de aprimoramento percentual, o modelo CBCLpe apresentou os valores mais altos (próximo ao modelo CBCLp), de 14%. Entretanto, destacam-se que em boa parte dos casos, observam-se aprimoramentos de até 50% a partir da metodologia proposta.

Figura 44 – Previsões de curto prazo para velocidade do vento, GHI, potência eólica e solar: aprimoramentos sobre índices de habilidade “SS4”.

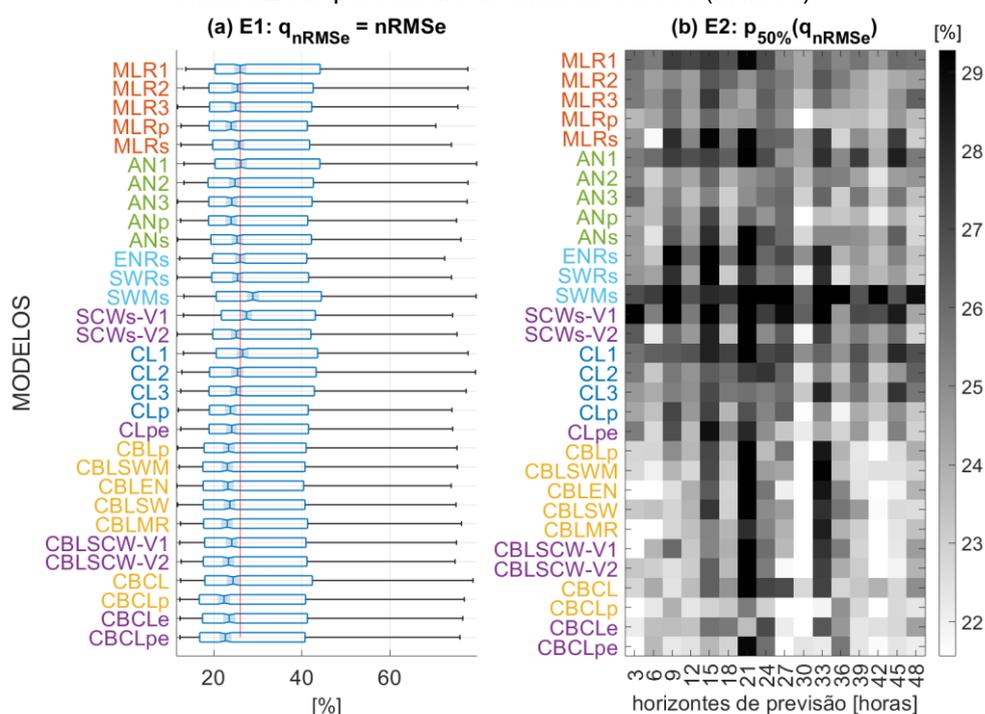


(a) Espaços E1: distribuições de valores de aprimoramento percentual sobre o modelo de referência (IBL) para os 768 testes realizados; (b) Espaços E3: pontuam-se os melhores dentre os 32 modelos em cada caso. As barras indicam a pontuação total em função do número de testes realizados. Fonte: autor.

Tal como em seções anteriores, as Figura 45 e Figura 46 apresentam os resultados gerais em relação ao nRMSe. Os resultados evidenciam perfil heterogêneo em função dos horizontes de previsão, com melhores resultados horizontes de previsão pares (+6, +12, +18 etc.). Além disso, observa-se redução na mediana das distribuições relativas aos modelos de combinação, tal como avaliado em relação ao SS4 (Figura 43). Em relação aos aprimoramentos obtidos, o modelo CBLSCW-V1

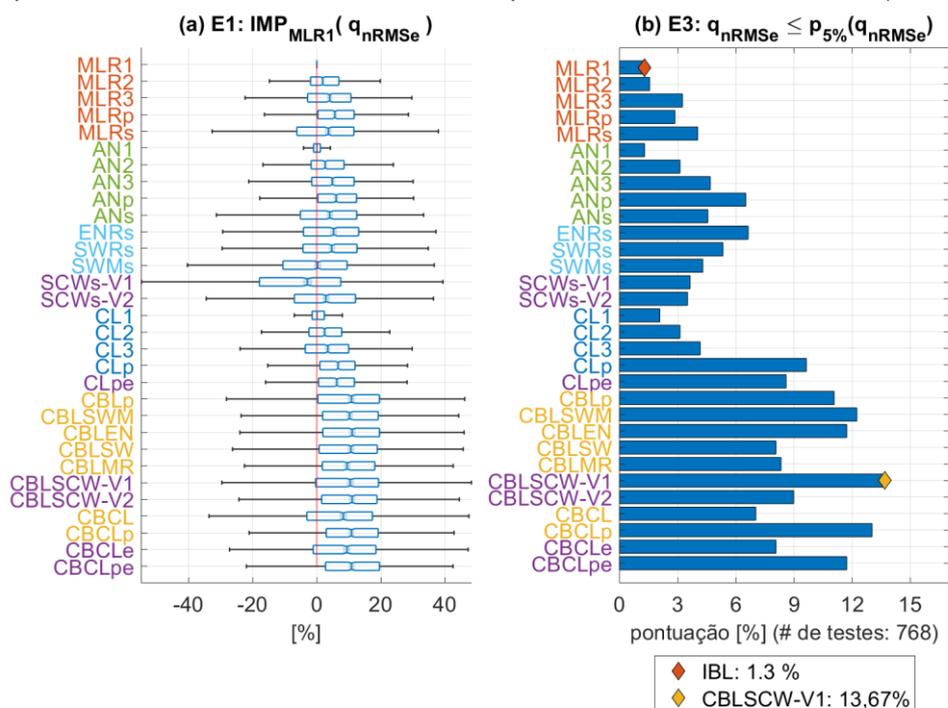
apresentou a pontuação mais alta, produzindo os melhores resultados em 13,67% dos testes realizados. Os resultados confirmam tendência que se mostrou em função das diferentes variáveis avaliadas, mostrando capacidade de reprodução de aspectos diversos do sinal observado em diferentes casos. Em comparação com os diversos modelos testados, a acurácia das previsões, em geral, pôde ser melhorada em função de erros sistemáticos (e.g., erros na amplitude de variações e erros no comportamento médio do sinal observado). Por outra parte, observou-se tendência de preservação da habilidade de reprodução da estrutura de fase e frequência do sinal observado, resultados que se expressam de forma evidente a partir da análise do índice de habilidade SS4 e métricas gerais de erro como o RMSe.

Figura 45 – Previsões de curto prazo para velocidade do vento, GHI, potência eólica e solar: RMSe normalizado pela média do sinal observado ($nRMSe$).



(a) Espaços E1: distribuições para os 768 testes; (b) Espaços E2: 48 testes por horizonte de previsão.
Fonte: autor.

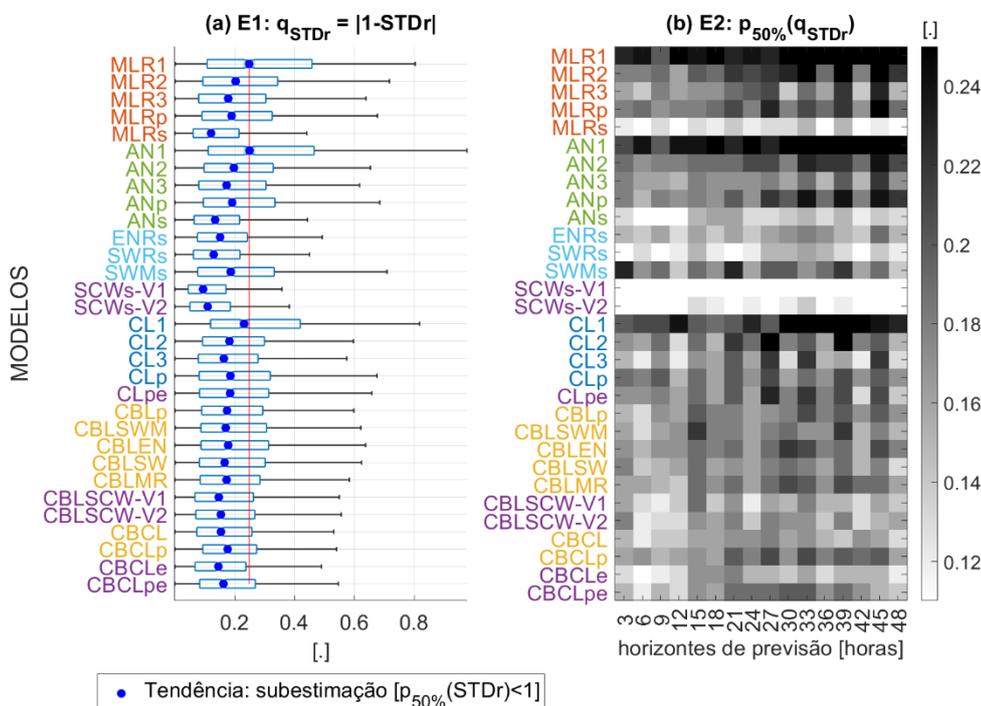
Figura 46 – Previsões de curto prazo para velocidade do vento, GHI, potência eólica e solar: aprimoramentos sobre RMSe normalizado pela média do sinal observado ($nRMSe$).



(a) Espaços E1: distribuições de valores de aprimoramento percentual sobre o modelo de referência (IBL) para os 768 testes realizados; (b) Espaços E3: pontuam-se os melhores dentre os 32 modelos em cada caso. As barras indicam a pontuação total em função do número de testes realizados. Fonte: autor.

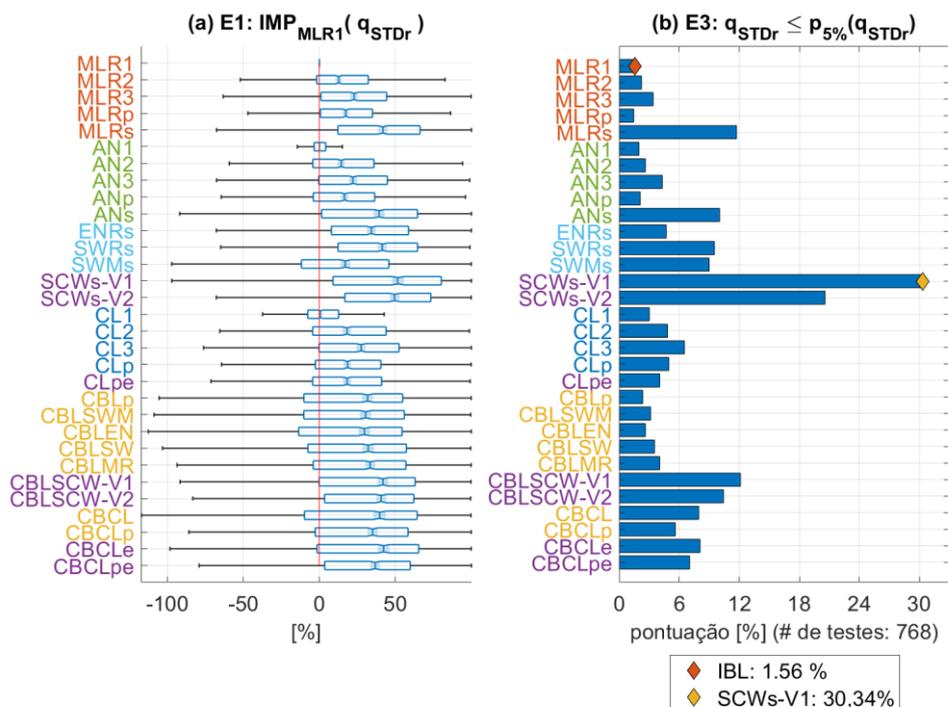
Para demonstrar a capacidade dos modelos propostos, considera-se também, nas Figura 47 e Figura 48, a avaliação dos resultados em relação à variância reproduzida ($STDr$; Tabela 5). Os resultados evidenciam que os modelos baseados em complementaridade espectral possuem maior capacidade de descrição da amplitude de variações do sinal observado (i.e., valores máximos e mínimos) quando comparados aos modelos semelhantes (grupo de modelos imediatamente acima em cada caso, tal como indicado pelas diferentes cores na abscissa). Os modelos SCWs-V1 e SCWs-V2 produzem os maiores aprimoramentos, cujas distribuições apresentam medianas de 51,6% e 50%, respectivamente. Além disso, as pontuações obtidas pelos mesmos modelos foram de 30,33% e 20,20%, indicando que em grande parte dos testes realizados, esses modelos produziram os melhores resultados. De forma similar, observa-se tendência de maiores pontuações em relação aos modelos de combinação baseados em complementaridade espectral (CBLSCW-V1, CBLSCW-V2, CBCLLe e CBCLpe). Tais resultados, sugerem que a aplicação de modelos baseados em complementaridade espectral produziu melhorias gerais nas previsões, com destaque à maior capacidade de reprodução da amplitude de variações do sinal observado.

Figura 47 – Previsões de curto prazo para velocidade do vento, GHI, potência eólica e solar: índices de habilidade “STDr”.



(a) Espaços E1: distribuições para os 768 testes; (b) Espaços E2: 48 testes por horizonte de previsão. Fonte: autor.

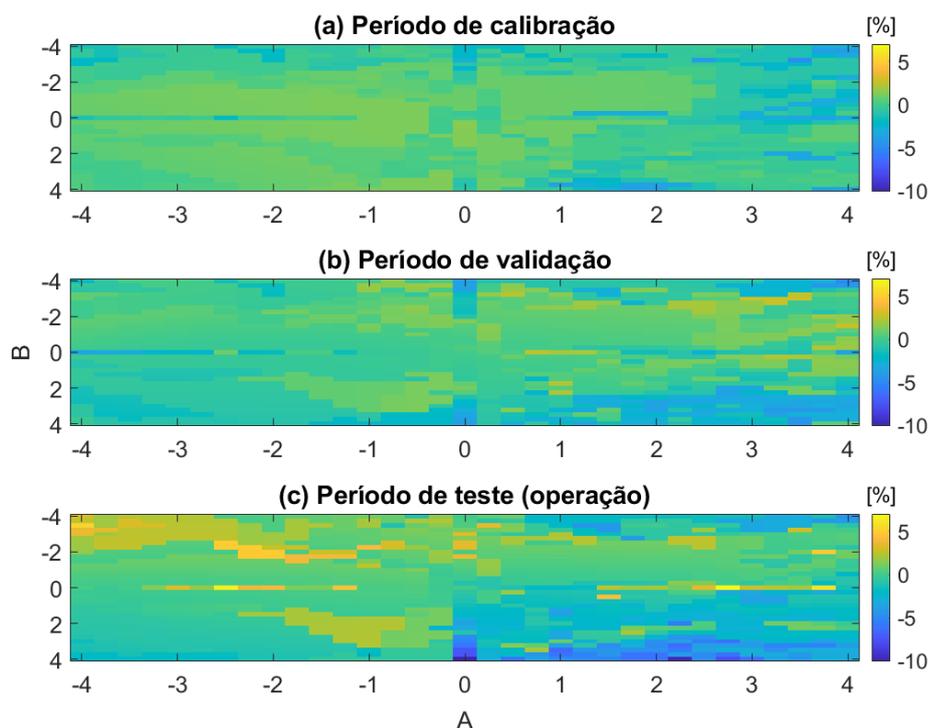
Figura 48 – Previsões de curto prazo para velocidade do vento, GHI, potência eólica e solar: aprimoramento sobre índice de habilidade “STDr”.



(a) Espaços E1: distribuições de valores de aprimoramento percentual sobre o modelo de referência (IBL) para os 768 testes realizados; (b) Espaços E3: pontuam-se os melhores dentre os 32 modelos em cada caso. As barras indicam a pontuação total em função do número de testes realizados. Fonte: autor.

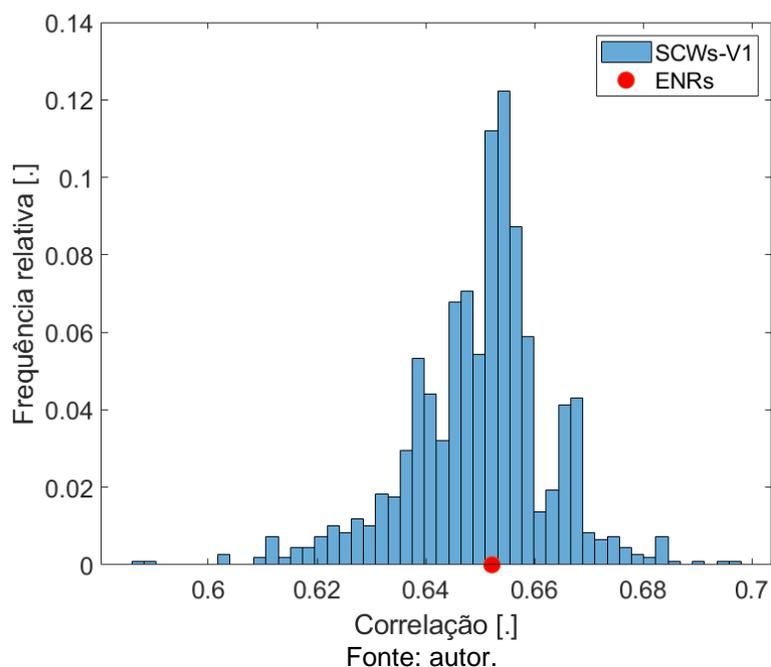
Apenas em um caso específico, para avaliar o potencial dos modelos SCW em relação à capacidade de reprodução da estrutura de fase e frequência, uma terceira variante do modelo (“V3”) foi submetida à análise de sensibilidade com respeito aos parâmetros A e B aplicados ao cálculo dos pesos espectrais (Eq. 34). A esse respeito, considerem-se os resultados mostrados na Figura 49, os quais foram obtidos para a estação anemométrica #6, considerando-se o mês de operação março de 2020, em horizonte de previsão para +3 horas à frente, configuração de treinamento tipo “A” (calibração com a primeira metade do conjunto de treinamento e validação na segunda metade). Os resultados mostrados em (a), (b) e (c) dizem respeito ao aprimoramento percentual da correlação obtida em relação ao modelo ENRs para os conjuntos de calibração (inferência de parâmetros), validação (otimização de hiperparâmetros) e teste (simulação da operação em tempo real). Na perspectiva oferecida pelo histograma da Figura 50, observa-se que, apesar da não trivialidade associada com a identificação das melhores soluções (em razão das mudanças – ou mesmo inversão – no perfil dos resultados percebidos nos diferentes períodos avaliados), a escolha apropriada de valores para A e B pôde trazer aprimoramentos acima de 5% em relação ao modelo base ENRs. Ressalta-se que tal resultado poderia ser obtido por meio de otimização, entretanto, aqui, tal análise é feita após o período de operação, não sendo um resultado válido para operação em tempo real. Por outro lado, tais resultados sugerem que, apesar do potencial da abordagem, a dificuldade demonstrada pelos modelos em convergir em soluções robustas se apresenta como a instabilidade na acurácia dos resultados, a exemplo do modelo dos SCWs-V1, o qual está habilitado para a aplicação de valores de $A, B \in [-3, 3]$, enquanto que a variante V2, a qual apenas busca valores entre -1 e 1, apresenta resultados mais estáveis.

Figura 49 – Análise de sensibilidade em relação aos coeficientes A e B aplicados ao cálculo dos pesos espectrais.



O gráfico mostra o aprimoramento percentual do índice de correlação de Pearson do modelo SCWs-V3 em relação ao modelo ENRs nos períodos de (a) calibração, (b) validação e (c) teste. Fonte: autor.

Figura 50 – Comparação entre as correlações dos modelos ENRs e os resultados da variante SCWs-V3 obtidos em análise de sensibilidade sobre os coeficientes A e B dos pesos espectrais.



6 CONCLUSÕES

No presente trabalho de tese, buscou-se avaliar metodologia desenvolvida com vistas a aplicação em previsão operacional (i.e., em tempo real) de variáveis de interesse para os setores de geração de energia eólica e solar. Nesse contexto, foram implementados e avaliados modelos estatísticos dispostos em três níveis, sendo o primeiro voltado para modelo para a seleção de domínio (ou de pontos de malha composta por saídas de modelo de larga escala), o segundo relativo aos modelos “individuais” ou de *downscaling* estatístico (partindo de previsões numéricas do tempo do GFS), e no terceiro, por fim, combinam-se as diferentes saídas produzidas no primeiro nível para o local de interesse. Tal estratégia permite melhor compreensão sobre vantagens e desvantagens em diferentes métodos, bem como o aprimoramento na acurácia de previsões produzidas por modelos consolidados na literatura e novos modelos propostos. Além disso, considerando que a qualidade das previsões possui diferentes aspectos (e.g., reprodução de valor médio, amplitude de variações, estrutura de fase e frequência etc.), avaliou-se a hipótese de que é possível melhorar ou modificar a habilidade de modelos estatísticos a partir da extração de atributos com base em análise de complementaridade espectral, pressupondo benefícios diversos à metodologia proposta.

Diferentes grupos de modelos foram considerados para o estudo, incluindo o mais simples (e.g., baseados em interpolação bilinear ou regressão multilinear) até modelos mais complexos (e.g., *clustering*, algoritmos de extração/seleção de atributos, modelos de combinação etc.). Dentre tais modelos, novos modelos foram propostos com base em métodos para a inferência de complementaridade espectral global (i.e., sobre todo o período observado) e local (i.e., avaliados para cada amostra do conjunto). Dessa forma, com base no cálculo dos atributos propostos, modificações foram realizadas a modelos apropriados para o objetivo almejado (e.g., modelos regressivos com arquitetura de parâmetros global e local). Três modificações principais foram propostas aos modelos: i) seleção de variáveis com base em índice de complementaridade espectral, que consiste em ponderação entre a relevância e redundância espectral entre sinais; ii) aplicação de pesos espectrais em método dos mínimos quadrados ponderados (*weighted least squares*); iii) aplicação da entropia espectral cruzada dos sinais para fins de agrupamento de padrões sinóticos. Sob diferentes recortes, a avaliação dos resultados dos métodos propostos foi realizada a

partir de comparação com modelos mais simples, bem como aqueles mais próximos em termos de complexidade (tendo como única distinção o fato de não serem aplicadas as modificações propostas).

Para a validação da metodologia proposta, foram considerados dados observacionais referentes a velocidade do vento, irradiância solar, potência eólica e solar fotovoltaica, aferidos em locais que se distribuem na Região Nordeste do Brasil. Além disso, foram avaliadas previsões para horizontes de previsão de até dois dias à frente (+48 horas), em passos de 3 horas. Em relação ao período avaliado, foram incluídos meses referentes a épocas de alta e baixa geração eólica/solar. Os resultados mostraram que a metodologia produziu resultados satisfatórios para a maior parte dos testes aplicados. Para todas as variáveis, os resultados mostraram maior acurácia em horizontes “pares” (+6, +12, +18 etc.), com valores de nRMSe em torno de 20% (considerando normalização pela média das observações). Para horizontes ímpares, por outro lado, os modelos apresentaram erros mais elevados, com nRMSe em torno de 30%.

Ao serem comparadas com as previsões para irradiância e potência solar, constatou-se que previsões realizadas para velocidade do vento e potência eólica demonstraram menor previsibilidade (com valores de SS4 em torno de 0,5), porém, com aprimoramentos em torno de 20% sobre os modelos de referência (IBL e MLR1). No caso das previsões para irradiância e potência solar, os índices de habilidade (SS4) se situaram em torno de 0,9, com a maior parte dos aprimoramentos sobre os modelos de referência de até 2%. Observou-se, ainda, aumentos nos erros (nRMSe) obtidos para as previsões de potência eólica, que na maioria dos casos foram acima de 40% da média da série observada, valores significativamente superiores àqueles obtidos para previsões de velocidade do vento (em torno de 20%). Para previsões de irradiância e potência solar, os valores de nRMSe se apresentaram entre 30% (horizontes pares) e 50% (horizontes ímpares) (considerando o período das noites para o cálculo da média aplicada à normalização). Avalia-se que seja possível reduzir os erros obtidos para sinais de potência a partir da adição de informações sobre outras variáveis que possam explicar variações no sinal de potência (e.g., temperatura ambiente). Por outro lado, modelos não lineares possuem maior potencial para associação entre saídas de modelos de larga escala e sinais de potência eólica/solar. No que diz respeito à avaliação comparativa entre os modelos, os resultados mostraram que a metodologia empregada é capaz de promover aprimoramentos em

torno de 10% sobre modelos mais simples. Note-se, no entanto, que tais valores de erro estão em conformidade com os valores tipicamente reportados na literatura em estudos similares (HOLMGREN; LORENZO; HANSEN, 2017; PÉREZ; NAVARRO, M. E. J.; CANTOS, 2016; ZHANG *et al.*, 2019).

A comparação entre os diferentes modelos avaliados revelou que os modelos de combinação foram capazes de aumentar a acurácia em relação aos modelos individuais, produzindo aprimoramentos significativos em relação aos modelos de referência na maior parte dos testes. Por outro lado, os modelos de combinação também se mostraram capazes de se adaptar aos modelos individuais (de *downscaling* estatístico) que produziram melhores resultados em casos diversos, conferindo robustez aos resultados finais.

Os novos modelos propostos, os quais se fundamentam em atributos de complementaridade espectral, produziram resultados importantes em relação aos demais modelos, demonstrando boa capacidade de reproduzir aspectos diversos da qualidade das previsões (e.g., estrutura de fase e frequência, amplitude de variações, valor médio etc.). Resultados complementares (em relação à qualidade das previsões) foram obtidos a partir do *downscaling* estatístico, os quais compuseram conjuntos de entrada para os modelos de combinação. A principal vantagem demonstrada por tais modelos consiste na habilidade de preservação da estrutura de fase e frequência, enquanto tende a reduzir que erros sistemáticos (e.g., bias ou variância reproduzida). Em comparação com modelos mais próximos (em termos de complexidade), os modelos propostos apresentaram pontuações maiores em diversos casos, indicando que tais modelos apresentaram os melhores resultados (em comparação com os demais modelos) para um maior número de testes realizados, melhorando a acurácia e robustez em comparação com outros modelos de combinação avaliados. Em corroboração com a hipótese fundamental deste trabalho de tese, os resultados obtidos sugerem que a metodologia proposta estatístico proposta foi beneficiada a partir da introdução de modelos baseados na complementaridade espectral entre os sinais relacionados.

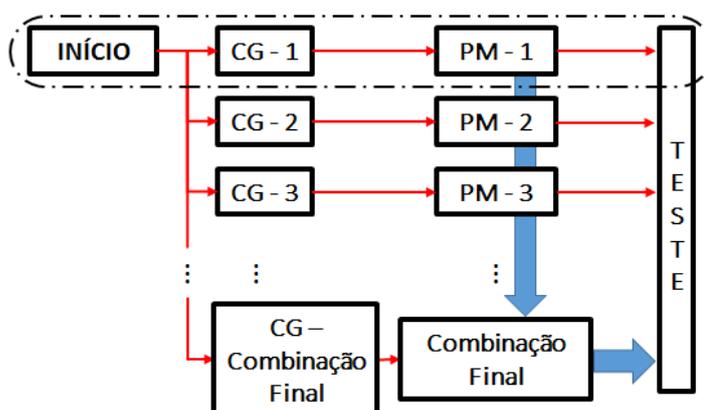
Por fim, como resultado dos testes avaliados neste trabalho, recomenda-se a aplicação da metodologia contendo os três níveis de combinação, considerando modelos de *downscaling* e combinação baseados em regressão em classificação e agrupamento de padrões, incluindo ainda modelos voltados para a seleção e extração de atributos, bem como modelos baseados em regularização (e.g., Stepwise

Regression, *Elastic Net*, *Clustering* etc.). Dentre todos os modelos avaliados, observou-se que os modelos baseados em pesos espectrais (“SCW”) demonstraram o maior equilíbrio entre aprimoramentos em acurácia e robustez frente aos demais modelos. Dessa forma, conclui-se que as técnicas baseadas em complementaridade espectral demonstraram capacidade de ressaltar relações distintas e relevantes de variáveis atmosféricas e de geração eólica e solar, contribuindo com resultados enriquecedores para a estratégia de combinação de modelos de *downscaling* estatístico avaliada.

7 PERSPECTIVAS DE DESENVOLVIMENTOS FUTUROS

Como perspectiva de trabalhos futuros para a metodologia proposta, soluções produzidas por configurações distintas podem ser combinadas por meio da adição de novas camadas de modelagem. Como exemplo, a Figura 51 ilustra a implementação do quarto nível à metodologia proposta neste trabalho, o qual permite combinar parametrizações para diferentes variáveis, locais e configurações de treinamento. Em tal abordagem, para futuros desenvolvimentos, propõe-se a aplicação de parâmetros obtidos para diversos aerogeradores ou sistemas fotovoltaicos, para enfim, realizar a previsão final para a central eólica e/ou solar.

Figura 51 – Exemplo de estratégia para a combinação de modelos no quarto nível.



Análise de sensibilidade relativa a diferentes configurações para a metodologia de *downscaling* estatístico e de combinação. Fonte: autor.

Em relação à aplicabilidade da metodologia proposta, propõe-se a investigação da utilidade dos métodos propostos em novas áreas, como em previsão probabilística, detecção e diagnóstico de falhas, dado que os resultados mostraram capacidade para a descrição de aspectos diversos dos sinais observados. Além disso, a partir dos novos modelos propostos se faz possível a avaliação do impacto da aplicação de filtros a diferentes componentes de frequência, isolando harmônicos característicos de fenômenos específicos (e.g., variações sazonais na relação entre larga e microescala). Os atributos propostos podem ser acoplados a algoritmos de modelos estatísticos diversos, incluindo modelos baseados em inteligência artificial. Nesse sentido, propõe-se a verificação de diferentes funções de transferência, incluindo modelos não lineares paramétricos (e.g., método de mínimos quadrados generalizado) ou não paramétricos (e.g., redes neurais).

Especificamente em relação à complementaridade espectral global, propõe-se a avaliação de algoritmos heurísticos voltados para a formação de conjuntos complementares. Para modelos regressivos, propõe-se a aplicação de saídas todas as saídas geradas a partir de variações em parâmetros associados com a complementaridade espectral (em lugar de recorrer aos algoritmos de otimização). Por outra parte, em relação aos modelos baseados em agrupamento de padrões, propõe-se a verificação de novos atributos baseados na entropia espectral cruzada, e.g., considerando-se a possibilidade de aplicação de filtros para a identificação e assimilação de processos específicos que podem ser extraídos dos sinais observados.

REFERÊNCIAS

ABEEÓLICA. ABEEólica ASSOCIAÇÃO BRASILEIRA DE ENERGIA EÓLICA E NOVAS TECNOLOGIAS ONSHORE OFFSHORE. 2021. p. 1–19.

ACCADIA, C. *et al.* Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. **Weather and Forecasting**, 2003. v. 18, n. 5, p. 918–932.

ADAM, S. P. *et al.* No free lunch theorem: A review. **Springer Optimization and Its Applications**, 2019. v. 145, p. 57–82.

AL-YAHYAI, S.; CHARABI, Y.; GASTLI, A. Review of the use of numerical weather prediction (NWP) models for wind energy assessment. **Renewable and Sustainable Energy Reviews**, 2010. v. 14, n. 9, p. 3192–3198.

ALLEN, J. B. Short Term Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform. **IEEE Transactions on Acoustics, Speech, and Signal Processing**, 1977. v. 25, n. 3, p. 235–238.

AMJADY, N.; KEYNIA, F.; ZAREIPOUR, H. Wind Power Prediction by a New Forecast Engine Composed of Modified Hybrid Neural Network and Enhanced Particle Swarm Optimization. **IEEE Transactions on Sustainable Energy**, jul. 2011. v. 2, n. 3, p. 265–276. Disponível em: <<http://ieeexplore.ieee.org/document/5713273/>>.

BACHER, P.; MADSEN, H.; NIELSEN, H. A. Online short-term solar power forecasting. **Solar Energy**, 2009. v. 83, n. 10, p. 1772–1783.

BAGHANAM, A. H. *et al.* Assessing the impact of climate change over the northwest of Iran: an overview of statistical downscaling methods. **Theoretical and Applied Climatology**, 2020. v. 141, n. 3–4, p. 1135–1150.

BARDOSSY, A.; BOGARDI, I.; MATYASOVSKY, I. Fuzzy rule-based downscaling of precipitation. **Theoretical and Applied Climatology**, 2005. v. 82, n. 1–2, p. 119–129.

BARNETT, T. P.; PREISENDORFER, R. W. Multifield Analog Prediction of Short-Term Climate Fluctuations Using a Climate State vector. **Journal of the Atmospheric Sciences**, 1978. v. 35, n. 10, p. 1771–1787.

BENAVIDES CESAR, L. *et al.* Review on Spatio-Temporal Solar Forecasting Methods Driven by In Situ Measurements or Their Combination with Satellite and Numerical Weather Prediction (NWP) Estimates. **Energies**, 14 jun. 2022. v. 15, n. 12, p. 4341. Disponível em: <<https://www.mdpi.com/1996-1073/15/12/4341>>.

BENESTAD, R. *et al.* Using statistical downscaling to assess skill of decadal predictions. **Tellus, Series A: Dynamic Meteorology and Oceanography**, 2019. v. 71, n. 1, p. 1–19.

BERGLAND, G. D. A guided tour of the fast Fourier Transform. **IEEE Spectrum**,

1969. v. 6, n. 7, p. 41–52.

BERNARDIN, F. *et al.* Stochastic downscaling method: Application to wind refinement. **Stochastic Environmental Research and Risk Assessment**, 2009. v. 23, n. 6, p. 851–859.

BERNHARDT, M. *et al.* High resolution modelling of snow transport in complex terrain using downscaled MM5 wind fields. **Cryosphere**, 2010. v. 4, n. 1, p. 99–113.

BORDONI, S.; STEVENS, B. Principal component analysis of the summertime winds over the Gulf of California: A gulf surge index. **Monthly Weather Review**, 2006. v. 134, n. 11, p. 3395–3414.

BÜRGER, G. Expanded downscaling for generating local weather scenarios. **Climate Research**, 1996. v. 7, n. 2, p. 111–128.

BUSENBARK, J. R. *et al.* Omitted Variable Bias: Examining Management Research With the Impact Threshold of a Confounding Variable (ITCV). **Journal of Management**, 2022. v. 48, n. 1, p. 17–48.

BUSUIOC, A.; TOMOZEIU, R.; CACCIAMANI, C. Statistical downscaling model based on canonical correlation analysis for winter extreme precipitation events in the Emilia-Romagna region. **International Journal of Climatology**, 2008. v. 28, n. 4, p. 449–464.

BUTTKUS, B. **Spectral Analysis and Filter Theory in Applied Geophysics**. [S.l.]: [s.n.], 2000.

CARVALHO, D. *et al.* WRF wind simulation and wind energy production estimates forced by different reanalyses: Comparison with observed data for Portugal. **Applied Energy**, 2014. v. 117, p. 116–126.

CATALÃO, J. P. S.; POUSINHO, H. M. I.; MENDES, V. M. F. Short-term wind power forecasting in Portugal by neural networks and wavelet transform. **Renewable Energy**, abr. 2011. v. 36, n. 4, p. 1245–1251. Disponível em: <<https://linkinghub.elsevier.com/retrieve/pii/S0960148110004477>>.

CHANDRASHEKAR, G.; SAHIN, F. A survey on feature selection methods. **Computers and Electrical Engineering**, 2014. v. 40, n. 1, p. 16–28.

CHATFIELD, C. **Time-series forecasting**. [S.l.]: CRC press, 2000.

CINELLI, C.; HAZLETT, C. Making sense of sensitivity: extending omitted variable bias. **Journal of the Royal Statistical Society. Series B: Statistical Methodology**, 2020. v. 82, n. 1, p. 39–67.

COSTA, A. *et al.* A review on the young history of the wind power short-term prediction. **Renewable and Sustainable Energy Reviews**, ago. 2008. v. 12, n. 6, p. 1725–1744. Disponível em: <<https://linkinghub.elsevier.com/retrieve/pii/S1364032107000354>>.

COULIBALY, P.; DIBIKE, Y. B.; ANCTIL, F. Downscaling precipitation and

temperature with temporal neural networks. **Journal of Hydrometeorology**, 2005. v. 6, n. 4, p. 483–496.

CURRY, C. L.; KAMP, D. VAN DER; MONAHAN, A. H. Statistical downscaling of historical monthly mean winds over a coastal region of complex terrain. I. Predicting wind speed. **Climate Dynamics**, 2012. v. 38, n. 7–8, p. 1281–1299.

DAMODAR, N. **Basic econometrics**. [S.l.]: The Mc-Graw Hill, 2004.

DARBELLAY, G. A.; VAJDA, I. Estimation of the information by an adaptive partitioning of the observation space. **IEEE Transactions on Information Theory**, 1999. v. 45, n. 4, p. 1315–1321.

DAUBECHIES, I.; LU, J.; WU, H. T. Synchrosqueezed wavelet transforms: An empirical mode decomposition-like tool. **Applied and Computational Harmonic Analysis**, 2011. v. 30, n. 2, p. 243–261.

DEE, D. P. *et al.* The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. **Quarterly Journal of the Royal Meteorological Society**, 2011. v. 137, n. 656, p. 553–597.

DIBIKE, Y. B.; COULIBALY, P. Temporal neural networks for downscaling climate variability and extremes. **Neural Networks**, 2006. v. 19, n. 2, p. 135–144.

DING, C.; PENG, H. Minimum redundancy feature selection from microarray gene expression data. **Journal of Bioinformatics and Computational Biology**, 2005. v. 3, n. 2, p. 185–205.

DING, S.; LI, R.; TAO, Z. A novel adaptive discrete grey model with time-varying parameters for long-term photovoltaic power generation forecasting. **Energy Conversion and Management**, 2021. v. 227.

DONG, Y. *et al.* Wind power forecasting based on stacking ensemble model, decomposition and intelligent optimization algorithm. **Neurocomputing**, 2021. v. 462, p. 169–184.

DOOL, H. M. VAN DEN. **Searching for analogues, how long must we wait?** **Tellus A**. Disponível em: <[http://onlinelibrary.wiley.com/doi/10.1034/j.1600-0870.1994.t01-2-00006.x/abstract%5Cnhttp://onlinelibrary.wiley.com/store/10.1034/j.1600-0870.1994.t01-2-00006.x/asset/j.1600-0870.1994.t01-2-00006.x/pdf?v=1&t=hvmgvsxv&s=2a28400ee9de7aff83cfc126403552c0fd8](http://onlinelibrary.wiley.com/doi/10.1034/j.1600-0870.1994.t01-2-00006.x/abstract%5Cnhttp://onlinelibrary.wiley.com/store/10.1034/j.1600-0870.1994.t01-2-00006.x/asset/j.1600-0870.1994.t01-2-00006.x.pdf?v=1&t=hvmgvsxv&s=2a28400ee9de7aff83cfc126403552c0fd8)>.

ENKE, W.; SPEKAT, A. Downscaling climate model outputs into local and regional weather elements by classification and regression. **Climate Research**, 1997. v. 8, n. 3, p. 195–207.

EPE. Balanço Energético Nacional (BEN) 2022: Ano base 2021 - Relatório Final. 2022. p. 264. Disponível em: <<http://www.epe.gov.br>>.

FENG, C. *et al.* A data-driven multi-model methodology with deep feature selection for short-term wind forecasting. **Applied Energy**, 2017. v. 190, p. 1245–1257.

FU, Y. *et al.* Assessment and prediction of regional climate based on a multimodel ensemble machine learning method. **Climate Dynamics**, 2023.

GAO, L.; SCHULZ, Karsten; BERNHARDT, Matthias. Statistical downscaling of ERA-interim forecast precipitation data in complex terrain using lasso algorithm. **Advances in Meteorology**, 2014. v. 2014.

GEORGE, J.; JANAKI, L.; PARAMESWARAN GOMATHY, J. Statistical Downscaling Using Local Polynomial Regression for Rainfall Predictions – A Case Study. **Water Resources Management**, 2016. v. 30, n. 1, p. 183–193.

GÓMEZ, D.; ROJAS, A. An empirical overview of the no free lunch theorem and its effect on real-world machine learning classification. **Neural Computation**, 2016. v. 28, n. 1, p. 216–228.

GOUTTE, C. Note on Free Lunches and Cross-Validation. **Neural Computation**, 1997. v. 9, n. 6, p. 1245–1249.

GRIFFIN, D. W.; LIM, J. S. Signal Estimation from Modified Short-Time Fourier Transform. **IEEE Transactions on Acoustics, Speech, and Signal Processing**, 1984. v. 32, n. 2, p. 236–243.

GRINSTED, A.; MOORE, J. C.; JEVREJEVA, S. Application of the cross wavelet transform and wavelet coherence to geophysical time series. **Nonlinear Processes in Geophysics**, 2004. v. 11, n. 5/6, p. 561–566.

GROCH, M.; VERMEULEN, J. Short-Term Ensemble NWP Wind Speed Forecasts using Mean-Variance Portfolio Optimization and Neural Networks. [S.l.]: IEEE, 2019. p. 1–6. Disponível em: <<https://ieeexplore.ieee.org/document/8783505/>>.

GUEYMARD, C. SMARTS2: a simple model of the atmospheric radiative transfer of sunshine: algorithms and performance assessment. **Florida Solar Energy Center**, 1995.

GURLEY, K.; KAREEM, A. **Applications of wavelet transforms in earthquake, wind and ocean engineering**. [S.l.]: Engineering structures, 1999.

GUTIÉRREZ, J. M. *et al.* Clustering methods for statistical downscaling in short-range weather forecasts. **Monthly Weather Review**, 2004. v. 132, n. 9, p. 2169–2183.

GUYON, ISABELLE; ELISSEEFF, A. An introduction to variable and feature selection. **Journal of machine learning research**, 2003. v. 3, p. 1157–1182. Disponível em: <<https://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf?ref=driverlayer.com/web>>. Acesso em: 05 de Jan. 2023>.

HAMMAMI, D. *et al.* Predictor selection for downscaling GCM data with LASSO. **Journal of Geophysical Research Atmospheres**, 2012. v. 117, n. 17.

HARPHAM, C.; WILBY, Robert L. Multi-site downscaling of heavy daily precipitation occurrence and amounts. **Journal of Hydrology**, 2005. v. 312, n. 1–4, p. 235–255.

HART, N. C. G.; GRAY, S. L.; CLARK, P. A. Detection of coherent airstreams using cluster analysis: Application to an extratropical cyclone. **Monthly Weather Review**, 2015. v. 143, n. 9, p. 3518–3531.

HE, R. R. *et al.* LASSO as a tool for downscaling summer rainfall over the Yangtze River Valley. **Hydrological Sciences Journal**, 2019. v. 64, n. 1, p. 92–104.

HEGERMILLER, C. A. *et al.* A multimodal wave spectrum-based approach for statistical downscaling of local wave climate. **Journal of Physical Oceanography**, 2017. v. 47, n. 2, p. 375–386.

HESSAMI, M. *et al.* Automated regression-based statistical downscaling tool. **Environmental Modelling and Software**, 2008. v. 23, n. 6, p. 813–834.

HIBON, M.; EVGENIOU, T. To combine or not to combine: selecting among forecasts and their combinations. **International Journal of Forecasting**, jan. 2005. v. 21, n. 1, p. 15–24. Disponível em:
<<https://linkinghub.elsevier.com/retrieve/pii/S0169207004000494>>.

HIDALGO, HUGO G.; DETTINGER, MICHAEL D.; CAYAN, D. R. **ownscaling with constructed analogues: Daily precipitation and temperature fields over the United States**. [S.l.]: [s.n.], 2008.

HOLMGREN, W.; LORENZO, A.; HANSEN, C. A Comparison of PV Power Forecasts Using PVLlib-Python. [S.l.]: IEEE, 2017. p. 1127–1131.

HORTON, P.; BRÖNNIMANN, S. Impact of global atmospheric reanalyses on statistical precipitation downscaling. **Climate Dynamics**, 2019. v. 52, n. 9–10, p. 5189–5211.

HORVATH, K.; BAJIĆ, A.; IVATEK-ŠAHDAN, S. Dynamical downscaling of wind speed in complex terrain prone to bora-type flows. **Journal of Applied Meteorology and Climatology**, 2011. v. 50, n. 8, p. 1676–1691.

HUTCHINSON, M. F. Stochastic space-time weather models from ground-based data. **Agricultural and Forest Meteorology**, 1995. v. 73, n. 3–4, p. 237–264.

HUTH, R. Statistical downscaling in central Europe: Evaluation of methods and potential predictors. **Climate Research**, 1999. v. 13, n. 2, p. 91–101.

INMAN, R. H.; PEDRO, H. T. C.; COIMBRA, C. F. M. Solar forecasting methods for renewable energy integration. **Progress in Energy and Combustion Science**, 2013. v. 39, n. 6, p. 535–576.

JAKOB THEMESSEL, M.; GOBIET, A.; LEUPRECHT, A. Empirical-statistical downscaling and error correction of daily precipitation from regional climate models. **International Journal of Climatology**, 2011. v. 31, n. 10, p. 1530–1544.

JIANG, Y. *et al.* Short-term wind speed prediction using time varying filter-based empirical mode decomposition and group method of data handling-based hybrid model. **Energy Conversion and Management**, 2020. v. 220.

JIMENEZ, P. A. *et al.* WRF-SOLAR: Description and clear-sky assessment of an augmented NWP model for solar power prediction. **Bulletin of the American Meteorological Society**, 2016. v. 97, n. 7, p. 1249–1264.

JOLLIFFE, I. T. **Principle component analysis**. [S.l.]: Springer Verlag, 1986.

JORDAN, D.; MIKSAD, R. W.; POWERS, E. J. Implementation of the continuous wavelet transform for digital time series analysis. **Review of Scientific Instruments**, 1997. v. 68, n. 3, p. 1484–1494.

JOVIĆ, A.; BRKIĆ, K.; BOGUNOVIĆ, N. A review of feature selection methods with applications. **2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2015 - Proceedings**, 2015. p. 1200–1205.

JUNG, J.; BROADWATER, R. P. Current status and future advances for wind speed and power forecasting. **Renewable and Sustainable Energy Reviews**, mar. 2014. v. 31, p. 762–777. Disponível em:
<<https://linkinghub.elsevier.com/retrieve/pii/S1364032114000094>>.

_____; TAM, K.-S. A frequency domain approach to characterize and analyze wind speed patterns. **Applied Energy**, mar. 2013. v. 103, p. 435–443.

KALNAY, E. *et al.* The NCEP/NCAR 40-year reanalysis project. **Bulletin of the American Meteorological Society**, 1996. v. 77, n. 3, p. 437–471.

KANG, H. *et al.* Statistical Downscaling of Precipitation in Korea Using Multimodel Output Variables as Predictors. **Monthly Weather Review**, 1 jun. 2009. v. 137, n. 6, p. 1928–1938. Disponível em:
<<http://journals.ametsoc.org/doi/10.1175/2008MWR2706.1>>.

KARTHIKEYA, B. R.; NEGI, P. S.; SRIKANTH, N. Wind resource assessment for urban renewable energy application in Singapore. **Renewable Energy**, 2016. v. 87, p. 403–414.

KIRCHMEIER, M. C.; LORENZ, D. J.; VIMONT, D. J. Statistical downscaling of daily wind speed variations. **Journal of Applied Meteorology and Climatology**, 2014. v. 53, n. 3, p. 660–675.

KOHAVI, R. *Et Al.* A study of cross-validation and bootstrap for accuracy estimation and model selection. [S.l.]: Ijcai, 1995. p. Vol. 2, 1137-1145.

KUMAR, A.; CHEN, M. Inherent predictability, requirements on the ensemble size, and complementarity. **Monthly Weather Review**, 2015. v. 143, n. 8, p. 3192–3203.

LAJNEF, T. *et al.* Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines. **Journal of Neuroscience Methods**, 2015. v. 250, p. 94–105.

LANDBERG, L.; WATSON, S. J. Short-term prediction of local wind conditions. **Boundary-Layer Meteorology**, 1994. v. 70, n. 1–2, p. 171–195.

LEE, J.; ZHAO, F. Global Wind Report | GWEC. **Global Wind Energy Council**, 2021. p. 75. Disponível em: <<http://www.gwec.net/global-figures/wind-energy-global-status/>>.

LEEDS, W. B.; STEIN, M. L.; MOYER., E. J. **Conditional Simulation of Future Climate Under Changing Temporal Covariance Structures**. Disponível em: <http://geosci.uchicago.edu/~moyer/MoyerWebsite/PrivateFiles_unlinked/Leeds_etal_CondSim.pdf>.

LI, X.; ZHANG, Xiaodong; WANG, Shuguang. A Hybrid Statistical Downscaling Framework Based on Nonstationary Time Series Decomposition and Machine Learning. **Earth and Space Science**, 2022. v. 9, n. 6.

LI, Yi; DING, A. A. Double-structured sparse multitask regression with application of statistical downscaling. **Environmetrics**, 2019. v. 30, n. 4.

LINDERSON, M. L.; ACHBERGER, C.; CHEN, D. Statistical downscaling and scenario construction of precipitation in Scania, southern Sweden. **Nordic Hydrology**, 2004. v. 35, n. 3, p. 261–278.

LIU, Hui *et al.* A hybrid model for wind speed prediction using empirical mode decomposition and artificial neural networks. **Renewable Energy**, 2012. v. 48, p. 545–556.

LOPES, J. A. P. *et al.* Integrating distributed generation into electric power systems: A review of drivers, challenges and opportunities. **Electric Power Systems Research**, 2007. v. 77, n. 9, p. 1189–1203.

LYDIA, M. *et al.* Linear and non-linear autoregressive models for short-term wind speed forecasting. **Energy Conversion and Management**, 2016. v. 112, p. 115–124.

MA, Y. *et al.* Short-term photovoltaic power forecasting method based on irradiance correction and error forecasting. **Energy Reports**, nov. 2021. v. 7, p. 5495–5509. Disponível em: <<https://linkinghub.elsevier.com/retrieve/pii/S2352484721007708>>.

MADDEN, R. A.; JULIAN, P. R. Detection of a 40–50 Day Oscillation in the Zonal Wind in the Tropical Pacific. **Journal of the Atmospheric Sciences**, 1971. v. 28, n. 5, p. 702–708.

MAO, Y.; MONAHAN, A. Linear and nonlinear regression prediction of surface wind components. **Climate Dynamics**, 2018. v. 51, n. 9–10, p. 3291–3309.

MARKOVICS, D.; MAYER, M. J. Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction. **Renewable and Sustainable Energy Reviews**, 2022. v. 161.

MEDSKER, LARRY R.; JAIN, L. C. **Recurrent neural networks**. [S.l.]: CRC Press LLC, 2001.

MEIGNEN, S.; OBERLIN, T.; PHAM, D. H. Synchrosqueezing transforms: From low- to high-frequency modulations and perspectives. **Comptes Rendus Physique**, 2019.

v. 20, n. 5, p. 449–460.

MENDES, D.; MARENGO, J. A. Temporal downscaling: A comparison between artificial neural network and autocorrelation techniques over the Amazon Basin in present and future climate change scenarios. **Theoretical and Applied Climatology**, 2010. v. 100, n. 3, p. 413–421.

MILLSTEIN, D. *et al.* Wind energy variability and links to regional and synoptic scale weather. **Climate Dynamics**, 2019. v. 52, n. 7–8, p. 4891–4906.

MONAHAN, A. H. Can we see the wind? Statistical downscaling of historical sea surface winds in the subarctic northeast Pacific. **Journal of Climate**, 2012. v. 25, n. 5, p. 1511–1528.

MORAES, C. F. W. De C. B. Procedimento objetivo para a garantia de qualidade de dados observacionais de vento em superfície no litoral do Rio Grande do Norte. **Procedimento objetivo para a garantia de qualidade de dados observacionais de vento em superfície no litoral do Rio Grande do Norte**, 2015. p. 89. Disponível em: <<https://repositorio.ufpe.br/handle/123456789/19714>>.

MUÑOZ-SABATER, J. *et al.* ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. **Earth System Science Data**, 2021. v. 13, n. 9, p. 4349–4383.

NAITAM, A.; MEGHANA, N.; SRIVASTAV, R. Multimodel classification and regression technique for the statistical downscaling of temperature. **Stochastic Environmental Research and Risk Assessment**, 2023.

NARAYANAN, V. A.; PRABHU, K. M. M. The fractional Fourier transform: Theory, implementation and error analysis. **Microprocessors and Microsystems**, 2003. v. 27, n. 10, p. 511–521.

NASCIMENTO, R. L. Energia solar no Brasil: situação e perspectivas. **Estudo Técnico**, 2017. p. 1–46. Disponível em: <[file:///C:/Users/gmols/Downloads/energia_solar_limp \(2\).pdf](file:///C:/Users/gmols/Downloads/energia_solar_limp%20(2).pdf)>.

NCEP. Global Forecast System (GFS). **National Centers for Environmental Prediction**, 2023. Disponível em: <<https://www.nco.ncep.noaa.gov/pmb/products/gfs/%3E>>.

OSMAN, Y. Z.; ABDELLATIF, M. E. Improving accuracy of downscaling rainfall by combining predictions of different statistical downscale models. **Water Science**, 2016. v. 30, n. 2, p. 61–75.

PAHLAVAN, H. A. *et al.* Improvement of multiple linear regression method for statistical downscaling of monthly precipitation. **International Journal of Environmental Science and Technology**, 2018. v. 15, n. 9, p. 1897–1912.

PEDROSA, G. Detecção e Diagnóstico de Falhas na Performance de Aerogeradores. 2016. p. 67. Disponível em: <<https://repositorio.ufpe.br/handle/123456789/22218>>.

PENG, X. *et al.* A short-term wind power prediction method based on deep learning

and multistage ensemble algorithm. **Wind Energy**, 2022. v. 25, n. 9, p. 1610–1625.

PÉREZ, J. J. M.; NAVARRO, M. E. J.; CANTOS, J. O. Reconstruction of thermal local signal from statistical downscaling (sd) through artificial neuronal network: detection of local patterns of change in Valencia Region (Spain). [S.l.]: [s.n.], 2016. p. 471–475. Disponível em: <<https://core.ac.uk/download/pdf/32327349.pdf>>.

PERRUCCI, V. P. **Análise da complementaridade entre diferentes técnicas estatísticas para melhoria na resolução espacial do comportamento do vento local**. [S.l.]: Universidade Federal de Pernambuco (UFPE), 2018.

PETTRIBÚ, L.; SOUZA, J. **Avaliação da qualidade de dados observacionais de vento e radiação solar**. [S.l.]: Universidade Federal de Pernambuco (UFPE), 2018.

PIERCE, D. W.; CAYAN, D. R. Downscaling humidity with Localized Constructed Analogs (LOCA) over the conterminous United States. **Climate Dynamics**, 2016. v. 47, n. 1–2, p. 411–431.

PIERRO, M. *et al.* Deterministic and Stochastic Approaches for Day-Ahead Solar Power Forecasting. **Journal of Solar Energy Engineering, Transactions of the ASME**, 2017. v. 139, n. 2.

PRYOR, S. C.; SCHOOF, J. T.; BARTHELMIE, R. J. Climate change impacts on wind speeds and wind energy density in Northern Europe: Empirical downscaling of multiple AOGCMs. **Climate Research**, 2005. v. 29, n. 3, p. 183–198.

RACHMAWATI, R. N.; SUNGKAWA, I.; RAHAYU, A. Extreme rainfall prediction using Bayesian quantile regression in statistical downscaling modeling. **Procedia Computer Science**, 2019. v. 157, p. 406–413.

RANSTAM, J.; COOK, J. A. LASSO regression. **British Journal of Surgery**, 2018. v. 105, n. 10, p. 1348.

REICHENBERG, L.; SIDDIQUI, A. S.; WOGGIN, S. Policy implications of downscaling the time dimension in power system planning models to represent variability in renewable output. **Energy**, 2018. v. 159, p. 870–877.

REN 1059/2023. Resolução Normativa ANEEL 1.059/2023. [s.d.]. Disponível em: <<https://www2.aneel.gov.br/cedoc/ren20231059.pdf>>.

RIBEIRO, M. H. D. M. *et al.* Efficient bootstrap stacking ensemble learning model applied to wind power generation forecasting. **International Journal of Electrical Power and Energy Systems**, 2022. v. 136.

RICHARDSON, C. W. Stochastic simulation of daily precipitation, temperature, and solar radiation. **Water Resources Research**, 1981. v. 17, n. 1, p. 182–190.

RIVINGTON, M. *et al.* Downscaling regional climate model estimates of daily precipitation, temperature and solar radiation data. **Climate Research**, 2008. v. 35, n. 3, p. 181–202.

ROUX, R. LE *et al.* Comparison of statistical and dynamical downscaling results from

the WRF model. **Environmental Modelling and Software**, 2018. v. 100, p. 67–73.

ROZAS-LARRAONDO, P.; INZA, I.; LOZANO, J. A. A method for wind speed forecasting in airports based on nonparametric regression. **Weather and Forecasting**, 2014. v. 29, n. 6, p. 1332–1342.

SALCEDO-SANZ, S. *et al.* A CRO-species optimization scheme for robust global solar radiation statistical downscaling. **Renewable Energy**, 2017. v. 111, p. 63–76.

SÁNCHEZ, I. Adaptive combination of forecasts with application to wind energy. **International Journal of Forecasting**, 2008. v. 24, n. 4, p. 679–693.

SCHULTZ, M. G. *et al.* Can deep learning beat numerical weather prediction? **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, 2021. v. 379, n. 2194.

SEMENOV, M. A.; BARROW, E. M. Use of a stochastic weather generator in the development of climate change scenarios. **Climatic Change**, 1997. v. 35, n. 4, p. 397–414.

SIFUZZAMAN, M.; ISLAM, M. RAFIQ; ALI, M. Z. **Application of wavelet transform and its advantages compared to Fourier transform**. International Journal and Magazine of Engineering, Technology, Management and Research. Disponível em: <<http://www.ijmetmr.com/olfebruary2016/BasimNasih-A-4.pdf>. Acesso em: 17 fev. 2023>.

SILVA, G. R. **Características de Vento da Região Nordeste: Análise, Modelagem e Aplicações para Projetos de Centrais Eólicas**. [S.I.]: Universidade Federal de Pernambuco (UFPE), 2003.

SSEKULIMA, E. B. *et al.* Wind speed and solar irradiance forecasting techniques for enhanced renewable energy integration with the grid: A review. **IET Renewable Power Generation**, 2016. v. 10, n. 7, p. 885–898.

STORCH, H. VON. Review of empirical downscaling techniques. [S.I.]: [s.n.], 2000.

SUMNER, J.; WATTERS, C. S.; MASSON, C. CFD in wind energy: The virtual, multiscale wind tunnel. **Energies**, 2010. v. 3, n. 5, p. 989–1013.

TANG, B. H.; BASSILL, N. P. Point downscaling of surface wind speed for forecast applications. **Journal of Applied Meteorology and Climatology**, 2018. v. 57, n. 3, p. 659–674.

TANG, J.; ALELYANI, S.; LIU, Huan. **Feature selection for classification: A review**. [S.I.]: Data classification: Algorithms and applications, 2014.

TAYLOR, K. E. Summarizing multiple aspects of model performance in a single diagram. **Journal of Geophysical Research Atmospheres**, 2001. v. 106, n. D7, p. 7183–7192.

TIBSHIRANI, R. Regression Shrinkage and Selection Via the Lasso. **Journal of the Royal Statistical Society: Series B (Methodological)**, 1996. v. 58, n. 1, p. 267–

288.

TIMBAL, B.; DUFOUR, A; MCAVANEY, B. An estimate of future climate change for western France using a statistical downscaling technique. **Climate Dynamics**, 2003. v. 20, p. 807–823. Disponível em: <<http://www.springerlink.com/content/940EKRKKEAAXFGR5>>.

TIMMERMANN, A. Chapter 4 Forecast Combinations. **Handbook of Economic Forecasting**, 2006. v. 1, p. 135–196.

TIPPETT, M. K.; DELSOLE, T. Constructed analogs and linear regression. **Monthly Weather Review**, 2013. v. 141, n. 7, p. 2519–2525.

TORRENCE, C.; COMPO, G. P. A Practical Guide to Wavelet Analysis. **Bulletin of the American Meteorological Society**, 1998. v. 79, n. 1, p. 61–78.

TROCCOLI, A.; MORCRETTE, Jean Jacques. Skill of direct solar radiation predicted by the ECMWF global atmospheric model over Australia. **Journal of Applied Meteorology and Climatology**, 2014. v. 53, n. 11, p. 2571–2588.

VACCARO, A. *et al.* An adaptive framework based on multi-model data fusion for one-day-ahead wind power forecasting. **Electric Power Systems Research**, mar. 2011. v. 81, n. 3, p. 775–782. Disponível em: <<https://linkinghub.elsevier.com/retrieve/pii/S0378779610002816>>.

VELEDA, D.; MONTAGNE, R.; ARAUJO, M. Cross-wavelet bias corrected by normalizing scales. **Journal of Atmospheric and Oceanic Technology**, 2012. v. 29, n. 9, p. 1401–1408.

VRAC, M.; STEIN, M.; HAYHOE, K. Statistical downscaling of precipitation through nonhomogeneous stochastic weather typing. **Climate Research**, 2007. v. 34, n. 3, p. 169–184.

WALTON, D. B. *et al.* A hybrid dynamical-statistical downscaling technique. Part I: Development and validation of the technique. **Journal of Climate**, 2015. v. 28, n. 12, p. 4597–4617.

WANG, H. Z. *et al.* Deep belief network based deterministic and probabilistic wind speed forecasting approach. **Applied Energy**, 2016. v. 182, p. 80–93.

WANG, Jianzhou; HU, J.; MA, K. Wind speed probability distribution estimation and wind energy assessment. **Renewable and Sustainable Energy Reviews**, 2016. v. 60, p. 881–899.

WANG, Jujie *et al.* An optimized complementary prediction method based on data feature extraction for wind speed forecasting. **Sustainable Energy Technologies and Assessments**, 2022. v. 52.

WANG, Ning *et al.* Wavelet Compression Technique for High-Resolution Global Model Data on an Icosahedral Grid. **Journal of Atmospheric and Oceanic Technology**, set. 2015. v. 32, n. 9, p. 1650–1667. Disponível em: <https://journals.ametsoc.org/view/journals/atot/32/9/jtech-d-14-00217_1.xml>.

WANG, Y. *et al.* A review of wind speed and wind power forecasting with deep neural networks. **Applied Energy**, 2021. v. 304.

WARD, J. H. Hierarchical Grouping to Optimize an Objective Function. **Journal of the American Statistical Association**, 1963. v. 58, n. 301, p. 236–244.

WELCH, P. D. The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms. **IEEE Transactions on Audio and Electroacoustics**, 1967. v. 15, n. 2, p. 70–73.

WETTERHALL, F.; HALLDIN, S.; XU, C. Y. Statistical precipitation downscaling in central Sweden with the analogue method. **Journal of Hydrology**, 2005. v. 306, n. 1–4, p. 174–190.

WILBY, R. *et al.* **Guidelines for Use of Climate Scenarios Developed from Statistical Downscaling Methods**. [S.I.]: IPCC Task Group on Data and Scenario Support for Impact and Climate Analysis (TGICA), 2004.

WILBY, R. L. *et al.* Statistical downscaling of general circulation model output: A comparison of methods. **Water Resources Research**, 1998. v. 34, n. 11, p. 2995–3008.

WILBY, R.; WIGLEY, T. Downscaling general circulation model output: A review of methods and limitations. **Progress in Physical Geography**, 1997. v. 21, n. 4, p. 530–548.

WILKS, D. S. **Statistical methods in the atmospheric sciences**. [S.I.]: Academic press, 2011.

WINSTRAL, A.; JONAS, T.; HELBIG, N. Statistical downscaling of gridded wind speed data using local topography. **Journal of Hydrometeorology**, 2017. v. 18, n. 2, p. 335–348.

WOOLDRIDGE, J. M. **Introductory econometrics: A modern approach**. [S.I.]: Cengage learning, 2015.

WÜTHRICH, K.; ZHU, Ying. Omitted Variable Bias of Lasso-Based Inference Methods: A Finite Sample Analysis. **Review of Economics and Statistics**, 2023. p. 1–16.

YANG, C.; WANG, Ninglian; WANG, Shijin. A comparison of three predictor selection methods for statistical downscaling. **International Journal of Climatology**, 2017. v. 37, n. 3, p. 1238–1249.

YANG, Y. COMBINING FORECASTING PROCEDURES: SOME THEORETICAL RESULTS. **Econometric Theory**, 5 fev. 2004. v. 20, n. 01. Disponível em: <http://www.journals.cambridge.org/abstract_S0266466604201086>.

YU, L.; LIU, Huan. Feature selection for high-dimensional data: A fast correlation-based filter solution. [S.I.]: [s.n.], 2003. p. 856–863.

ZHANG, Gang *et al.* Solar forecasting with hourly updated numerical weather

prediction. **Renewable and Sustainable Energy Reviews**, 2022. v. 154.

ZHANG, Xinmin *et al.* A Solar Time Based Analog Ensemble Method for Regional Solar Power Forecasting. **IEEE Transactions on Sustainable Energy**, 2019. v. 10, n. 1, p. 268–279.

ZHANG, Yu; LI, Yanting; ZHANG, Guangyao. Short-term wind power forecasting approach based on Seq2Seq model using NWP data. **Energy**, 2020. v. 213.

ZHANG, Yuzhen; LIU, J.; SHEN, W. A Review of Ensemble Learning Algorithms Used in Remote Sensing Applications. **Applied Sciences (Switzerland)**, 2022. v. 12, n. 17.

ZHOU, W. *et al.* Current status of research on optimum sizing of stand-alone hybrid solar-wind power generation systems. **Applied Energy**, 2010. v. 87, n. 2, p. 380–389. Disponível em: <<http://dx.doi.org/10.1016/j.apenergy.2009.08.012>>.

ZIEGEL, E. R.; ENDERS, W. **Applied Econometric Time Series**. [S.l.]: [s.n.], 1995. V. 37.

ZORITA, E. *et al.* Stochastic characterization of regional circulation patterns for climate model diagnosis and estimation of local precipitation. **Journal of Climate**, 1995. v. 8, n. 5, p. 1023–1042.

ZORITA, Eduardo; STORCH, Hans VON. The analog method as a simple statistical downscaling technique: Comparison with more complicated methods. **Journal of Climate**, 1999. v. 12, n. 8 PART 2, p. 2474–2489.