

**UNIVERSIDADE FEDERAL DE PERNAMBUCO
DEPARTAMENTO DE ENERGIA NUCLEAR**

**COMISSÃO NACIONAL DE ENERGIA NUCLEAR
CENTRO REGIONAL DE CIÊNCIAS NUCLEARES DO NORDESTE**

Programa de Pós-Graduação em Tecnologias Energéticas e Nucleares

VALENTIN PASCHOAL PERRUCI

**ANÁLISE DE COMPLEMENTARIDADE ENTRE
DIFERENTES TÉCNICAS ESTATÍSTICAS PARA AUMENTO
NA RESOLUÇÃO ESPACIAL DO COMPORTAMENTO DO
VENTO LOCAL**

Orientador: Profa. Dra. Olga de Castro Vilela

Coorientador: Prof. Dr. Alexandre Costa

Recife

2018

VALENTIN PASCHOAL PERRUCI

**ANÁLISE DE COMPLEMENTARIDADE ENTRE
DIFERENTES TÉCNICAS ESTATÍSTICAS PARA AUMENTO
NA RESOLUÇÃO ESPACIAL DO COMPORTAMENTO DO
VENTO LOCAL**

Dissertação submetida ao Programa de Pós-Graduação em Tecnologias Energéticas e Nucleares para obtenção do título de Mestre em Ciências, Área de Concentração: Fontes Renováveis de Energia.

Orientador: Profa. Dra. Olga de Castro Vilela

Coorientador: Prof. Dr. Alexandre Costa

Recife

2018

Catálogo na fonte
Bibliotecário Carlos Moura, CRB-4 / 1502

P461a

Perruci, Valentin Paschoal.

Análise da complementaridade entre diferentes técnicas estatísticas para aumento na resolução espacial do comportamento do vento local. / Valentin Paschoal Perruci. - Recife: O Autor, 2018.

98 f. : il., tabs.

Orientadora: Profa. Dra. Olga de Castro Vilela.

Coorientador: Prof. Dr. Alexandre Carlos Araújo da Costa.

Dissertação (mestrado) – Universidade Federal de Pernambuco. CTG. Programa de Pós-Graduação em Tecnologias Energéticas e Nucleares, 2018.

Inclui referências e apêndices.

1. Energia eólica. 2. Região Nordeste do Brasil. 3. Downscaling estatístico. 4. Combinação de modelos. 5. Análise espectral. I. Vilela, Olga de Castro, orientadora. II. Costa, Alexandre Carlos Araújo da, coorientador. III. Título.

CDD 621.45 (22. ed.)

UFPE
BDEN/2018-18

VALENTIN PASCHOAL PERRUCI

**ANÁLISE DE COMPLEMENTARIDADE ENTRE
DIFERENTES TÉCNICAS ESTATÍSTICAS PARA AUMENTO
NA RESOLUÇÃO ESPACIAL DO COMPORTAMENTO DO
VENTO LOCAL**

Dissertação apresentada ao Programa de Pós-Graduação em Tecnologias Energéticas e Nucleares da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciências, Área de Concentração: Fontes Renováveis de Energia.

Aprovada em: 27/02/2018

BANCA EXAMINADORA:

Prof. Dr. André Maciel Netto

Profa. Dra. Doris Regina Aires Veleda

Prof. Dr. Tsang Ing Ren

AGRADECIMENTOS

Agradeço primeiramente a Deus pela vida como um todo, o que inclui toda a trajetória percorrida em associação com pessoas admiráveis, que participam direta ou indiretamente de cada momento do dia a dia;

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 da Graduação e/ou Mestrado: ao CNPq pelo apoio financeiro no âmbito do projeto EOLIPREV (Processo 406984/2013-1), à RNP pelo apoio no âmbito do projeto HPC4E (<https://hpc4e.eu/the-project/work-plan/wp4>) e ao CTGAS-ER/PETROBRAS pelo apoio no âmbito do projeto Aprimoramento de Aerogerador de 2 MW para Adequação às Condições Eólicas Brasileiras (Programa de P&D Tecnológico do Setor de Energia Elétrica da ANEEL); 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 notável crescimento da energia eólica no Brasil aumenta a demanda tanto por ferramentas mais sofisticadas para a avaliação do recurso eólico como ferramentas para o controle do sistema elétrico nacional em vista do aumento na participação de uma fonte cuja disponibilidade é intermitente. Nesse contexto, o presente trabalho dedica-se a uma ferramenta que bem pode atender às duas aplicações citadas em função de ajustes específicos. Por exemplo, dados de previsão numérica do tempo (empregados como entrada à ferramenta proposta) são de interesse para o despacho da produção de centrais eólicas e controle do sistema elétrico. Por sua vez, dados climáticos de reanálise (empregados como entrada à ferramenta proposta) são de interesse para a avaliação do recurso eólico com vistas aos estudos de viabilidade sobre a implantação de centrais eólicas. Particularmente, este trabalho avalia diversas técnicas estatísticas tipicamente adotadas em estratégias de aumento de resolução espacial (*downscaling*) e combinação de modelos com vistas à modelagem do vento próximo aos locais de interesse. Para isso, dados provenientes de 13 estações anemométricas distribuídas na Região Nordeste do Brasil (região que concentra a maior parte da capacidade eólica instalada no país) são empregados juntamente com simulações (reanálises) geradas por modelos de circulação geral da atmosfera (GCMs). Além disso, nesse trabalho, uma análise espectral é proposta como artifício auxiliar em estratégias de combinação para discriminar as aptidões/limitações de cada modelo avaliado em diferentes faixas de frequência (ou escalas de variação temporal). Para tanto, empregaram-se técnicas baseadas no ajuste de modelos individuais tais como a regressão multilinear, redes neurais, análogos e agrupamento de padrões sinóticos para, em seguida, combinar as saídas de tais modelos por meio de técnicas de regressão. De forma geral, os resultados mostram que as técnicas empregadas de *downscaling* estatístico e combinação são capazes de reproduzir satisfatoriamente o comportamento em baixas e médias frequências dos sinais observados pelas estações anemométricas, tendo obtido melhorias significativas em relação a abordagens mais simples (e.g., interpolação). Os resultados também mostram que, em geral, a combinação de saídas produziu estimativas mais confiáveis que as saídas individuais dos modelos. Entretanto, ainda há limitações a serem superadas quanto à reprodução do comportamento em altas frequências dos sinais observados pelas estações anemométricas. Por fim, resultados promissores obtidos com uma análise semiobjetiva (semiautomática) aqui proposta indicam que é possível aprimorar, através da definição apropriada de “complementaridade espectral”, procedimentos de combinação com respeito à confiabilidade das estimativas e ao custo computacional associado ao ajuste dos modelos.

Palavras-chave: Energia Eólica. Região Nordeste do Brasil. Downscaling Estatístico. Combinação de Modelos. Análise Espectral.

ABSTRACT

The remarkable growth of wind energy in Brazil increases the demand for both more sophisticated tools for the evaluation of the wind resource and tools for the control of the national electric system in view of the increase in the participation of a source whose availability is intermittent. In this context, the present work devotes itself to a tool that can well serve the two applications mentioned in function of specific adjustments. For example, numerical weather forecast data (employed as input to the proposed tool) are of interest for the dispatch of wind power production and control of the electrical system. On the other hand, climatic data of reanalysis (used as input to the proposed tool) are of interest for the evaluation of the wind resource with a view to feasibility studies on the implantation of wind farms. In particular, this work evaluates several statistical techniques typically adopted in strategies of increase of spatial resolution (downscaling) and combination of models with a view to modeling the wind near the places of interest. For this, data from 13 anemometric stations distributed in the Northeast Region of Brazil (region that concentrates most of the wind capacity installed in the country) are used together with simulations (reanalysis) generated by general circulation models of the atmosphere (GCMs). In addition, in this work, a spectral analysis is proposed as an auxiliary artifice in combination strategies to discriminate the aptitudes / limitations of each model evaluated in different frequency bands (or time variation scales). To that end, techniques based on the adjustment of individual models such as multilinear regression, neural networks, analogs and synoptic pattern grouping were used to then combine the outputs of such models by means of regression techniques. In general, the results show that the techniques used for statistical downscaling and combination are able to reproduce satisfactorily the behavior at low and medium frequencies of the signals observed by the anemometric stations, having obtained significant improvements in relation to simpler approaches (eg, interpolation). The results also show that, in general, the combination of outputs produced more reliable estimates than the individual outputs of the models. However, there are still limitations to be overcome regarding the reproduction of the behavior at high frequencies of the signals observed by anemometric stations. Finally, promising results obtained with a semi-automatic (semiautomatic) analysis proposed here suggest that it is possible to improve, through the appropriate definition of "spectral complementarity", combination procedures with respect to the reliability of the estimates and the computational cost associated with the adjustment of the models.

Keywords: Wind Energy. Northeastern Brazil. Statistical Downscaling. Combination of Models. Spectral Analysis.

LISTA DE ILUSTRAÇÕES

Figura 2.1. Ilustração da Disposição Espacial Típica de GCMs.	15
Figura 2.2. Arquiteturas de Parâmetros de Modelos Estatísticos	20
Figura 4.1. Ilustração da Estratégia de Ação.	28
Figura 4.2. Ilustração da metodologia empregada.....	29
Figura 4.3. Interpolação Bilinear.	32
Figura 4.4. Ilustração de Rede Neural tipo Percéptron de Múltiplas Camadas.....	41
Figura 4.5. Wavelet Analítica de Morlet Adimensional (ψ_0).....	49
Figura 4.6. Resolução de Tempo-Frequência.....	50
Figura 5.1. Localização de Estações Anemométricas e de Centrais Eólicas no Brasil.	55
Figura 5.2. Caso 1: Diagrama de Taylor (escala horária).....	61
Figura 5.3. Caso 1: Diagramas de Dispersão e Histograma de Erros (escala horária).....	64
Figura 5.4. Caso 1: Superposição de Séries Temporais (escala horária).....	65
Figura 5.5. Caso 1: Diagrama de Taylor (escala diária).....	66
Figura 5.6. Caso 1: Diagramas de Dispersão e Histograma de Erros (escala diária).	68
Figura 5.7. Caso 1: Superposição de Séries Temporais (escala diária).....	69
Figura 5.8. Caso 1: Erros Espectrais Absoluto e Relativo.	71
Figura 5.9. Caso 2: Diagrama de Taylor (escala horária).....	73
Figura 5.10. Caso 2: Diagrama de Taylor (escala diária).....	76
Figura 5.11. Caso 2: Erros Espectrais Absoluto e Relativo.	78
Figura A.1. Caso 2: Diagramas de Dispersão e Histograma de Erros (escala horária).....	92
Figura A.2. Caso 2: Superposição de Séries Temporais (escala horária).....	93
Figura A.3. Caso 2: Diagramas de Dispersão e Histograma de Erros (escala diária).	94
Figura A.4. Caso 2: Superposição de Séries Temporais (escala diária).....	94

LISTA DE TABELAS

Tabela 4.1. Configuração de domínio adotada para cada modelo individual.....	30
Tabela 4.2. Configuração geral de técnicas de agrupamento.	39
Tabela 5.1. Informações sobre Estações Anemométricas Ilustradas na Figura 5.1.	57
Tabela 5.2. Caso 1: configurações resultantes para CL1, CL2 e CL3.....	63
Tabela 5.3. Caso 2: configurações resultantes para CL1, CL2 e CL3.....	75
Tabela B.1. Melhores Resultados em Todas as Estações Consideradas.	96

SUMÁRIO

1 INTRODUÇÃO	11
2 CONCEITOS PRELIMINARES.....	14
2.1 Modelos de Circulação Geral Atmosférica e Dados de Reanálise.....	14
2.2 Aumento da Resolução (<i>Downscaling</i>) do Vento Local.....	16
2.2.1 Abordagem Dinâmica.....	17
2.2.2 Abordagem Empírica.....	18
2.2.3 Modelos Estatísticos: Arquiteturas de Parâmetros	19
3 REVISÃO BILIOGRÁFICA	22
4 METODOLOGIA E MODELOS	28
4.1 Técnicas Empregadas em Estratégias de <i>Downscaling</i> Estatístico	31
4.1.1 Interpolação Bilinear (IBL)	31
4.1.2 Regressão Multilinear (MLR)	33
4.1.3 Análise de Componentes Principais (PCA).....	34
4.1.4 Análogos (AN)	35
4.1.5 Agrupamento de Padrões Sinóticos (CL)	37
4.1.6 Redes Neurais (NN).....	40
4.2 Combinação de Modelos	41
4.3 Análise no Domínio da Frequência	42
4.3.1 Transformada de Fourier (TF).....	44
4.3.2 Transformada de Fourier de Curto-Prazo (TFCP).....	46
4.3.3 Transformada <i>Wavelet</i> Contínua (TWC).....	48
4.3.4 Complementaridade Espectral.....	52
5 RESULTADOS E DISCUSSÃO	55
5.1 Estudo de Caso 1: Melhores Performances Globais	59
5.2 Estudo de Caso 2: Melhores Performances Relativas à Interpolação Bilinear	72
6 CONCLUSÕES E PERSPECTIVAS	79
REFERÊNCIAS	81

APÊNDICE A – CASO 2: RESULTADO DO MODELO CBLB	92
APÊNDICE B – ESTUDOS DE CASO: RESULTADOS GERAIS.....	96

1 INTRODUÇÃO

Diversas políticas de incentivo às energias renováveis foram criadas nos últimos anos visando a diversificação da matriz elétrica nacional (e.g., Resoluções Normativas ANEEL nº 482/2012 e 687/2015). Impulsionado por tais medidas, o setor eólico tem observado rápida expansão, o que tem estimulado a sofisticação das técnicas empregadas para modelar o vento nos locais de interesse, tendo em vista que as incertezas associadas ao comportamento de tal fonte impactam diretamente na operação de centrais eólicas (e.g., previsão de despacho de potência na rede elétrica), bem como em estudos de viabilidade técnico-econômica para a instalação de novas centrais (BATHURST et al., 2002; BARTHELMIE et al., 2008; BADGER et al., 2014).

A modelagem do vento local frequentemente recorre ao emprego de dados provenientes de modelos gerais da circulação atmosférica (*general circulation models*, GCMs – KALNAY et al., 1996), os quais simulam o comportamento de diversas variáveis atmosféricas para todo o globo terrestre e com ampla cobertura temporal (dezenas de anos). Entretanto, as maiores limitações dos GCMs se dão com respeito à baixa capacidade que eles apresentam para descrever o comportamento do vento na escala espacial própria de uma central eólica (ordem de centenas de metros), devido à insuficiente resolução horizontal disponível (ordem de centenas de quilômetros). Há diversos casos em que são necessárias informações sobre os fenômenos tipicamente associados com a escala local de resolução espacial, visto que, em tais escalas o comportamento do vento é sensível à orografia, vegetação, obstáculos etc.

A fim de superar as barreiras impostas por essas limitações, costumam-se empregar diversas técnicas de aumento de resolução conhecidas como técnicas de *downscaling*, técnicas estas que relacionam variáveis atmosféricas nas escalas sinóptica e regional (variáveis preditoras) com variáveis de interesse na escala local (preditandos) (WILBY e WIGLEY, 1997). A natureza das técnicas de *downscaling* pode ser numérica (*downscaling* dinâmico), empírica (*downscaling* estatístico) ou híbrida. As diversas aplicações propostas deram origem a técnicas que possuem diferentes aptidões, bem como limitações.

Com o crescente número de proposições para modelos aplicados em estratégias de *downscaling*, é natural que maiores esforços sejam necessários para se determinar que modelos melhor se adequam em cada caso. Alguns modelos podem apresentar maior habilidade na reprodução de características estatísticas específicas da série temporal de referência (e.g., variância reproduzida, correlação linear etc.) e nem sempre eles são melhores em termos gerais,

ou seja, podem apresentar um erro geral de ajuste superior aos demais. Isso pode levar a situações em que tais modelos sejam considerados ineficazes e seus resultados sejam subutilizados. Apesar do grande número de estratégias de combinação já propostos, há um número relativamente baixo de publicações sobre combinação de saídas de modelos de *downscaling* estatístico. Em todo caso, sabe-se que a introdução de modelos inadequados para a combinação pode resultar em uma performance inferior às individuais (TIMMERMAN, 2006; SÁNCHEZ, 2008)

Com respeito à modelagem do vento local baseada em *downscaling* estatístico, não há, até o presente momento, publicações em periódicos de alto índice de impacto sobre métodos que permitam avaliar a complementaridade entre saídas de modelos microescalares, derivados das estratégias de *downscaling* estatístico, especialmente no que se refere às suas estruturas de fase (deslocamento temporal de eventos periódicos) e frequência (distribuição da variância do sinal em faixas de frequência). Nesse âmbito, é comum o emprego de métodos de seleção de variáveis baseados no fator de correlação (e.g., YU e LIU 2003; KANG et al., 2008). Porém, este estatístico não é apropriado para discriminar informações relativas às frequências do sinal (diferentes escalas de variação temporal) uma vez que infere características globais entre séries temporais (TAYLOR, 2001). Nesse contexto, métodos locais podem ser propostos para destacar, com maior nível de detalhes, a disposição geral entre um sistema formado por saídas de modelos distintos, tendo como referência um comportamento alvo (observações).

A metodologia apresentada neste estudo tem como objetivo fundamental o de dar suporte ao aprimoramento dos procedimentos usuais para a avaliação do comportamento da velocidade do vento local em pontos localizados na Região Nordeste do Brasil, que conta com 80% da capacidade instalada da energia eólica brasileira (ABEEólica, 2017). Para tal, são consideradas 13 estações anemométricas distribuídas na região, juntamente com dados de simulação atmosférica de reanálise provenientes do programa ERA-Interim (<https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-interim>). O presente trabalho avalia e expande o conhecimento sobre técnicas para o *downscaling* estatístico tipicamente adotadas para estimar o recurso eólico. Dessa forma, espera-se que o estudo sirva de base para a superação das barreiras existentes com respeito às práticas usuais voltadas para a energia eólica (e.g., estimação de recurso com baixa acurácia), contribuindo tanto com estudos de viabilidade técnico-econômica (os quais normalmente empregam dados climáticos de reanálise), quanto com atividades operacionais em centrais eólicas (os quais normalmente empregam dados de previsão numérica). Em mais detalhes, apresenta-se aqui uma metodologia que permite avaliar o desempenho de modelos diversos, bem como de combinações das saídas,

as quais estão voltadas para a obtenção de estimativas mais confiáveis. Além disso, também se propõe um método para a análise da complementaridade entre as saídas, ou, em outras palavras, do potencial aproveitamento das aptidões únicas de modelos avaliados.

Mais especificamente, será proposta uma análise semiobjetiva (semiautomática) para o estudo da complementaridade entre diferentes saídas de modelos de *downscaling* estatístico para a descrição do vento local em termos das distribuições de energia espectral dos sinais gerados contrastados com a do sinal observado. Em outras palavras, a técnica irá servir como indicador da compatibilidade (avaliada no domínio da frequência) entre as saídas individuais pretendidas para reproduzir o sinal observado. A resultante de tal análise consiste em um maior detalhamento da habilidade demonstrada nas saídas de modelos por meio de informações extraídas de estimativas de densidade de energia espectral de séries temporais de magnitude do vento (observadas e estimadas). Com isso, é possível restringir o número total de combinações entre modelos por meio de um critério preestabelecido, o que potencialmente reduz o custo computacional demandado por técnicas de otimização do sistema de variáveis para combinação (e.g., seleção objetiva de variáveis regressoras). Além disso, estudos futuros podem ser propostos para a elaboração de algoritmos de treinamento baseados na complementaridade espectral entre os preditores, permitindo que mesmo aspectos sutis (ocultos em representação temporal) sejam considerados, o que potencialmente trará maior confiabilidade às estimativas de técnicas de *downscaling* estatístico e combinação.

Ao todo, o trabalho está dividido em seis seções principais: 1) Introdução; 2) Conceitos Preliminares, em que são introduzidos conceitos básicos relacionados com a aplicação de técnicas de *downscaling* para descrever comportamentos locais; 3) Revisão Bibliográfica, na qual são exploradas as abordagens mais relevantes e atuais relacionadas com o estudo proposto nesta dissertação; 4) Metodologia e Modelos; 5) Resultados e Discussão, na qual são apresentados e discutidos dois estudos de caso notáveis; 6) Conclusões e Perspectivas. No final do trabalho, encontra-se uma lista de referências, seguida pelos Anexos A e B. O Anexo A apresenta figuras relacionadas com a discussão apresentada na seção de resultados. Na sequência, o Anexo B apresenta os resultados gerais e uma breve discussão sobre todos os estudos de caso realizados.

2 CONCEITOS PRELIMINARES

Esta seção apresenta conceitos fundamentais para o entendimento do contexto geral deste trabalho.

A subseção 2.1 trata dos principais modelos de circulação geral da atmosfera (GCMs), os quais são desenvolvidos por renomados centros de previsão meteorológica como o ECMWF (*“European Centre for Medium-Range Weather Forecasts”*). As saídas dos GCMs têm sido amplamente empregadas em estudos relacionados com a modelagem de grandezas atmosféricas.

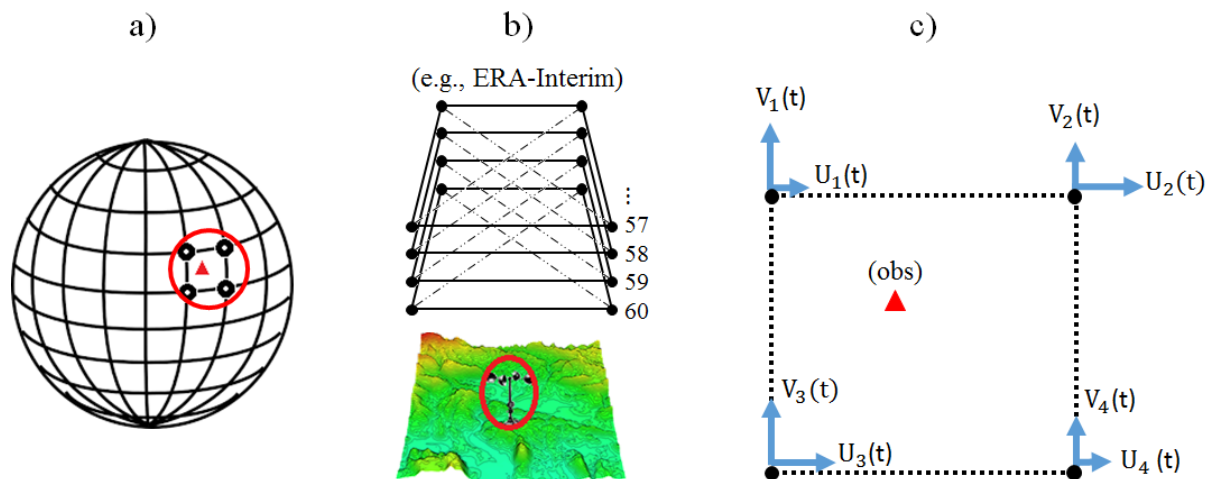
A subseção 2.2 aborda as técnicas de aumento de resolução espacial (*“downscaling”*) empregadas para acoplar as escalas sinópticas (variáveis provenientes dos GCMs) e local (variáveis observadas). A aplicação destas técnicas são várias, alguns exemplos são: previsão de curto prazo (e.g., COSTA et al., 2008), estudos sobre efeitos de mudanças climáticas e caracterização do clima local com vistas à avaliação de recurso disponível, esta última se refere principalmente aos casos de estudos de viabilidade de geração de energia elétrica a partir das fontes renováveis como a eólica, solar ou hídrica. As diferentes limitações que surgem nas diversas aplicações de técnicas de *downscaling* estimulam a proposição de modelos com diferentes aptidões, uma vez que são concebidos para reproduzir características específicas (e.g., valor médio e variabilidade) das variáveis de interesse, a depender da aplicação (WILBY et al., 2004), o que impulsiona a busca por maneiras de combinar as diferentes saídas a fim de tirar maior proveito dos resultados individuais.

2.1 Modelos de Circulação Geral Atmosférica e Dados de Reanálise

Os modelos de circulação geral da atmosfera são modelos numéricos que resolvem as equações de conservação de massa, energia e momento de várias grandezas atmosféricas, como o campo de velocidades de vento, altura geopotencial, temperatura e pressão superficial, dentre outras variáveis. As saídas desses modelos são disponibilizadas com ampla cobertura temporal e espacial, contemplando dezenas de anos de simulações para todo o globo terrestre. Por esse motivo, tais saídas servem como aproximações preliminares para o estado atmosférico. Entretanto, esses modelos não costumam apresentar resoluções espaciais satisfatórias para serem aplicados diretamente como aproximações para o comportamento, em escala local, de variáveis como o vento, que possuem uma componente estocástica bastante expressiva.

A Figura 2.1 ilustra a disposição dos GCMs sob a perspectiva global e local. Esses modelos possuem cobertura horizontal de simulação bastante ampla, cobrindo em todo o globo terrestre, porém, com baixa resolução espacial para aplicações regionais e locais (áreas circuladas nas Figuras 2.1a e 2.1b, respectivamente). Portanto, costuma-se empregar malhas horizontais com um determinado número de pontos próximos ao local de interesse (e.g., os quatro pontos mais próximos). Normalmente, os estados atmosféricos são simulados para vários níveis de modelo, que compreendem desde níveis mais próximos à superfície terrestre até as camadas mais altas da atmosfera (e.g., 60 níveis de modelo do programa ERA-Interim; Fig. 2.1b). A Figura 2.1c apresenta a vista superior de uma malha horizontal, contendo os 4 pontos mais próximos ao local de interesse (ainda com referência à Figura 2.1b). Considere que, para cada ponto de simulação estão disponíveis diversas variáveis atmosféricas geradas por um determinado GCM, como: pressão, temperatura, umidade relativa do ar, radiação solar, etc. O estado atmosférico (ou padrão sinóptico) é definido, em um instante de tempo t , pelo conjunto de todas as variáveis consideradas na formação do sistema estudado. A Figura 2.1c ilustra um padrão sinóptico formado pelas componentes zonal (“U”) e meridional (“V”) da velocidade do vento nos pontos de malha considerados. Note que a convenção adotada no ERA-Interim para a definição do campo de velocidades de vento diz que “a componente U é positiva para um fluxo oeste-leste (vento leste) e a componente V é positivo para o fluxo sul-norte (vento norte)” (ECMWF, 2018). Após ajustes apropriados, costuma-se relacionar tais componentes a dados anemométricos, os quais normalmente estão disponíveis em sistemas de projeção polares (magnitude e direção).

Figura 2.1. Ilustração da Disposição Espacial Típica de GCMs.



(a) Representação global de malhas horizontais; (b) Zoom: vista lateral. A discretização vertical é representada por meio de malhas horizontais, que descrevem o comportamento atmosférico em diversos níveis; (c) Zoom:

vista superior. É representado um estado (ou “padrão”) sinóptico contendo as componentes zonal e meridional da velocidade do vento (U e V), dadas em um instante t . Fonte: Própria

Há três principais produtos de modelos de circulação geral da atmosfera: análise, reanálise e previsão. Os dados de análise gerados por modelos operacionais (i.e., em tempo real) possuem uma base heterogênea devido às constantes atualizações nas parametrizações dos modelos operacionais e utilização de dados observacionais com distribuição espacial irregular. Diferentemente da análise, a reanálise é um procedimento não-operacional em que se emprega um único modelo (mais apropriado) para todo o período de simulação, formando uma base de dados consistente. Espera-se que estes tenham um erro associado menor do que aquele relativo à análise. Os dados de previsão, por sua vez, são dados relativos ao estado futuro das variáveis atmosféricas (AMS, 2012). No presente estudo, são empregados dados de reanálise provenientes de diferentes GCMs para a modelagem do vento local.

O ERA-Interim é uma reanálise atmosférica global do ECMWF (*European Centre for Medium-Range Weather Forecasts*) com cobertura temporal na ordem de 40 anos. Ele é um produto do ECMWF que inclui uma análise variacional em 4 dimensões com 12 horas de janela de análise. Possui um banco de dados com uma resolução espacial de aproximadamente 80 km e conta com 60 níveis verticais, sendo o nível 60 o mais próximo à superfície. O banco de dados do ERA-Interim é gratuito para utilizações não comerciais (DEE et al., 2011).

2.2 Aumento da Resolução (*Downscaling*) do Vento Local

As técnicas de aumento de resolução, ou técnicas de *downscaling*, são amplamente empregadas para fins de acoplamento entre variáveis descritas em diferentes resoluções, i.e., em diferentes escalas espaciais e temporais. Essas técnicas estão classificadas em técnicas dinâmicas, empíricas e híbridas. Na primeira abordagem mencionada, são resolvidas numericamente leis de conservação de massa, momento e energia que regem o comportamento das variáveis em questão, sendo assim possível realizar o acoplamento direto entre a escala sinóptica (baixa resolução espacial) e a microescala (alta resolução espacial) ou mesoescala (resolução espacial intermediária) (seção 2.2.1). Uma importante vantagem dessas técnicas é que, de maneira geral, uma vez validadas, essas não necessitam de observações para que se obtenham simulações para outras regiões com condições similares.

A segunda abordagem busca estabelecer relações empíricas entre os conjuntos preditando (microescala) e preditor (macroescala) (WILBY et al., 1998), o que habitualmente

se realiza através de técnicas regressivas, de classificação de padrões sinóticos, ou estocásticas (seção 2.2.2). Os modelos de natureza híbrida são aqueles que combinam, em diferentes etapas, as técnicas dinâmicas e estatísticas (e.g., BOÉ et al., 2006) (seção 2.2.3). A principal utilidade das técnicas estatísticas é a capacidade de simular satisfatoriamente o comportamento de variáveis complexas, uma vez que houver uma relação estatística significativa entre as variáveis que são inter-relacionadas. Entretanto, é importante ressaltar que esta abordagem está limitada à disponibilidade e qualidade de dados que compõem os conjuntos preditor e preditando.

A seção 2.2.3 traz comentários sucintos sobre as diferentes arquiteturas de parâmetros adotadas para a calibração de modelos estatísticos.

2.2.1 Abordagem Dinâmica

As abordagens dinâmicas para as técnicas de *downscaling* dinâmico, podem ser classificadas em dois grupos principais: a modelagem da camada limite planetária (PBL, *Planetary Boundary Layer*, e.g., LANDBERG e WATSON, 1994) para realizar o 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 e 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.

A partir da lei de conservação do momento linear é possível obter expressões simplificadas para o comportamento do vento horizontal (dentre outras variáveis) na PBL, desprezando-se os termos de menor importância por via de análise dimensional (PEDLOSKY, 1984). A PBL é a porção atmosférica que está sujeita aos efeitos térmicos (trocas de calor) e mecânicos (orografia, obstáculos, etc.) que ocorrem a partir da interação com a superfície. Essa camada se subdivide em duas outras camadas: a camada superficial (tipicamente na ordem de centenas de metros de altura), na qual o vento varia verticalmente segundo um perfil logarítmico, e a camada externa de Ekman, localizada acima da anterior (LANDBERG e WATSON, 1994). O vento horizontal na camada externa de Ekman sofre variações na direção e magnitude devido a efeitos de fricção (VERKAIK et al., 2006). Acima da PBL situa-se a atmosfera livre, camada na qual o efeito da superfície pode ser desprezado e a componente horizontal resultante é denominada de vento geostrófico. Nessa condição, fenômenos característicos se manifestam em escalas globais, devido ao maior valor de livre caminho médio

das partículas fluidas, realidade bastante diferente daquela que se apresenta na superfície. A forma mais simples sob a qual tais modelos podem ser concebidos leva em consideração que os efeitos de troca de calor com a superfície terrestre são desprezíveis, caso denominado de regime neutro (LANDBERG e WATSON, 1994). Nos casos de regime não neutro (LANDBERG, 2001), nos quais o efeito térmico na PBL é expressivo, há um fluxo de vento ascendente ou descendente, a depender do sentido da troca de calor (da superfície para o ar ou do ar para a superfície, respectivamente). Esse fluxo perturba o perfil de cisalhamento vertical do vento e deve ser considerado em zonas de alta instabilidade, como no caso da região do Nordeste do Brasil, situada em baixas latitudes (próxima à linha do equador).

Os modelos regionais (RCMs, *Regional Circulation Models*; e.g., GRELL et al., 1994), possuem maior complexidade do que modelos microescalares (e.g., modelos baseados na PBL, mencionados no parágrafo acima) com respeito à complexidade do sistema de variáveis com os quais operam. Os RCMs são modelos similares aos GCMs, com a particularidade de serem projetados para descrever o comportamento atmosférico em malha mais refinada, a partir da qual o modelo incorpora informações com maior nível de detalhes sobre as condições do terreno (e.g., orografia). Outro ponto que o diferencia dos GCMs é que estes modelos não empregam procedimentos de assimilação de dados observacionais, ao invés disso, os RCMs recebem como parâmetro de entrada as simulações do GCM referentes aos pontos situados no contorno do domínio em questão. Os RCMs possibilitam a realização de simulações satisfatórias para regiões com terreno complexo, sendo necessários os dados observacionais apenas para fins de validação. As maiores desvantagens desse grupo são o grande esforço computacional requerido e o alto grau de complexidade associado ao detalhamento do sistema físico necessário para as simulações.

2.2.2 Abordagem Empírica

Em geral, as técnicas estatísticas são bastante flexíveis quanto à sua aplicabilidade, uma vez que os métodos de ajuste disponíveis permitem a proposição de um incontável número de modelos, os quais, por sua vez, relacionam variáveis diversas. Por este motivo, na maior parte dos casos não é necessário realizar grandes mudanças na concepção dos modelos empíricos para que bons resultados sejam obtidos. O mesmo não acontece na adaptação de modelos dinâmicos para regiões de alta instabilidade atmosférica, devido à complexidade associada às soluções analíticas de expressões diferenciais que descrevem fenômenos locais. Entretanto, é importante atentar aos cuidados necessários para a escolha de fatores como: a) domínio

empregado; b) variáveis preditoras; c) modelos apropriados para o estudo pretendido. A esse respeito, Wilby et al. (2004) propuseram um guia de boas práticas para a realização de *downscaling* estatístico.

A abordagem empírica possui a vantagem de produzir bons resultados levando em consideração sistemas bastante simplificados, devido à flexibilidade que as arquiteturas dos modelos estatísticos possuem, consequentemente demandando pouco esforço computacional (de maneira geral), quando comparados com os RCMs. Costuma-se subdividir o grupo de *downscaling* estatístico entre os subgrupos de técnicas de regressão, classificação e geradores estocásticos de tempo. No primeiro, procedimentos de otimização de parâmetros são adotados na determinação de funções de transferência (lineares ou não-lineares). O segundo busca identificar classes que definam diferentes estados atmosféricos, o que possibilita a parametrização de funções especializadas (locais) para as ocorrências classificadas. O terceiro subgrupo, por sua vez, descreve o comportamento das variáveis atmosféricas como processos estocásticos.

2.2.3 Modelos Estatísticos: Arquiteturas de Parâmetros

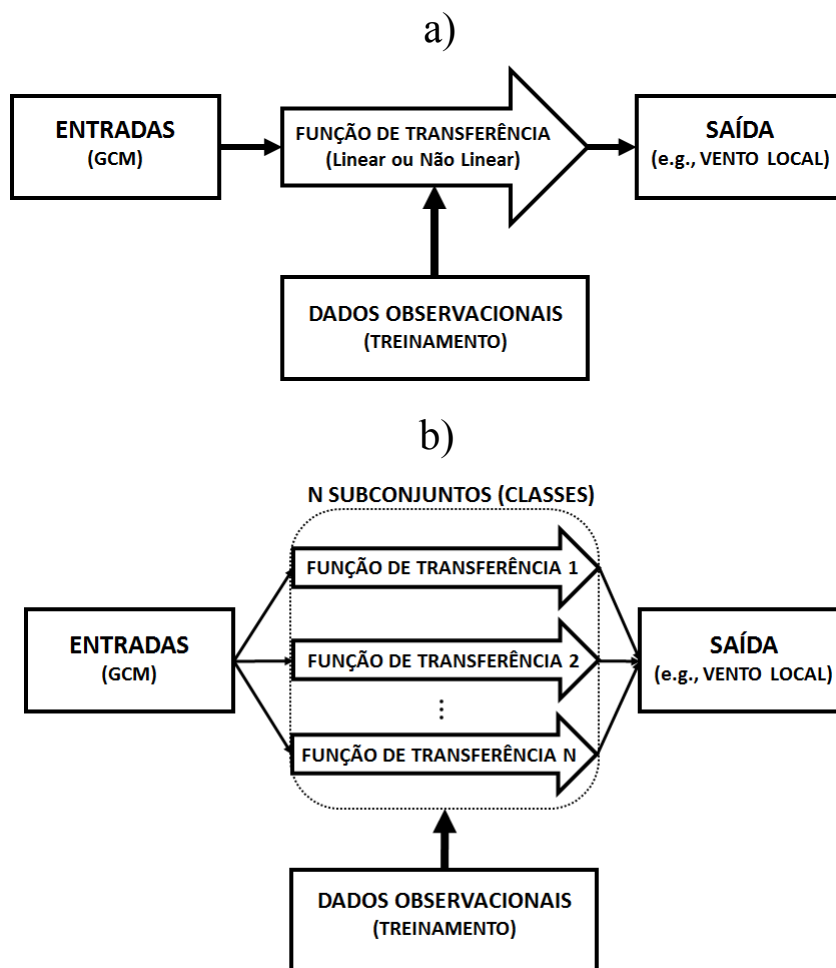
Há duas principais abordagens para a parametrização de modelos estatísticos: global e local (Figura 2.2). No caso específico de *downscaling* estatístico, normalmente os dados de simulação atmosférica provenientes dos GCMs são relacionados com dados observacionais (e.g., velocidades de vento) através de funções de transferência ajustadas em função da minimização do gradiente de erro global (calculado para todo o período disponível na série temporal). Nesse caso, considera-se que o modelo possui uma arquitetura global de parâmetros (Figura 2.2a).

Em contraste com a abordagem global, arquiteturas locais possuem parâmetros variáveis no tempo, os quais são especializados para subconjuntos específicos (classes) formados a partir dos dados disponíveis. Isso quer dizer que, nesse caso, ajustes são realizados em função da minimização de gradientes de erro locais¹ (Figura 2.2b). Tipicamente, os subconjuntos são caracterizados por métodos de classificação que identificam N grupos de estados atmosféricos (subperíodos da série temporal). O intuito das técnicas de classificação é o de definir um ou vários subconjuntos, a partir dos quais são compreendidos os padrões de estados atmosféricos mais relevantes. Em seguida, costuma-se aplicar técnicas regressivas ou

¹ O termo “local”, neste caso, diz respeito à dimensão temporal, i.e., subconjuntos de instantes específicos das séries temporais.

estocásticas aos grupos individualmente, de forma a compor uma estimativa final por meio da união das N estimativas geradas para cada subconjunto classificado.

Figura 2.2. Arquiteturas de Parâmetros de Modelos Estatísticos



Dados derivados de GCMs são relacionados com o vento local por meio de funções de transferência (lineares ou não lineares); a) Abordagem Global, uma função de transferência é ajustada para todo o período de treinamento;

b) Abordagem Local, diversas funções de transferência são especializadas para subconjuntos, ou classes, formadas a partir do conjunto de variáveis preditoras (dados de simulação numérica). Fonte: Própria.

Para a abordagem global e local, normalmente se adota um conjunto específico para a calibração das funções de transferência, e outro apenas para validar os parâmetros resultantes. Caso sejam considerados bem ajustados, o modelo poderá então ser adotado para estimar o comportamento da variável em função apenas das variáveis de entrada (simulações proveniente de GCMs). No caso de modelos de parâmetros locais, costuma-se adotar técnicas para a identificação de classes de padrões sinóticos (ver Fig. 2.1c) para estimar o comportamento local de variáveis atmosféricas com maior acurácia.

A metodologia empregada para a parametrização de modelos de *downscaling* estatístico representada na Figura 2.2 também pode ser adaptada para ilustrar processos típicos de combinação de modelos. Note que cada modelo representado pela figura produz uma saída, ou uma estimativa para a variável de interesse (e.g., vento local). As séries estimadas pelos modelos podem ser então consideradas como entrada, para um novo processo de calibração de funções de transferência análogo aos adotados anteriormente, gerando assim uma estimativa final combinada.

A Seção 3 trará uma discussão mais aprofundada sobre técnicas de *downscaling* estatístico. Algumas das técnicas discutidas nesta seção foram implementadas no presente trabalho de dissertação, estando essas detalhadas na Seção 4 (Metodologia), e seus resultados apresentados e discutidos na Seção 5 e no Apêndice A (Resultados Individuais de Modelos).

A revisão bibliográfica, apresentada na próxima seção, busca levantar o estado-da-arte de técnicas e modelos relacionados com a metodologia proposta. Particularmente, o tema deste trabalho está focado na aplicação de técnicas de *downscaling* estatístico e estratégias de combinação para a obtenção de estimativas suficientemente confiáveis para o comportamento da velocidade do vento nas proximidades de um local de interesse.

3 REVISÃO BIBLIOGRÁFICA

Discutidos de forma preliminar na Subseção 2.2, as técnicas de *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. Wilby e Wigley (1997) e Xu (1999) apresentaram revisões sobre as principais técnicas desse grupo disponíveis à época.

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. Lam (1983) e Li e Heap (2008) apresentam uma vasta gama de abordagens e aplicações. 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; PRYOR et al., 2005; BERNHARDT et al., 2010). As ponderações geralmente se dão com respeito às posições no plano horizontal, entretanto, é possível generalizar o método incluindo outras informações relevantes para o estudo em questão. Como exemplo, particularmente nos casos de modelagem do vento próximo a superfícies de orografia complexa, mostrou-se que a elevação é um importante fator a ser considerado (PALOMINO e MARTÍN, 1995). Outros importantes tipos de interpolação são: técnicas de vizinhos mais próximos (SNELL et al., 2000); interpolação multiquádrica (DEGAETANO e BELCHER, 2007); e “krigagem” (OLIVER e WEBSTER, 1990; BIAU et al., 1999). Essa última apresenta um maior nível de complexidade, devido ao fato de que é baseada na teoria da variável regionalizada, a qual descreve as variações espaciais por meio de abordagens estocásticas.

As técnicas de regressão também são comumente empregadas em modelos de arquitetura global. Essas apresentam um maior nível de sofisticação quando comparadas a interpolações devido ao fato de possuírem procedimentos de otimização de parâmetros, que minimizam um gradiente de erro de ajuste entre os conjuntos preditando e preditor em questão. Tais técnicas estabelecem, entre os conjuntos preditando e preditor, relações lineares ou não lineares. Uma das técnicas lineares mais comuns é a regressão multilinear, frequentemente empregada em estudos de comparação de modelos, tendo em vista que apesar de serem de implementação relativamente simples, costumam apresentar resultados satisfatórios (MURPHY, 1999; CURRY et al., 2012). Outras abordagens relacionadas incluem a regressão

linear não paramétrica e regressão sujeita a restrições. Sobre a primeira, Rozas-Larraondo e Lozano (2014) propuseram um modelo baseado em regressão linear ponderada, na qual os pesos são obtidos a partir da estimação de densidade de kernels o que, segundo os autores, resulta em uma abordagem mais intuitiva, o que simplifica sua interpretação. A exemplo da segunda, Bürger (1996) apresentou um método regressivo com restrições à variância preservada da variável local. Nos três últimos estudos mencionados, resultados satisfatórios foram obtidos com relação a outras técnicas mais complexas, apesar do menor grau de complexidade das abordagens propostas.

As redes neurais artificiais são bastante difundidas dentre as técnicas de regressão não linear, sendo consideradas superiores a outras técnicas comuns de regressão (VON STORCH et al., 2000). Isso se deve à capacidade adaptativa do método na definição de uma forma funcional, sendo realizada em função dos próprios dados de treinamento e servem como aproximações de funções não lineares complexas, diferentemente de outras técnicas de regressão, nas quais se define à priori a forma funcional do modelo (e.g., linear, polinomial, etc.). A arquitetura em redes neurais 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 (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 (DAWSON e WILBY, 2001; SCHOOFF e PRYOR, 2001; CANNON e WHITFIELD, 2002). Outra abordagem do tipo *feed-forward* menos usual emprega funções de base radial (HARPHAM e WILBY, 2005). Outras abordagens mais complexas são usualmente adotadas quando se busca ajustar informações da dependência temporal em um determinado sistema para melhor descrever comportamentos transientes. A esse respeito, há duas principais abordagens: rede neural temporal (COULIBALY et al., 2005; DIBIKE e COULIBALY, 2006; MENDES e MARENGO, 2010) e rede neural recursiva (*recurrent neural network*, MEDSKER e JAIN, 2001). Essa última envolve algoritmos mais complexos, que demandam maiores esforços computacionais e, portanto, são menos usadas.

Com arquitetura de parâmetros locais, diversos modelos têm sido propostos através da aplicação de técnicas conhecidas como classificação de padrões sinópticos (WILBY e WIGLEY, 1997). A técnica baseada em “análogos”, proposta inicialmente por Lorentz (1969), está entre as mais simples de classificação, e, assim como a interpolação, serve como modelo de referência para técnicas mais complexas, sendo até mesmo equivalente ou superior em alguns casos (ZORITA e VON STORCH, 1999). O termo “análogos” se refere ao par de

instantes em que os respectivos estados sinópticos “próximos” entre si, sendo um relativo ao instante alvo, a ser estimado, e outro relativo ao banco de dados. É comum o emprego da distância euclidiana para identificar a proximidade entre os padrões sinópticos (i.e., instantes análogos; e.g., BARNETT e PREISENDORFER, 1978; TIMBAL et al., 2003), porém há outras distâncias mais genéricas disponíveis na literatura, como a distância euclidiana ponderada e distância de Mahalanobis (WILKS, 2011). Como função de transferência que os relaciona, costuma-se empregar a identidade (e.g., ZORITA e VON STORCH, 1999; WETTERHALL et al., 2005) ou regressão linear (e.g., HIDALGO et al., 2008). Tippet e DelSole (2013) compararam essa última com uma regressão linear pura e constataram que ambas são similares com respeito às suas aptidões (capacidade de reproduzir as observações) e limitações (e.g., sobreajustes, pressupostos relacionados às regressões lineares).

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. De maneira geral, há duas limitações especificamente 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 (VAN DEN DOOL, 1994).

As técnicas de agrupamento (*clustering*) são bastante aplicadas em modelos de arquitetura local, estando dentre as mais difundidas do grupo e 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 o mais internamente homogêneos e externamente heterogêneos quanto for possível, 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; WARD, 1963; HART et al., 2015). 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 e SPEKAT, 1997; GUTIÉRREZ et al., 2004). Outro exemplo de *clustering* hierárquico é a técnica conhecida como árvore de classificação e regressão (CART). A análise CART emprega dados de variáveis sinópticas (proveniente de GCMs) e locais em um procedimento recursivo no qual se obtém uma árvore de decisão binária, cujos nós de decisão definem classes que representam os estados sinópticos característicos (ZORITA et al., 1995; ENKE e SPEKAT, 1997; ZORITA e VON STORCH, 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-objetivos². 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 et al., 1995; DUBOIS e PRADE, 1996; BARDOSSY et al., 2005).

Os geradores estocásticos de tempo são modelos probabilísticos que inicialmente foram empregados para fins agrícolas, hidrológicos e de preservação ambiental (IPCC, 2013). Atualmente têm sido bastante explorados para fins de descrição de variáveis em escala local (BERNARDIN et al., 2009; WILKS, 2010) e estudos de impactos causados por mudanças climáticas (SEMENOV e BARROW, 1997; WILKS, 1999). Esses modelos produzem séries temporais de várias variáveis (vento horizontal, precipitação, temperatura máxima e mínima diária, radiação solar e umidade) com comprimento ilimitado para um determinado local, por meio de inferência estocástica (HUTCHINSON, 1995). A metodologia proposta por

² 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.

Richardson (1981) está entre as principais ainda empregadas. Nela, a ocorrência ou não de precipitação diária pode ser descrita através um processo markoviano com dois estados: úmido e seco. Uma crítica à essa abordagem é que a mesma não reproduz satisfatoriamente eventos extremos como os períodos de seca prolongados e chuvas de longa duração. A fim de superar essa limitação, SWGs do tipo “serial” foram propostos. Esse último ajusta funções de distribuição de probabilidade condicional a partir dos dados observacionais de treinamento (RACSKO et al., 1991).

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 filtrar sistemas de dados multidimensionais através da análise de componentes principais (“*Principal Component Analysis*”, PCA; e.g., JOLLIFFE, 1986; BORDONI e STEVENS, 2006), análise de correlações canônicas (e.g., GYALISTRAS et al. 1994; BUSUIOC et al., 2008), ou ambas (HUTH, 1999). Estas técnicas formam novas bases para o sistema em questão sob as quais se costumam aplicar técnicas de regressão e classificação (por exemplo, MLR, NN, análogos, *clustering*, etc.) para inter-relacionar os conjuntos preditando e preditor. Maiores detalhes a respeito dessas técnicas estão disponíveis em Wilks (2011). Outras abordagens de compressão de dados mais complexas incluem: 1) seleção de atributos associada com teoria da informação mútua (AMJADY et al., 2011); 2) decomposição espectral por meio de transformada *wavelet* (CATALÃO et al., 2011; WANG et al., 2015); 3) decomposição em modos empíricos (EMD, *Empirical Mode Decomposition*; LIU et al., 2012).

Os diferentes modelos propostos para simular o comportamento de variáveis atmosféricas (e.g., vento, pressão, índices de precipitação, etc.) possuem aptidões e limitações diferentes, no que se refere à habilidade em reproduzir características das observações (WILBY et al., 1998). Para melhor aproveitar as vantagens de abordagens distintas, metodologias foram propostas para a combinação de diferentes modelos, as quais podem ser divididas em dois grupos principais (em função do objetivo desejado): combinação voltada para aprimoramento e combinação voltada para adaptação (YANG, 2004). Na primeira busca-se obter uma ou várias funções de transferência, lineares ou não lineares, das quais resulta uma estimativa mais confiável com relação às saídas individuais (e.g., HIBON e EVGENIOU, 2005; COSTA et al., 2008; KANG et al., 2008). Diversos trabalhos demonstraram um aumento na acurácia de estimativas combinadas por meio de técnicas regressivas, tais como: regressão multilinear, redes neurais, lógica difusa, dentre outras. Jung e Broadwater (2014) apresentam uma revisão

mais detalhada sobre a aplicação dessas técnicas à previsão de recurso eólico (velocidade do vento e/ou potência de despacho). A segunda, por sua vez, está voltada para a obtenção de um sinal de saída similar ou superior à saída do modelo que apresente o melhor desempenho dentre os combinados. Regressões com parâmetros variáveis ou técnicas de aprendizagem de máquinas podem ser empregadas para esse fim (SÁNCHEZ, 2008; VACCARO et al., 2011). Como exemplo, Sánchez (2008) apresenta um modelo de combinação adaptativo com ponderações às performances das saídas individuais.

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 (YU e LIU 2003; KANG et al., 2008). 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 (LEEDS et al., 2013; JUNG e TAM, 2013). Isso impulsiona a busca por métodos que permitam avaliar a compatibilidade entre diferentes preditores através das distribuições de energia espectral entre as saídas dos modelos, entretanto, com relação ao *downscaling* estatístico especificamente, atualmente há um número relativamente baixo de publicações relevantes sobre o tema.

A análise proposta neste estudo permite destacar a relação de complementaridade entre modelos de *downscaling* estatístico e de combinação. Para isso, serão parametrizados e validados modelos diversos baseados em estratégias de regressão (linear e não linear) e classificação de padrões sinóticos. A variedade de modelos adotada nesse estudo, os quais se baseiam em diferentes pressupostos, se dá para que um amplo conjunto de saídas de modelos seja avaliado em estratégias de combinação, o que, em tese, confere maior acurácia à estimativa final (SÁNCHEZ, 2008). Nesse trabalho, dois procedimentos de seleção de variáveis foram empregados para a combinação dos modelos: PCA e ordenamento de saídas em função do índice de habilidade alcançado (TAYLOR, 2001). Por fim, a análise final se fundamenta em procedimentos de transformação para o domínio baseados em transformada de Fourier e transformada *wavelet*, visando a obtenção de informações sutis que ajudem a explicar a relação entre as aptidões entre os modelos, quer sejam únicas ou redundantes. Os modelos aqui implementados que se destacam possuem aptidões únicas, o termo “complementaridade”, nesse caso, se traduz na potencial união entre as tais virtudes.

4 METODOLOGIA E MODELOS

A metodologia a ser adotada possui três etapas principais (Figura 4.1). A primeira consiste na parametrização de diversos modelos de downscaling estatístico, os quais empregam técnicas de regressão pura (abordagem global) e de classificação de padrões sinóticos (abordagem local). O segundo bloco é relativo às estratégias de combinação aplicadas para associar os conjuntos mencionados. Por último, será realizada a análise de complementaridade entre as saídas individuais e combinadas.

Figura 4.1. Ilustração da Estratégia de Ação.



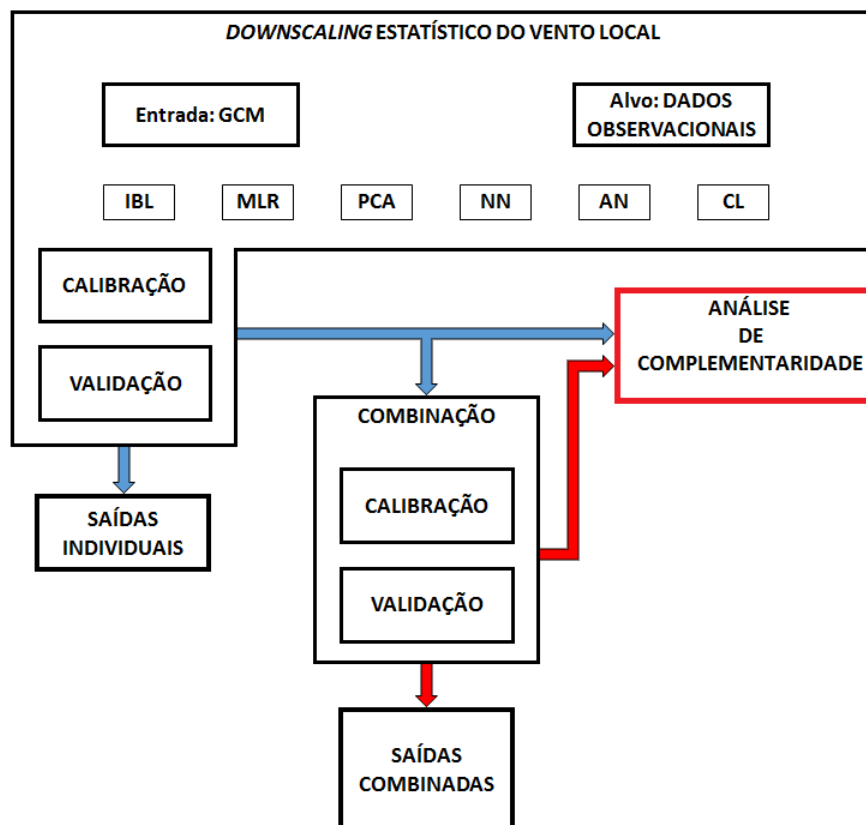
Fonte: Própria.

A Figura 4.2, adiante, ilustra com mais detalhes a metodologia empregada. O bloco na parte superior da imagem representa o GCMs adotado no estudo (ERA-Interim), dos quais serão obtidos os conjuntos de variáveis preditoras para os modelos de *downscaling* estatístico. O grupo multidisciplinar de técnicas de *downscaling* estatístico avaliadas na primeira etapa deste estudo inclui interpolação bilinear (IBL), regressões lineares múltiplas (MLR), análise de componentes principais (PCA), redes neurais (NN), análogos (AN) e *clustering* (CL) e variantes. Os modelos são calibrados e validados através de uma estratégia de validação cruzada (*cross validation approach*, e.g., HUTH, 1999), na qual se calibra o modelo sob um período distinto daquele empregado na validação do resultado. De posse dos parâmetros dos modelos devidamente ajustados, realizam-se simulações individuais (saídas do primeiro bloco na Figura 4.1, e bloco de “saídas individuais” na Figura 4.2) para o período destinado à validação. Em seguida, as simulações obtidas são comparadas com os dados observacionais correspondentes ao mesmo período simulado. Os detalhes técnicos sobre os modelos adotados são apresentados na Seção 4.1.

Ainda com respeito às saídas individuais (modeladas por técnicas de *downscaling*), neste trabalho, o conjunto de preditores examinado para cada local de interesse é composto por dados de reanálise ECMWF ERA-INTERIM, os quais são aplicados com uma resolução horizontal de 0,75 graus (para latitude e longitude). Os domínios aplicados são compostos pelos 10 níveis de modelo mais próximos à superfície (51 a 60), bem como pelo nível superficial

(interpolação para uma altura fixa a 10 m acima do solo). As variáveis do GCM consideradas nesse estudo são as componentes zonal e meridional (“U” e “V”, respectivamente; ver Fig. 2.1c). A grandeza estimada nesse estudo é a magnitude da velocidade do vento, dada por $M = \sqrt{U^2 + V^2}$.

Figura 4.2. Ilustração da metodologia empregada.



(IBL: Interpolação Bilinear; MLR: Regressão Multilinear; PCA: Análise de Componentes Principais; Análise de Correlações Canônicas; Redes Neurais; AN: Análogos; CL: Agrupamento; AC: Análise de Complementaridade).
Fonte: Própria.

Com respeito à disposição horizontal dos domínios empregados para os modelos de *downscaling* estatístico, malhas contendo os 4, 16, e 36 pontos mais próximos à posição de interesse foram consideradas (e.g., Fig. 2.1b). Além disso, uma quarta configuração de domínio foi composta a partir da união entre os 11 níveis mencionados considerando as malhas horizontais com 36 pontos, totalizando assim 396 pontos (em disposição tridimensional). A Tabela 4.1 apresenta as configurações citadas, bem como os modelos relacionados. Os modelos associados às três primeiras configurações mostradas na tabela (4, 16 e 36 pontos de malha) são calibrados para cada nível individualmente, sendo assim, há 11 saídas produzidas por cada um

desses modelos. Na última configuração (396 pontos), o ajuste se dá para todos os níveis simultaneamente, havendo, portanto, apenas uma saída disponível para os respectivos modelos.

Tabela 4.1. Configuração de domínio adotada para cada modelo individual.

Domínios			
4 pontos	16 pontos	36 pontos	396 (36x11) pontos
IBL, MLR1, AN1, CL1, NN1	MLR2, AN2, CL2, NN2	MLR3, AN3, CL3	MLRp, ANp

Os domínios de 4, 16 e 36 pontos estão dispostos como malhas horizontais (em onze níveis). O domínio de 396 pontos consiste na união dos onze níveis (configuração tridimensional). IBL: Interpolação Bilinear; MLR: Regressão Linear Múltipla; CL: *Clustering*; AN: Análogos; NN: Redes Neurais. O Subscrito “p” na nomenclatura dos modelos faz referência ao emprego da análise de componentes principais (PCA). Fonte: Própria.

O índice de habilidade (“*skill score*”) adotado nesse trabalho para quantificar o desempenho geral dos modelos é definido de acordo com a Equação 1, onde R significa a correlação entre a saída do modelo e a observação e $\hat{\sigma}$ representa o desvio padrão da estimativa normalizado pelo desvio padrão das observações. Com o índice de habilidade SS4, os ajustes dos modelos sobre os dados observacionais são classificados dentre valores de zero até um, i.e., desde o pior até o melhor índice de qualidade, respectivamente (TAYLOR, 2001). O SS4 é adotado nesse estudo como critério para a classificação da qualidade geral de saídas dos modelos em todas as etapas. Isto é, o SS4 é adotado para quantificar a qualidade das estimativas geradas por diferentes configurações (variantes) de modelo, bem como os resultados obtidos para cada variante ajustada sobre conjuntos de preditores diferentes (e.g., níveis de modelo do GCM).

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

Como pode ser visto na Figura 4.2, as técnicas de combinação são parametrizadas a partir das saídas individuais de modelos estatísticos e de observações sobre o período de treinamento (conjunto de calibração), e avaliadas sobre o período de validação considerado para cada local de avaliado. As técnicas de combinação incluem técnicas de ajuste global baseadas em MLR e NN, as quais são associadas a dois procedimentos de seleção de variáveis. No primeiro, a seleção de variáveis se dá no âmbito de redução da dimensionalidade do sistema através de análise de componentes principais (PCA). No segundo, é adotado um procedimento

de busca do qual resulta o ordenamento de possíveis conjuntos de saídas em função dos respectivos valores de SS4 alcançados.

A conversão entre os domínios de tempo e frequência serão abordados tanto de forma global quanto local em três abordagens principais. Na abordagem global, propõe-se fazer uso da Transformada de Fourier aplicada diretamente a todo o período de tempo disponível, além disso, espectros globais suavizados serão obtidos. Duas abordagens locais (no domínio do tempo) são propostas para a suavização do espectro global. A primeira se dá por meio do emprego da Transformada de Fourier de Curto Prazo (TFCP). Na segunda, a Transformada *Wavelet* Contínua (TWC) é adotada. Essa última possibilita uma análise de tempo-frequência com múltiplas resoluções de frequência. O procedimento para a avaliação de complementaridade espectral proposto se baseia na estimação de densidades de energia espectral cruzada (entre modelos e observações) e densidades de energia auto espectral das observações. Mais especificamente, o erro relativo entre parâmetro baseados nos conceitos de espectro cruzado e auto espectro é calculado para cada modelo em função das frequências, e também de tempo (no caso da TFCP e TWC). Os detalhes sobre essas abordagens são apresentados na Seção 4.3.

4.1 Técnicas Empregadas em Estratégias de *Downscaling* Estatístico

Nesta seção, são apresentadas as técnicas aplicadas aos modelos de downscaling estatístico (Tabela 4.1), as quais também são empregadas no procedimento de combinação de modelos, como será apresentado na Seção 4.2.

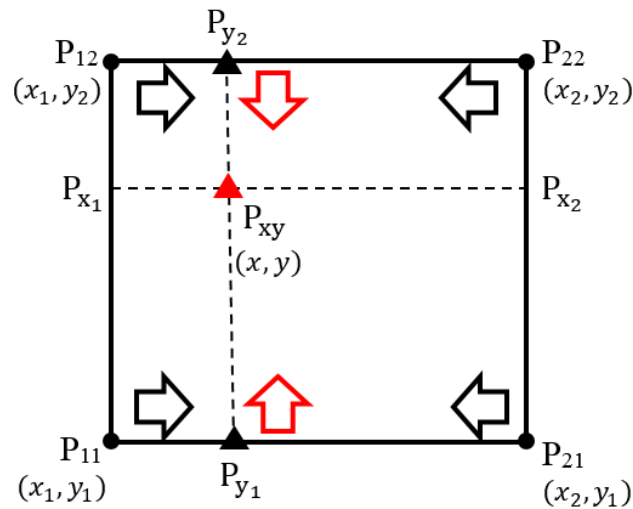
4.1.1 Interpolação Bilinear (IBL)

O modelo IBL representa a aplicação da técnica de interpolação bilinear à malha contendo os quatro pontos mais próximos à posição da torre anemométrica de interesse, sendo considerado neste texto como modelo de referência (o mais simples dentre os avaliados). Essa técnica, apesar de bastante empregada para extrapolação espacial de saídas de modelos de circulação geral da atmosfera, tende a subestimar variabilidade dos processos atmosféricos na microescala, devido à própria concepção da mesma, a qual consiste em uma suavização das variáveis em questão (ACCADIA et al., 2003). O método representa o cálculo da média ponderada pelas distâncias, e, por isso, resulta em um aumento nos valores mínimos locais e

uma redução nos valores máximos locais com respeito ao preditando (i.e., subestima a variância do sinal observado).

A interpolação bilinear, quando aplicada a variáveis atmosféricas, parte do pressuposto que tais variáveis são função apenas das posições em que se encontram no espaço. Ilustrada na Figura 4.3, a técnica faz uso da malha que contém os quatro pontos mais próximos à posição de interesse (P_{11} , P_{12} , P_{21} e P_{22}). O cálculo pode ser subdividido em três etapas, as quais consistem em interpolações lineares simples, sendo duas na direção da abscissa (setas horizontais) e uma terceira na direção da ordenada (setas verticais) ou vice-versa. Em outras palavras, uma interpolação linear simples é realizada, para o ponto de interesse (P_{xy}), fazendo-se uso de dois pontos previamente interpolados (P_{x_1} e P_{x_2} ou P_{y_1} e P_{y_2}).

Figura 4.3. Interpolação Bilinear.



Fonte: Própria.

Desta forma, uma estimativa para a magnitude do vento local na posição de interesse será dada por (Equação 2):

$$f(P_{xy}) = f(x, y) = \left(\frac{y_2 - y}{y_2 - y_1} \right) f(P_{y_1}) + \left(\frac{y - y_1}{y_2 - y_1} \right) f(P_{y_2}) \quad (2)$$

Os valores de $f(P_{y_1})$ e $f(P_{y_2})$ são obtidos através dos quatro pontos da malha do GCM que contornam a posição de interesse, tal como a seguir (Eqs. 3 e 4):

$$f(P_{y_1}) = f(x, y_1) = \left(\frac{x_2 - x}{x_2 - x_1} \right) f(P_{11}) + \left(\frac{x - x_1}{x_2 - x_1} \right) f(P_{21}) \quad (3)$$

$$f(P_{y_2}) = f(x, y_2) = \left(\frac{x_2 - x}{x_2 - x_1} \right) f(P_{12}) + \left(\frac{x - x_1}{x_2 - x_1} \right) f(P_{22}) \quad (4)$$

4.1.2 Regressão Multilinear (MLR)

Ao todo, três variantes de modelo baseadas em regressão multilinear “pura”, no sentido de serem aplicados diretamente às configurações de domínio apresentadas na Tabela 4.1, sem a aplicação de qualquer técnica auxiliar (e.g., pré-processamento de dados, classificação de padrões sinóticos), resultando em três variantes de modelo: MLR1, MLR2 e MLR3. Vale ressaltar que outros modelos implementados nesse estudo aplicam regressões multilíneares em associação com análise de componentes principais (MLRp), análogos (AN1, AN2, AN3, ANp), agrupamento de padrões sinóticos (CL1, CL2 e CL3) e combinação de modelos (CBLb e CBLp).

Em síntese, essa técnica parametriza funções lineares empíricas que descrevem uma variável de interesse (e.g., vento local) em função de k variáveis regressoras (e.g., saídas de GCMs) através de $k + 1$ parâmetros lineares β_i . De maneira geral, as variáveis regressoras consideradas em todos os modelos mencionados provêm de alguma maneira do GCMs, sendo a combinação a abordagem mais indireta nesse contexto. O preditando, em todos os casos, contém necessariamente informações dos dados observacionais anemométricos. O sistema assim estabelecido pode ser representado tal como abaixo:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Em que:

y_i – dados observacionais; β_j – parâmetros do modelo; x_{ij} – variáveis regressoras; ε_i – erro associado ao modelo.

A Equação 5 apresenta o sistema descrito em notação matricial (GELADI e KOWALSKI, 1986):

$$Y = X\beta + \epsilon \quad (5)$$

A otimização dos parâmetros ($\hat{\beta}$) pode ser obtida a partir da minimização do erro quadrático (ϵ_q), definido como:

$$\epsilon_q = \sum_{i=1}^n \epsilon_i^2 = \epsilon' \epsilon = (Y - X\beta)'(Y - X\beta)$$

Assim, tem-se que:

$$\frac{\partial \epsilon_q}{\partial \beta} = -2X'Y + 2X'X\hat{\beta} = 0$$

Resolvendo a expressão acima para $\hat{\beta}$ de acordo com a expressão abaixo, obtém-se a seguinte expressão matricial (6):

$$\hat{\beta} = (X'X)^{-1}X'Y \quad (6)$$

A série estimada será dada pela expressão abaixo (7):

$$\hat{Y} = X\hat{\beta} \quad (7)$$

Diferentemente da interpolação bilinear (seção 4.1.1), a regressão multilinear é parametrizada através da minimização do erro de ajuste. Por isto, espera-se que os modelos baseados nessa técnica apresentem melhores ajustes com relação à interpolação bilinear, cuja parametrização não considera o comportamento do vento local.

4.1.3 Análise de Componentes Principais (PCA)

Modelos baseados em regressão multilinear, análogos e combinação (MLRp, ANp e CBLp, respectivamente; ver Tabela 4.1) foram submetidos à análise de componentes principais (PCA, “*Principal Component Analysis*”), ou funções empíricas ortogonais, como também é conhecida a técnica. Essa técnica é bastante utilizada quando se busca a compactação da informação contida em um grande conjunto de dados, uma vez que reduz a dimensionalidade do sistema de dados ao eliminar informação redundante, preservando a variância original na medida do possível. Isso é de vital importância quando há um grande esforço computacional no processamento de tais dados (JOLLIFFE, 1986).

Considera-se um espaço multidimensional formado por m variáveis preditoras com dimensão temporal t ($R_{t \times m}$). As componentes principais são vetores que indicam as direções sobre as quais são projetados os eixos da base ortogonal deste espaço. Estas direções (componentes), sujeitas à restrição de serem ortogonais entre si, ficam determinadas pela maximização da variância que as mesmas comportam. Tais componentes e suas variâncias correspondentes são representadas, respectivamente, pelos autovetores e autovalores da matriz de correlação ($\Sigma_{m \times m}$) do espaço em questão. Mantém-se o menor número n de componentes ($P_{m \times n}$) de forma que o erro de reconstrução seja inferior a 5% (e.g., GUTIÉRREZ et al., 2004). A projeção do espaço original nas n componentes principais ($X_{t \times n}$, Eq. 8), o espaço reconstruído por tais componentes ($R_{t \times m}^*$, Eq. 9) e o erro de reconstrução (E_R , Eq. 10) são definidos de acordo com as equações (8), (9) e (10), respectivamente:

$$X_{t \times n} = \tilde{R}_{t \times m} \times P_{m \times n} \quad (8)$$

$$R_{t \times m}^* = \tilde{X}_{t \times n} \times P_{n \times m}' \quad (9)$$

$$E_R = \sqrt{\frac{1}{M} \frac{1}{T} \cdot \sum_{j=1}^M \sum_{i=1}^T \frac{(R_{i \times j}^* - \tilde{R}_{i \times j})^2}{R_{i \times j}^2}} \quad (10)$$

O termo $\tilde{R}$ representa a matriz de dados preditores e componentes principais preservadas estando ambas padronizadas, ou seja, com médias nulas e desvios unitários. “ M ” e “ T ” indicam o número de variáveis regressoras e número de passos temporais (“*timesteps*”), respectivamente.

4.1.4 Análogos (AN)

As variantes de modelos baseado em análogos estão dispostas na Tabela 4.1 (AN1, AN2, AN3, ANp). A técnica de classificação de padrões sinóticos baseada em análogos está entre as mais simples (ZORITA e VON STORCH, 1999). Ela parte do princípio que estados sinóticos (F , e.g., Fig. 2.1c) similares da atmosfera produzem observações análogas na microescala (escala local). Desta forma, o preditando S é simulado em um instante t , a partir das variáveis preditoras (análogos) em um instante distinto u . A escolha do instante u , ou instante análogo, se dá através do grau de similaridade entre os estados sinóticos, como a seguir:

$$\min ||F(u) - F(t)||$$

Os análogos são comumente aplicados a sistemas multidimensionais, de forma que se faz conveniente o emprego da análise de componentes principais para reduzir a dimensionalidade do problema. Dessa forma, o conjunto de variáveis regressoras padronizadas, i.e., com média nula e variância unitária, pode ser representado no espaço de componentes principais (CP), como representa a equação abaixo (11):

$$A(t, k) = \sum_{j=1}^p b_j(t) \cdot c_j(k) + \epsilon(t) \quad (11)$$

Em que: k – índice de ponto de malha; t – índice temporal; $b_j(t)$ – amplitude da anomalia, em t , projetada na j -ésima CP; $c_j(k)$ – coeficiente (autovetor) da j -ésima CP relativo ao ponto de malha de índice k ; $\epsilon(t)$ – variabilidade não descrita pelas p CPs, em t . Note que $p < \max(k)$, devido à redução da dimensionalidade do sistema por meio da eliminação das componentes com menor variância associada.

Os coeficientes $B(t) = b_1, b_2, \dots, b_p(t)$ são conhecidos como “coeficientes de expansão” da CP correspondente. O análogo será obtido através de uma ocorrência de anomalia cujos coeficientes de expansão $Z = z_1, z_2, \dots, z_p$, estando descritos no mesmo sistema de coordenadas das componentes, minimizam a seguinte relação (12):

$$\sum_{i=1}^p (z_i - b_i(t))^2 \quad (12)$$

A obtenção das estimativas $\hat{S}$ se dá, por exemplo, considerando-se a identidade como a função que relacionam o preditando com o preditor (Eq. 13; WETTERHALLET al., 2005):

$$\hat{S}(t) = S(u) \quad (13)$$

Em que: $S(u)$ – preditando no instante análogo u .

Neste estudo, considera-se uma expressão mais geral para a estimativa, a qual pode ser representada como a seguir (14):

$$\hat{S}(t) = f(S(u), X(u), X(t)) \quad (14)$$

Na qual: $X(u)$ – conjunto de variáveis preditoras no instante análogo u ; $X(t)$ – conjunto de variáveis preditoras no instante atual t ; f – função empírica. No presente estudo, a função f foi parametrizada através de uma regressão multilinear (Seção 4.1.2, Equação 6) que toma como parâmetros de entrada os conjuntos preditores $S(u)$, $X(u)$ e $X(t)$.

Além das abordagens propostas nas equações (13) e (14), neste estudo, foram avaliadas também as relações entre preditando e preditor dadas por $\hat{S}(t) = f(S(u))$, $\hat{S}(t) = f(X(u))$ e $\hat{S}(t) = f(S(u), X(u))$. Após testes, verificou-se que a arquitetura de parâmetros que apresenta melhores resultados é aquela representada na Equação 14, sendo portanto, a arquitetura adotada para todos os modelos AN.

4.1.5 Agrupamento de Padrões Sinópticos (CL)

Esta seção apresenta o procedimento adotado para o ajuste de modelos baseados em agrupamento de padrões sinópticos, CL1, CL2 e CL3 (Tabela 4.1). Tais modelos, diferenciam-se dos demais por abordagem de parâmetros locais especializados para períodos específicos, os quais são identificados em função de comportamento atmosféricos similares.

Para isso, neste estudo, são consideradas duas abordagens: (a) agrupamento hierárquico aglomerativo (AHC, *Agglomerative Hierarchical Clustering*; SCHOOFF e PRYOR, 2001; VRAC et al., 2007); (b) procedimento não hierárquico (NHC, *Non-Hierarchical Clustering*) baseado no algoritmo *k-means* (e.g., GUTIÉRREZ et al., 2004); No AHC, o número de clusters C é estabelecido à posteriori, em função de um critério de parada de fusões. No NHC, por sua vez, C é decidido à priori, em função das classes adotadas na inicialização³. No método proposto, as classes adotadas para a inicialização do NHC são obtidas através de função do acoplamento com as saídas do AHC (N centróides).

A estimação do vento local é realizada através de regressões lineares múltiplas que relacionam a magnitude do vento simulado pelo GCM e a magnitude do vento observado (Equação 14). O procedimento aqui proposto consiste na aplicação de parâmetros

³ 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.

realizadas aos pares de *clusters* relacionados com o menor valor de D obtido em cada nível hierárquico, até que seja formado um único grupo contendo todas as amostras do sinal analisado, naturalmente associado ao maior valor de D .

Tabela 4.2. Configuração geral de técnicas de agrupamento.

a) Distâncias entre padrões (x_t e x_s)	Equações
Manhattan	$d = \sum_{j=1}^n x_{tj} - x_{sj} $
Euclideana	$d = \left[\sum_{j=1}^n (x_{tj} - x_{sj})^2 \right]^{1/2} = \ x_t - x_s\ $
Cosseno	$d = 1 - \frac{x_t x_s'}{\sqrt{(x_t x_t') (x_s x_s')}}$
Correlação	$d = 1 - \frac{(x_t - \bar{x}_t)(x_s - \bar{x}_s)'}{\sqrt{(x_t - \bar{x}_t)(x_t - \bar{x}_t)' \cdot (x_s - \bar{x}_s)(x_s - \bar{x}_s)'}}$
b) Dissimilaridades (D) entre c_p e c_q	Algoritmos
Vizinhos mais Próximos (VP)	$D(c_p, c_q) = \min \left(d(c_{pu}, c_{qv}) \right), i \in \{i, \dots, n_p\}, j \in \{1, \dots, n_q\}$
Vizinhos mais Distantes (VD)	$D(c_p, c_q) = \max \left(d(c_{pu}, c_{qv}) \right), i \in \{i, \dots, n_p\}, j \in \{1, \dots, n_q\}$
Vínculo Médio	$D = \frac{1}{n_p n_q} \sum_{u=1}^{n_p} \sum_{v=1}^{n_q} d(c_{pu}, c_{qv})$
Vínculo de mínima variância agregada (Método de Ward)	$D = \sqrt{\frac{2n_p n_q}{(n_p + n_q)}} \ \bar{c}_p - \bar{c}_q\ $

(a) Métricas adotadas AHC e NHC; (b) Algoritmos de Hierarquização de Dissimilaridades. Na primeira parte (a), os padrões sinópticos x_t e x_s são vetores n -dimensionais contendo as variáveis da malha do GCM para os instantes t e s . Na segunda parte (b), c_p e c_q representam classes arbitrárias, as quais contém n_p e n_q elementos associados, respectivamente. Note que c_{pu} e c_{qv} correspondem, respectivamente, ao u -ésimo e v -ésimo elemento de c_p e c_q . Com respeito ao método de Ward, $\bar{c}$ corresponde ao centroide, ou médias, tomadas a cada dimensão j dos padrões (x) pertencentes a uma classe c (vetor n -dimensional). Note que o método de Ward apenas se aplica a distâncias euclidianas.

Neste trabalho, propõe-se um processo iterativo de validação cruzada para a determinação de uma configuração ótima para N , métricas de distâncias, algoritmo de fusão e T (e.g., GUO et al., 2012). Para cada configuração considerada, estimativas sobre o período de validação são avaliadas com respeito ao SS4. Desse modo, a configuração cujo resultado apresenta o maior SS4 é adotada. Aqui, recomenda-se que seja determinado um critério de desempate em função da configuração mais robusta (e.g., análises de sensibilidade). Todavia, de forma pragmática, adotou-se um critério simples, o qual estabelece que a configuração ótima, em caso de empate, será a que apresentar estrutura mais simplificada (menor número de classes significativas, menor valor de limite inferior T , etc.).

4.1.6 Redes Neurais (NN)

De maneira geral, as redes neurais consistem em estruturas compostas por unidades de processamento, ou “neurônios”, os quais compõem uma rede de neurônios interconectados através de “pesos sinápticos”, geralmente dispostos em camadas (Figura 4.4). Há três tipos de camadas: 1) camada de entrada, cujo número de neurônios coincide com o número de variáveis preditores; 2) camadas internas ou escondidas, cada qual com um número de neurônios pré-definido; 3) camada de saída, com o número de neurônios igual ao número de saídas.

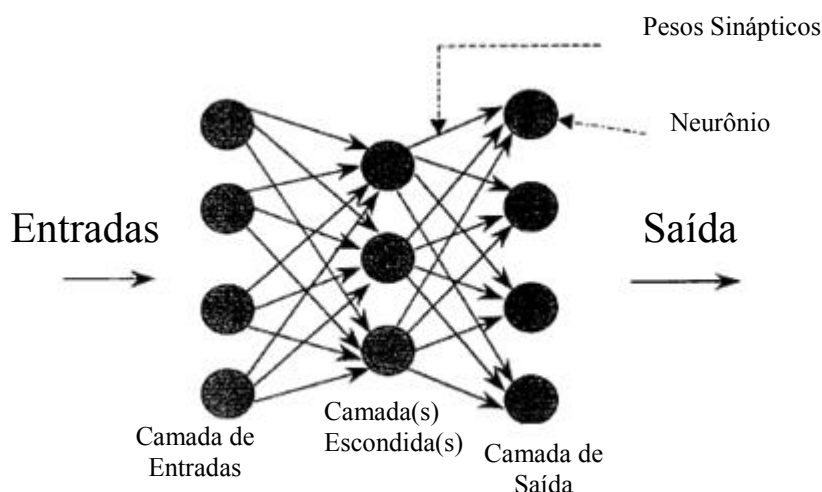
A Figura 4.4 ilustra a estrutura do *perceptron* de múltiplas camadas (*Multi-Layer Perceptron*, DAWSON e WILBY, 2001; SCHOOFF e PRYOR, 2001; CANNON e WHITFIELD, 2002). Essa possui arquitetura de propagação de informações unidirecional ou *feed-forward*, ou seja, informações são transferidas de uma camada para a seguinte sem retropropagação. De forma geral, os pesos sinápticos são definidos por meio de “funções de ativação” e ajustados em função de um determinado “algoritmo de treinamento”, sendo comum o emprego de funções sigmóides e o algoritmo de Levenberg–Marquardt (CATALÃO e POUSINHO, 2011).

Sistemas de redes neurais foram implementados com vistas ao ajuste dos modelos propostos a relações não lineares entre os domínios considerados (saídas do GCM) e os dados observacionais. A Tabela 4.1 apresenta detalhes sobre as variantes de modelos e configurações de domínio empregadas nessa análise (NN1 e NN2).

Arquiteturas do tipo *feedforward* foram adotadas no ajuste dos modelos NN1 e NN2. Embora seja pertinente a realização de uma busca exaustiva ou objetiva para uma configuração

ótima, neste estudo, configurações estáveis foram fixadas empiricamente, após testes⁵ (não exaustivos). Nessa estrutura, empregam-se 4 camadas internas, com um número decrescente de neurônios por camada. Para o NN1, o número de neurônios por camada interna é 4, 3, 3 e 2. Para o NN2, o número de neurônios por camada interna é 16, 8, 4 e 2. Outras configurações incluem 1) algoritmo de treinamento: Levenberg-Marquardt; 2) Função de ativação: tangente hiperbólica; 3) Número de inicializações de parâmetros aleatórios por época de treinamento: 200; 4) critério de parada para treinamento: erro quadrático médio ou 200 épocas de treinamento. Equivalentemente aos modelos MLR, os modelos baseados em NN aplicam parâmetros ajustados globalmente.

Figura 4.4. Ilustração de Rede Neural tipo Percéptron de Múltiplas Camadas.



Fonte: Adaptado de Sailor et al. (1999).

4.2 Combinação de Modelos

Como já foi mencionado, as saídas combinadas (ver Figura 4.2) são obtidas por meio de técnicas de regressão linear e não linear (Seções 4.1.2 e 4.1.6, respectivamente) que tomam como entrada os resultados obtidos pelos modelos apresentados na Seção 4.1. Ao todo, três modelos foram implementados, sendo dois lineares (CBLp e CBLb) e um não linear (CBnL).

O modelo CBLp aplica um conjunto preditor amplo, contendo todas as saídas dos modelos mostrados na Tabela 4.1. Para esse caso, aplica-se a análise de componentes principais (Seção 4.1.3) para a redução da dimensionalidade do sistema sobre o qual é ajustado o MLR.

⁵ Foi verificado que outras configurações podem produzir resultados similares, entretanto, há um maior esforço computacional associado.

O resultado do CBLb diz respeito à parametrização que melhor se ajusta aos dados observacionais, com respeito ao SS4, dentre aquelas obtidas por meio de MLRs ajustados para todos os K subconjuntos possíveis. O valor de K é calculado tal como na Equação 16, em que $\binom{N_s}{k}$ representa a combinação das N_s saídas dos modelos apresentados na Tabela 4.1, tomada k a k . Nessa abordagem, apenas uma saída é considerada para cada modelo, sendo aquela cujo nível apresenta o melhor resultado com respeito ao SS4. Assim, da Equação 16, resultam 1696 combinações de saídas possíveis ($N_s = 14$). Como critério de desempate, considera-se o nível mais próximo à superfície.

$$K = \sum_{k=2}^{N_s} \binom{N_s}{k} \quad (16)$$

O modelo baseado em regressão não linear foi implementado através de sistema de redes neurais (Seção 4.1.6). O CBnL é ajustado em função do mesmo subconjunto aplicado ao CBLb. Apenas o nível ótimo é aplicado para os modelos aplicados sobre os domínios com 4, 16 e 36 pontos, i.e., subconjuntos formados do total de 14 saídas são avaliados. As demais configurações são as mesmas empregadas para os modelos NN1 e NN2, exceto pelo número total de neurônios e sua distribuição nas camadas escondidas. A configuração das múltiplas camadas adotada para o CBnL é composta por 4 camadas internas com 7, 4, 4 e 2 neurônios (nessa ordem).

4.3 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. As ocorrências em escala local são influenciadas por diversos fatores, como a presença de obstáculos, corpos d’água, ocupação do terreno, dentre outros (SILVA, 2003). Por esse motivo, flutuações rápidas apenas detectáveis em resoluções locais podem ser associadas a ruídos. Isso pode ocorrer não apenas com as flutuações rápidas, mas com eventos em geral cujas periodicidades características não são exploradas de forma apropriada, fazendo com que tais eventos sejam subaproveitados em processos de

parametrização de modelos (WILBY et al., 2004; THEMESSEL et al., 2011; ZORITA e VON STORCH, 1999).

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 relevância 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*; WELCH, 1967; BERGLAND, 1969; MADDEN e JULIAN, 1971). O algoritmo FFT resulta em números complexos que transportam informações sobre a amplitude e fase de funções de onda harmônica que compõem o sinal, estando relacionado com a densidade espectral da “energia” contida na série temporal. Uma importante limitação dessa técnica é que pressupõe um comportamento estacionário da variável em questão, sendo seu uso inadequado para preservar informações de fenômenos transientes (GURLEY e KAREEM, 1999). Existem diversas variáveis que, assim como o vento próximo à superfície local, apresentam características não estacionárias, devido a uma importante componente estocástica associada. 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 Prazo⁶ (TFCP, *Shor-Term Fourier Transform*; e.g., ALLEN, 1977; GRIFFIN e LIM, 1984) e Transformada de Fourier Fracional⁷ (*Fractional Fourier Transform*; e.g., ALMEIDA, 1994; NARAYANAN e PRABHU, 2003), viabilizando com isso análises de tempo-frequência.

As funções *wavelet* formam bases de projeção para o domínio da frequência com a particularidade de possuírem representação finita no domínio do tempo e no domínio da frequência, sendo assim mais apropriadas para análises de comportamentos transientes (SIFUZZAMAN, et al., 2009; VELEDA et al., 2012; LIU et al., 2014). Outra importante vantagem dessas técnicas é que, diferentemente das técnicas baseadas em TFCP, 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

⁶ 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.

⁷ Generalização da transformada de Fourier clássica pela introdução de um termo arbitrário no expoente complexo que define uma rotação do plano tempo-frequência.

resolução, sendo janelas mais amplas aplicadas para descrever baixas frequências e janelas mais curtas para descrever altas frequências (JORDAN et al., 1997; TORRENCE e COMPO, 1998).

A análise espectral aqui proposta se dá com vistas à captação da informação sobre a complementaridade espectral em um determinado conjunto de saídas de modelo. O termo complementaridade, aqui, também pode ser entendido como a representatividade, sob a ótica do domínio da frequência, que se apresenta em um grupo de saídas de modelos candidatos à combinação. A análise no domínio da frequência permite que critérios quantitativos sejam estabelecidos para avaliar a similaridade entre dois sinais (e.g., espectro cruzado ou coerência espectral). Tal similaridade pode ser interpretada como o nível de compatibilidade entre um determinado par de preditor e preditando, e será mais explorada na Seção 4.3.4 (Análise de Complementaridade Espectral).

Três abordagens são propostas neste estudo, sendo uma global e duas locais. A abordagem global se dá por meio da aplicação direta da transformada rápida de Fourier (FFT, Seção 4.3.1) sobre as séries temporais. Dessa forma, uma estimativa da densidade de energia espectral média é obtida. Evidentemente, o espectro global não preserva informação sobre a evolução temporal do sinal considerado, entretanto, serve como abordagem de referência para outras técnicas mais sofisticadas. Como já mencionado na Seção 2.4, uma importante restrição associada com a transformada de Fourier é que essa pressupõe comportamentos estacionários sobre o sinal avaliado. Por isso, periodogramas⁸ foram estimados através de transformadas de Fourier de curto prazo (STFT, Seção 4.2.2) e de transformadas *wavelet* (Seção 4.2.3), com vistas à obtenção implícita de informações sobre efeitos transientes (não estacionários) que possam existir.

4.3.1 Transformada de Fourier (TF)

Resumidamente, a Transformada de Fourier (TF, Equação 17) pode ser entendida como uma operação matemática que projeta um sinal real (e.g., série temporal de vento) para uma base exponencial complexa $e^{-i2\pi ft}$, cujas partes reais e imaginárias podem ser representadas por uma combinação de infinitos termos de cosseno e seno (através da identidade de Euler: $e^{i\theta} = \cos \theta + i \sin \theta$). O valor complexo resultante possui representação no domínio da frequência através das amplitudes (coeficientes), frequências e fases associadas. A

⁸ Diagramas empregados em análise espectrais “localizadas” no domínio tempo, ou seja, para períodos específicos. Tais análises recebem o nome de análises de tempo-frequência.

transformada contínua de Fourier pode ser expressada tal como na Equação 17 (BERGLAND, 1969):

$$X(f) = \int_{-\infty}^{\infty} x(t)e^{-i2\pi ft} dt \quad (17)$$

Na equação acima (17), $x(t)$, f e t correspondem, respectivamente, a um sinal contínuo avaliado no domínio do tempo, a frequência e o tempo. Em notação discreta (DFT, *Discrete Fourier Transform*), a Equação 17 pode ser representada tal como em (18). Note que, neste texto, atribui-se o índice zero ao primeiro elemento de uma série discreta. Como já mencionado no início da Seção 4.3, normalmente se emprega o algoritmo heurístico FFT para estimar os coeficientes da transformada. O algoritmo reduz o número de operações para um número na ordem de $O(N \log_2 N)$, em contraste com as N^2 operações necessárias para o cálculo da DFT diretamente (BERGLAND, 1969).

$$X_n(j) = \frac{1}{N} \sum_{k=0}^{N-1} x_n(k) e^{-\left(\frac{i2\pi jk}{N}\right)} \quad j = 0, 1, \dots, N-1 \quad (18)$$

Em que $x_n(k)$ – sinal discreto no domínio do tempo; k – índice de tempo; j – índice de frequência; N – número de amostras do sinal x_n . O subscrito n indica notação discreta.

Note que, matematicamente, transformadas de sinais reais estão definidas até a frequência máxima $f_m = 1/\Delta t$, na qual Δt representa o período entre as amostras. A frequência máxima f_m pode ser obtida fazendo $j = N-1$ na Equação 18. Entretanto, de acordo com o teorema da amostragem de Shannon-Nyquist (STOICA e MOSES, 2005), a análise espectral de sinais reais pode ser realizada até uma frequência máxima de $f_m/2$, uma vez que o intervalo entre $f_m/2$ e f_m contém informação duplicada (“espelhada”) do intervalo entre f_m/N até $f_m/2$. Consequentemente, um sinal com período de amostragem de no máximo $\Delta t/2$ é necessário para que seja possível representar o espectro até f_m . Assim, em (18), considera-se o limite superior de j igual a $\widetilde{[N/2]}$, ou o valor inteiro (índice) que divide o espectro na mediana do intervalo de frequências.

A densidade de energia espectral pode ser calculada diretamente em função da Equação 18 (em notação discreta), como a seguir (19):

$$PSD(j) = \alpha \cdot |X_n(j)|^2 \quad (19)$$

Na Equação 19, α é um fator de normalização. Costuma-se adotar $\alpha = 2/(Nf_m)$, o qual normaliza pelo número de amostras e dimensiona (i.e., converte para valores de densidade) a energia espectral calculada por meio da Equação 19. Note que em tal fator, o numerador decorre em virtude do teorema de Shannon-Nyquist previamente mencionado e do teorema de Parseval. Esse último afirma que a energia total contida no domínio do tempo é conservada após conversão para o domínio da frequência. É importante ressaltar que fatores de normalização não influenciam resultados de análises relativas de espectro, uma vez que, nesse caso, a distribuição de energia espectral é “renormalizada” por outro valor de referência.

4.3.2 Transformada de Fourier de Curto-Prazo (TFCP)

Como discutido na Seção 2.4, o procedimento TFCP permite que informações temporais sejam preservadas através dos coeficientes da FFT aplicada sobre períodos de tempo mais curtos. Nesse caso, os períodos sobre os quais se aplica a FFT são definidos através de uma janela móvel (w) que é deslocada por todo o intervalo disponível. O resultado final pode ser demonstrado em um espectrograma⁹. A redução do intervalo temporal avaliado torna a análise mais sensível a efeitos transientes (i.e., variações rápidas), o que impacta no pressuposto de estacionariedade intrínseco à FFT. Por esse motivo, em abordagens locais, costuma-se aplicar uma função para a suavização (ou diminuição) dos extremos aos períodos enquadrados pela janela móvel (ALLEN, 1977). Esse procedimento confere ao sinal avaliado por FFT um comportamento “pseudo-periódico”, e é realizada para compensar a incapacidade do modelo em assimilar com comportamentos não-estacionários. Entretanto, deve-se observar que a redução do intervalo temporal de avaliação reduz a resolução de frequência.

O método (TFCP) pode ser representado em notação contínua, tal como na Equação 20, em que f e t representam parâmetros adimensionais de frequência e tempo, respectivamente:

$$X(f, t) = \int_{-\infty}^{\infty} w(t - \tau)x(\tau)e^{-i2\pi f\tau} d\tau \quad (20)$$

⁹ Diagramas empregados em análises de tempo-frequência, ou a visualização da distribuição da energia espectral em função do tempo e da frequência.

Para a representação em notação discreta, considere M intervalos definidos em função de janelas móveis, as quais compreendem T amostras e estão opcionalmente superpostas de acordo com passo temporal D . O número de segmentos (M) é igual ao maior valor inteiro $m + 1$ possível que satisfaz a restrição: $mD + T \leq N$ (Equação 21).

$$M = \max(m | mD + T \leq N) + 1, \quad m \in \mathbb{Z}^+ \quad (21)$$

Os M segmentos x_m (Eq. 22) são multiplicados por uma função de janela $w(v)$, no intuito de forçar um comportamento periódico nos intervalos mencionados, uma vez que a conversão da TF se dá sob o pressuposto de estacionariedade. Neste estudo, $w(v)$ foi estabelecido através de uma função de Hamming (ALLEN, 1977), definida em função de um número de elementos T e os índices v correspondentes¹⁰ (Equação 23). Essa função de janela reduz os valores nas extremidades da janela, forçando um aspecto semelhante ao de uma onda no período compreendido, mais compatível com o pressuposto de estacionariedade. O espectrograma (X_n) resultante da TFCP é obtido através da aplicação da TF aos termos $w(v)x_m(v)$, como mostra a Equação 24.

$$x_m(v) = x_n(v + mD), \quad v = 0, 1, \dots, T - 1; \quad m = 0, 1, \dots, M - 1 \quad (22)$$

$$w(v) = 0,54 - 0,46 \cdot \cos\left(\frac{2\pi v}{T}\right) \quad (23)$$

$$X_n(j, k) = \frac{1}{T} \sum_{v=0}^{T-1} w(v)x_m(v)e^{-\left(\frac{i2\pi jk}{T}\right)}, \quad j = 0, 1, \dots, N - 1 \quad (24)$$

De forma similar à Equação 19, para o cálculo da densidade de energia espectral em função do tempo com a TFCP (PSD , Eq. 25), costuma-se empregar um fator de normalização β tal como abaixo (Eq. 26). Note que a média dos valores quadráticos de $w_n(v)$ foi introduzida no denominador para compensar a redução de intensidade do espectro causada por w_n (WELCH, 1967). Entretanto, tal como mencionado na seção anterior (4.3.1), tais parâmetros de normalização não são considerados em análises relativas de espectro.

$$PSD(j, t) = \beta \cdot |X_n(j, k)|^2 \quad (25)$$

¹⁰ $\{v \in \mathbb{Z}^+ | 0 \leq v \leq T - 1\}$

$$\beta = \frac{2}{f_m \sum_v [w_n(v)^2]} \quad (26)$$

Uma importante limitação da TFCP é que essa força um tempo de resposta T sobre o sinal avaliado, sendo tal valor fixo para todas as faixas de frequência. Dessa forma, os eventos com recorrência de período típico maior que T podem ser subestimados. Nesse caso, o tempo de resposta atua como um filtro de baixa frequência (passa-alta). Outra limitação se dá com respeito a variações de frequência, uma vez que podem não estar adequadamente isoladas em um tempo de resposta T , que dessa vez atua como filtro passa-baixa.

4.3.3 Transformada Wavelet Contínua (TWC)

Diferentemente do TFCP, o emprego da Transformada Wavelet Contínua (TWC) permite a avaliação da distribuição de energia espectral em múltiplas resoluções. Na essência, *wavelets* fundamentam-se em funções adimensionais com propriedades particulares, que também podem ser interpretadas como critérios de admissibilidade para a definição de tal função (ψ_0). Uma dessas propriedades é a representação (ou “energia”) finita (ou “localizada”) tanto no domínio do tempo como no domínio da frequência (ver Figura 4.5). A segunda diz que “*funções wavelet devem oscilar mantendo o valor esperado nulo*”. De forma geral, essas duas propriedades são as mais importantes particularidades da transformada *wavelet*, que recebem o nome de “*ondoleta*” devido ao aspecto finito de tais funções.

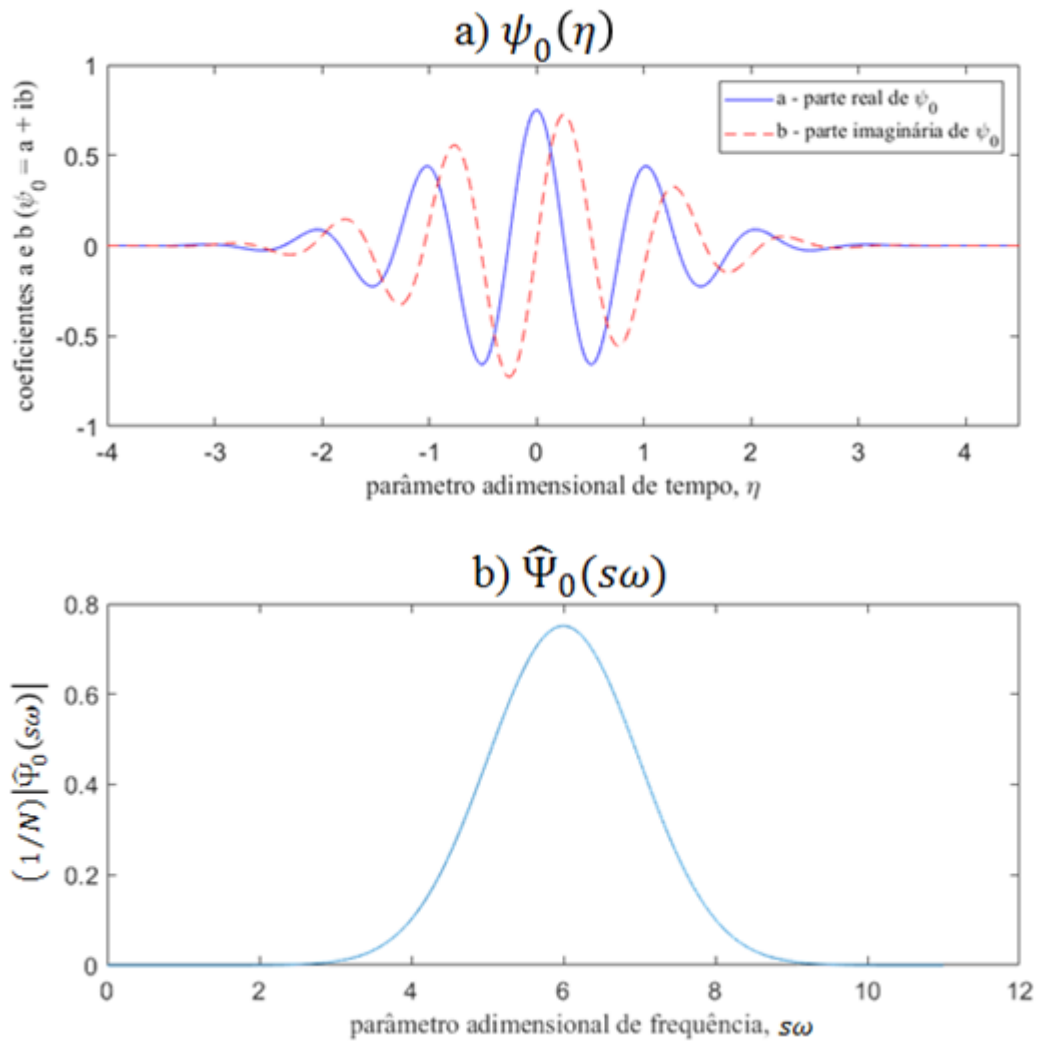
Neste estudo, foi empregada a função analítica de Morlet como *wavelet* “mãe” ψ_0 (27). Para isso, considera-se um parâmetro adimensional de frequência $\omega_0 = 6$, valor normalmente adotado e que satisfaz os critérios de admissibilidade mencionados anteriormente (JORDAN et al., 1997). A Figura 4.5 apresenta ψ_0 em função do parâmetro adimensional de tempo η , bem como sua representação no domínio da frequência $\hat{\Psi}_0$ (28), a qual é obtida por meio da TF (18) aplicada sobre ψ_0 . Mais informações teóricas e técnicas sobre esse assunto estão disponíveis em Torrence e Compo (1998).

$$\psi_0(\eta) = \pi^{-\frac{1}{4}} \exp(i\omega_0\eta) \exp\left(-\frac{\eta^2}{2}\right) \quad (27)$$

$$\hat{\Psi}_0(s\omega) = \begin{cases} \pi^{-1/4} \exp\left(\frac{-(s\omega - \omega_0)^2}{2}\right), & \omega > 0 \\ 0, & \omega \leq 0 \end{cases}$$

(28)

Figura 4.5. Wavelet Analítica de Morlet Adimensional (ψ_0).



(a) representação temporal; (b) representação no domínio da frequência. As curvas foram traçadas considerando um total de $N = 1000$ valores de η ($-5 \leq \eta \leq 5$).

Fonte: Própria.

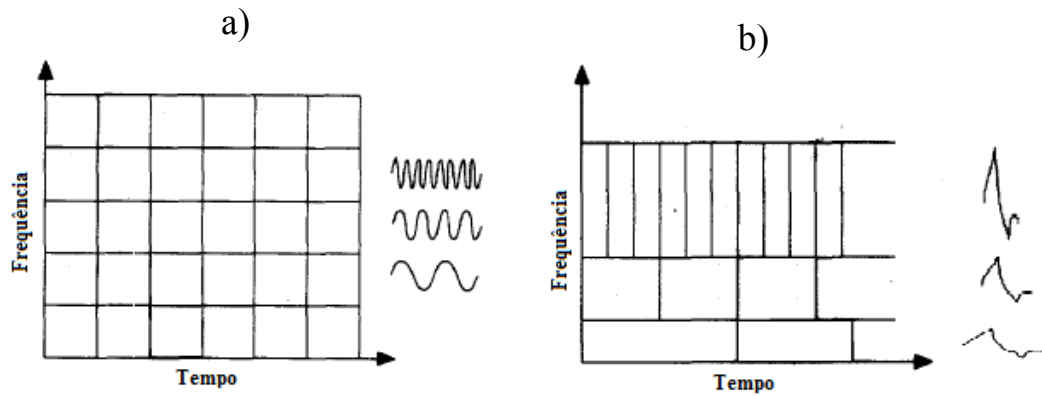
Funções variantes ψ são definidas como versões dimensionais de ψ_0 . Essa modificação ocorre por meio de uma escala (s) e de um deslocamento (τ), os quais definem curtas janelas temporais para descrever altas frequências, e janelas mais longas para baixas frequências do

espectrograma (Equação 29). O termo de normalização $s^{-1/2}$ é adicionado para que Ψ tenha energia unitária em função da escala.

$$\Psi(s, k) = \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} x(\tau) \cdot \psi\left(\frac{\tau - t}{s}\right) d\tau \quad (29)$$

A figura 4.6 ilustra as diferenças de resoluções obtidas em cada uma das abordagens, sendo a primeira com um tempo de resposta fixo (TFCP, Fig. 4.6a), e a segunda com o tempo de resposta variável (TWC, Fig. 4.6b).

Figura 4.6. Resolução de Tempo-Frequência.



(a) TFCP, tempo de resposta fixo; (b) TWC, tempo de resposta variável. As curvas localizadas à direita dos gráficos servem como exemplos ilustrativos de funções que compõem a base do sistema em diferentes faixas de frequência para cada caso (e.g., termos senoidais e *wavelet* de Daubechies).

Adaptado de: (GRAPS, 1995).

Ainda com referência à Figura 4.6, as funções que servem como base de projeção para a transformada estão representadas ao lado dos gráficos em cada caso (TFCP e TWC) sobre diferentes faixas de frequência. Note que o aumento ou a diminuição da frequência altera substancialmente o formato da base *wavelet*. Apesar disso, o efeito dessa mudança repercute de forma localizada na região de tempo-frequência, o que não acontece com a base senoidal, uma vez que essa se estende indefinidamente no domínio do tempo. Por esse motivo, a projeção viabilizada pela TWC, em tese, é mais apropriada para a avaliação de comportamentos transientes (GRAPS, 1995).

De forma geral, a TWC do sinal x_n pode ser expressa em notação discreta, tal como na Equação 30. Os parâmetros s e k indicam, respectivamente, o fator de escala e o índice de

tempo. Nessa expressão, o termo ψ_n^* indica o complexo conjugado do vetor discreto ψ_n (variante dimensional de ψ_0 ; Equação 31), sendo esse determinado por s e k .

$$\Psi_n(s, k) = \sum_{k'=0}^{N-1} x_n(k') \psi_n^* \left(\frac{(k' - k)\Delta t}{s} \right) \quad (30)$$

$$\psi_n \left(\frac{(k' - k)\Delta t}{s} \right) = \left(\frac{\Delta t}{s} \right)^{\frac{1}{2}} \psi_0 \left(\frac{(k' - k)\Delta t}{s} \right) \quad (31)$$

Assim como com a DFT, o procedimento descrito na Equação 30 pode demandar um elevado esforço computacional, entretanto, o número de operações necessárias para o cálculo dos coeficientes da transformada pode ser reduzido através de algoritmo baseado em FFT. A simplificação se fundamenta no teorema da convolução, do qual resulta que uma aproximação da TWC Ψ_n pode ser aproximada por meio da transformada inversa de Fourier (base de projeção $e^{i\omega_j k \Delta t}$) do produto entre a TF do sinal observado x_n (Equação 18) e a TF das funções *wavelet* ψ_n (Equação 33) (JORDAN et al., 1997; TORRENCE e COMPO, 1998). A frequência angular ω_j é obtida de acordo com a Equação 34. Assim como na Seção 4.3.1, o índice $\widetilde{[N/2]}$ divide o espectro na mediana do intervalo de 0 até $N - 1$. Note que, assim como discutido nas seções anteriores (4.3.1 e 4.3.2), os fatores de normalização das Equações 19, 26 e 32 serão suprimidos por análise de espectro relativo.

$$\Psi_n(s, k) = \frac{1}{N} \sum_{j=0}^{N-1} X_n(j) \hat{\Psi}_n(s\omega_j) e^{i\omega_j k \Delta t} \quad (32)$$

$$\hat{\Psi}_n(s\omega_j) = \left(\frac{2\pi s}{\Delta t} \right)^{\frac{1}{2}} \hat{\Psi}_0(s\omega_j) \quad (33)$$

$$\omega_j = \begin{cases} \frac{2\pi j}{N\Delta t} & , j = 0, 1, 2, \dots, \widetilde{[N/2]} \\ -\frac{2\pi j}{N\Delta t} & , j = \widetilde{[N/2]} + 1, \dots, N - 1 \end{cases} \quad (34)$$

A TWC também pode ser expressa em função de “pseudo-frequências” (f_p), ou valores de frequência representativos para a função *wavelet*¹¹, no lugar de escalas s . A escala e a pseudofrequência estão relacionados através da Equação 35, em que Δt é o período de amostragem e λ é chamado de “comprimento de onda de Fourier”, o qual, por sua vez, é calculado de acordo com a Equação 36 para a *wavelet* de Morlet.

$$f_p = \frac{1}{s\Delta t\lambda} \quad (35)$$

$$\lambda = \frac{4\pi s}{\omega_0 + \sqrt{2 + \omega_0^2}} \quad (36)$$

Uma importante limitação da TWC ocorre naturalmente do processo de variação de escalas. Para altos valores de s , a dilatação em ψ_n , juntamente com o deslocamento dado por valores de k suficientemente baixos ou altos as funções de transferência passam a estar definidas em regiões próximas às laterais do gráfico, denominadas “cone de influência”. Em outros termos, considere o intervalo efetivo em η , sobre o qual a função adimensional $\psi_0(\eta)$ mantém as características fundamentais de funções *wavelet*: $[-L, L]$. Com isso, pode ser demonstrado que a função modificada (ψ_n) possui um intervalo efetivo definido em $[|(k' - k)\Delta t| - sL, |(k' - k)\Delta t| + sL]$. Assim, para cada escala s , o conjunto de valores de k tal que $|(k' - k)\Delta t| \leq sL$ delimitam os coeficientes confiáveis da transformada, i.e., situados na região do cone de influência. Neste estudo, os coeficientes situados fora dessa região são desconsiderados durante a análise espectral.

4.3.4 Complementaridade Espectral

Como mencionado no início da Subseção 4.3, a complementaridade neste estudo é avaliada em termos da representatividade que cada modelo apresenta sob a perspectiva do domínio da frequência, cujas transformadas (Z) são estimadas, neste trabalho, por meio das Equações 18 (Seção 4.3.1), 24 (Seção 4.3.2) e 32 (Seção 4.3.3). Para isso, considera-se o cálculo do produto entre as magnitudes das transformadas Z_X e Z_Y , em que “ X ” representa cada uma das saídas dos modelos e “ Y ” a série temporal observada (Equação 37). O significado do cálculo proposto é similar ao da densidade de energia espectral cruzada entre os sinais, uma vez que

¹¹ Cada função *wavelet* está associada a valores de frequência, diferentemente de funções de onda harmônica da TF, que possuem um único valor de frequência associado.

ênfatiça, para cada faixa de frequência f , as amplitudes comuns aos dois sinais. Na Equação 37, o acento em $\tilde{Z}$ indica a normalização pelo máximo em Z_Y , para que seja estabelecida uma base comparativa entre os resultados obtidos através da TF, TFCP e TWC, bem como para cada modelo. É importante observar que, por esse motivo, os valores do espectro avaliado são adimensionais (relativos). Além disso, como artifício para facilitar a inspeção, aplica-se a raiz de quatro grau para ressaltar os menores valores, mantendo a importância relativa entre magnitudes espectrais.

$$A_{XY}(f) = (|\tilde{Z}_X(f)| \cdot |\tilde{Z}_Y(f)|)^{1/4} \quad (37)$$

Assim, propõe-se neste trabalho que a complementaridade seja avaliada por meio da comparação absoluta (Eq. 38) e relativa (Eq. 39) entre $A_{XY}(f)$ e $A_{YY}(f)$. Note que, assim como o subscrito “XY” faz referência ao aspecto cruzado entre saídas de modelos e observações, o subscrito “YY” faz referência ao auto espectro do sinal observado (que corresponde à substituição dos termos “X” na Equação 37).

$$E_A(f) = |A_{XY}(f) - A_{YY}(f)| \quad (38)$$

$$E_R(f) = \frac{|A_{XY}(f) - A_{YY}(f)|}{A_{YY}(f)} \quad (39)$$

Em caráter inicial, esse estudo está voltado para a avaliação do aspecto global dos espectros obtido por meio das transformadas propostas (Subseções 4.3.1-3). Dito isso, considera-se a aplicação da TFCP e TWC como artifício para a suavização do espectro global através do cálculo da média aritmética sobre a dimensão temporal (e.g., WELCH, 1967). É importante ressaltar que as suavizações conferidas pela TFCP e TWC levam em consideração os efeitos transientes estimados.

O cálculo do Erro Espectral Absoluto (EA, Eq. 38) permite avaliar a aptidão geral dos modelos em cada faixa de frequência f , uma vez que considera a diferença absoluta entre o espectro estimado (A_{XY}) e o espectro observado (A_{YY}). Em relação ao sinal observado, os valores de E_A enfatizam diferenças em faixas de críticas de frequência, ou seja, faixas associadas com as amplitudes espectrais mais expressivas (ou de maior energia espectral associada).

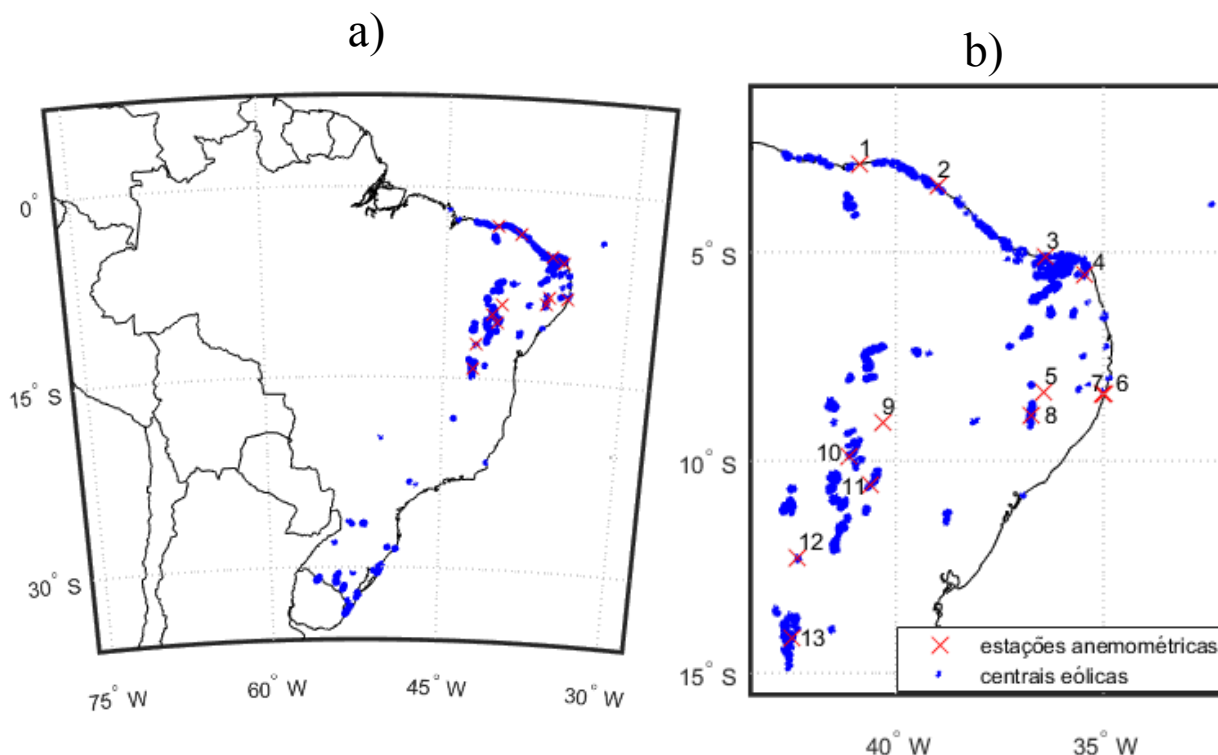
O Erro Espectral Relativo (ER, Eq. 39), por sua vez, destaca aspectos mais sutis da aptidão dos modelos presentes em faixas de frequência menos expressivas (com menor energia espectral associada). Nessas faixas, o valor de diferença absoluta é significativamente menor, assim, informações relevantes podem ser subaproveitadas.

Essa análise pode ainda ser expandida para a dimensão temporal, por meio dos coeficientes da TFCP e TWC, funcionalidade particularmente útil em abordagens adaptativas (modelos com parâmetros variáveis).

5 RESULTADOS E DISCUSSÃO

Resultados para 13 estações anemométricas distribuídas na Região Nordeste do Brasil foram obtidos e são apresentados nesse estudo (Figura 5.1). Dois casos destacados nessa seção: (a) Caso 1, local cujos modelos apresentam, em geral, os maiores valores de SS4 (Eq. 1); (b) Caso 2, local com o modelo que apresenta o maior aumento no valor de SS4 relativo ao da interpolação (técnica de referência). Resultados gerais de demais localizações serão apresentadas em Apêndice B, no final do texto.

Figura 5.1. Localização de Estações Anemométricas e de Centrais Eólicas no Brasil.



(a) Mapa do Brasil; (b) Zoom na Região Nordeste. Os pontos azuis representam centrais eólicas e os índices correspondem às localizações das estações anemométricas empregadas neste estudo. A Tabela 1 fornece mais informações a respeito dos índices indicados no mapa.

Fonte: Própria.

O comportamento do vento é modelado para alturas típicas de turbinas eólicas em função dos dados observacionais considerados para cada estação anemométrica (Tabela 5.1). Três instituições fornecem dados de observação do vento: Secretário de Infraestrutura do Estado do Ceará (SEINFRA-CE¹²), Instituto Nacional de Pesquisa Espacial (INPE¹³) e

¹² disponível em: <<http://www.seinfra.ce.gov.br/index.php/downloads/category/6-energia>>.

¹³ disponível em: <<http://sonda.ccst.inpe.br/>>.

Operador do Sistema Brasileiro (ONS¹⁴), que também é um parceiro especial no projeto HPC4E. Além disso, também foram empregados dados fornecidos pela Universidade Federal de Pernambuco, derivados do Projeto Pilacas (2014).

Um procedimento de garantia de qualidade é realizado sobre os dados observacionais com vistas à identificação de dados ausentes e comportamentos potencialmente anômalos nas séries temporais originais (SHAFER ET AL., 2000; GRAYBEAL ET AL., 2004; JIMÉNEZ ET AL., 2010; MORAES, 2015). Para esse estudo, foram aplicados aos dados testes globais (que avaliam o comportamento geral) e locais (que avaliam os instantes ou subperíodos individualmente). Os instantes das séries temporais identificados potencialmente anômalos foram desconsiderados nessa análise. O percentual de dados ausentes e potencialmente anômalos identificado para cada uma das bases de dados empregadas no Caso 1 e no Caso 2 está distribuído ao longo de todo o período considerado em cada base de dados (estudos de caso), uma vez que foram subdivididas (i.e., analisadas separadamente) as séries temporais que apresentaram um intervalo contínuo com um percentual de dados ausentes ou potencialmente anômalos maior que 10%. Deve-se ressaltar, entretanto, que a remoção sistemática de dados pode impactar na representatividade dos espectros estimados, os quais estão relacionados com a análise de complementaridade espectral proposta. Todavia, em cada caso estudado, os modelos foram submetidos a uma mesma base de dados (i.e., sujeitos às mesmas limitações em relação a uma possível má condição dos dados), e a comparação entre os espectros estimados se dá com respeito a aspectos relativos (adimensionais), estando, com isso, fora do escopo desse trabalho a avaliação da representatividade das componentes espectrais obtidas.

A primeira parte (Seção 5.1) apresenta os resultados das saídas individuais e saídas combinadas sem análise de complementaridade (ver Figura 2). Os resultados foram obtidos para as 13 estações anemométricas apresentadas no início da Seção 4. Dois casos são destacados e detalhados na Seção 5.1. O Caso 1 trata dos melhores resultados, tanto em termos de SS4, quanto em termos do aprimoramento relativo ao IBL, em escala horária. O Caso 2 apresenta os resultados que atingiram o maior aprimoramento do SS4 com respeito ao IBL em regime horário e diário.

A segunda parte (Seção 5.2) apresenta os resultados preliminares obtidos para a análise de complementaridade espectral para os mesmos casos apresentados na Seção 5.1. Tais resultados são verificados em função dos espectros cruzados e coerência espectral entre os diversos sinais intercomparados (preditores e preditando).

¹⁴ mais informações em: <ons.org.br/>

Tabela 5.1. Informações sobre Estações Anemométricas Ilustradas na Figura 5.1.

Índice da Estação Anemométrica	Altura (m) (AGL)	Fonte de Dados
#1	60,4	SEINFRA-CE
#2	60,4	SEINFRA-CE
#3	80,0	ONS
#4	89,0	ONS
#5	50,0	INPE
#6	42,0	UFPE
#7	60,0	UFPE
#8	80,0	ONS
#9	50,0	INPE
#10	80,0	ONS
#11	80,0	ONS
#12	80,0	ONS
#13	80,0	ONS

Fonte: Própria.

Dentre as 13 estações anemométricas consideradas neste estudo, duas foram selecionadas para uma discussão detalhada dos resultados (Figura 5.1). O Apêndice, no final do texto, apresenta de forma resumida os resultados para as demais estações anemométricas.

- Caso 1, estação anemométrica #1, cujos dados são fornecidos pela Secretaria de Infraestrutura do Estado do Ceará (SEINFRA-CE). Os resultados desse caso atingem valores de SS4 que estão dentre os mais altos. Além disso, esta estação também apresenta o maior aprimoramento sobre a interpolação no geral, para o caso horário. Por esses dois motivos, essa estação anemométrica foi adotada para o Estudo de Caso 1 (Seção 5.1). Um total de 1,11% dos dados foi identificado como ausentes ou potencialmente anômalos durante o processo de garantia de qualidade.
- Caso 2, estação anemométrica #3, dados fornecidos pelo Operador Nacional do Sistema (ONS). Nesse caso, os modelos obtêm o maior aumento relativo ao SS4 do modelo de interpolação IBL tanto horário, quanto em regime diário. Um total de 14,91% dos dados foi identificado como ausente ou potencialmente anômalo durante o processo de garantia de qualidade.

Todos os modelos apresentados na Seção 4.1 nos dois casos selecionados serão avaliados. Neste sentido, duas escalas para a resolução temporal estão sendo abordadas: 1) escala horária (considerando apenas os instantes de tempo a cada 6 horas, 00Z 06Z 12Z 18Z); 2) escala diária (considerando apenas a média diária). Além disso, os detalhes sobre o melhor modelo para cada caso e cada escala de tempo são mostrados (na Seção 5.1 para o Caso 1, e no Anexo A para o Caso 2).

A qualidade das estimativas podem ser quantificadas com respeito a quatro características estatísticas principais: 1) a amplitude das variações (por meio de desvios padrão); 2) estrutura global de frequência/fase (por meio da correlação entre as saídas do modelo e a observação). 3) raiz do erro quadrático médio das anomalias (RMSD¹⁵, “*Root-Mean-Square-Deviation*”); 4) índice de habilidade (*skill-score*; e.g., SS4). Tal como proposto por Taylor (2001), a relação entre as três primeiras¹⁶ permite a visualização (simultânea) em um gráfico de coordenadas polares (conhecidos como “diagramas de Taylor”). Os diagramas de Taylor são adotados aqui como ferramenta para a avaliação da qualidade geral das saídas de modelo com respeito às observações (Figuras 5.2 e 5.9, para resultados em escala horária; Figuras 5.5 e 5.10, para resultados em escala diária).

Também serão mostrados gráficos de dispersão, histogramas de erro e superposição de séries temporais (sob perspectivas global e locais), uma vez que o diagrama de Taylor não é apropriado para a avaliação de “*bias*”, ou tendências de superestimação ou subestimação do valor esperado.

É importante mencionar que os gráficos de superposição apresentados têm na abscissa (dimensão temporal), índices temporais ou “*time-steps*”. Note que há um percentual de dados ausentes ou removidos por método de garantia de qualidade, portanto, a sucessão de índices da série temporal pode representar um passo temporal maior que o período de amostragem empregado (i.e., os instantes que correspondem a dados ausentes ou removidos são omitidos). Tais dados ausentes impactam não apenas a parametrização de modelos, como também a representatividade espectral do sinal. Entretanto, de forma pragmática, neste estudo, tal efeito é desprezado, uma vez que a análise se baseia em espectros relativos e que o número de dados removidos, em geral, é pequeno. Ademais, normalmente são empregadas séries temporais com um percentual considerável de períodos ausentes para ajustes de modelos estatísticos, os quais

¹⁵ $RMSD = \left\{ \frac{1}{N} \sum_{i=1}^N [(y_i - \bar{y}) - (\hat{y}_i - \bar{\hat{y}})]^2 \right\}^{1/2}$, na qual N – número de amostras; y e $\hat{y}$ – vetores com N amostras observadas e estimativas, respectivamente; $\bar{y}$ e $\bar{\hat{y}}$ – valores médios de y_i e $\hat{y}_i$, respectivamente.

¹⁶ É possível demonstrar que relação entre a correlação, desvios padrão e RMSD é análoga à expressão da lei dos cossenos: $RMSD_{xy}^2 = \sigma_x^2 + \sigma_y^2 - 2\sigma_x\sigma_y\text{corr}_{xy}$. Essa relação motivou o autor a criar o diagrama.

podem ser posteriormente combinados. O pressuposto da validade de análises relativas de espectro está baseada no fato de que todos os modelos são parametrizados sobre o mesmo banco de dados, havendo equivalência com respeito às limitações dos dados durante o ajuste de cada modelo. Entretanto, é importante ressaltar que diferentes modelos, com diferentes níveis de complexidade, podem ser capazes de assimilar a característica geral dos dados de maneira diferente, sendo esse um objeto de estudo nesse trabalho.

Resultados gerais das demais estações anemométricas, além daquelas selecionadas para os Casos 1 e 2, são mostrados no Apêndice B.

5.1 Estudo de Caso 1: Melhores Performances Globais

Com já mencionado, esta subseção apresenta resultados preliminares do Caso 1, nos quais os modelos estão atingindo os melhores resultados de forma geral (em critério estabelecido através de SS4). O local está representado na Figura 5.1 pelo índice #1. O procedimento de validação cruzada divide o conjunto de dados completo em dois períodos distintos para treinamento e validação de todos os modelos. O período adotado para o treinamento dos modelos data desde o início das amostras até o índice que compreende $2/3$ do comprimento completo da sequência temporal N . A validação, por sua vez, é realizada sobre o período que data desde o término do intervalo de treinamento até o instante final (contemplando aproximadamente $N/3$ amostras). Para o caso 1, o conjunto de treinamento se estende desde 18 de agosto de 2004 até 11 de outubro de 2005; O conjunto de validação abrange o período de 11 de outubro de 2005 até 8 de maio de 2006 (observe que todos os resultados nesta subseção - e no próximo - referem-se apenas ao período de validação).

A Figura 5.2 apresenta os resultados do Caso 1, em escala horária (médias tomadas a cada 6 horas), por meio de um diagrama de Taylor, como mencionado anteriormente. Para ler tal diagrama, considere os pontos distribuídos no plano polar da figura, os quais representam séries temporais de velocidade do vento na posição da estação anemométrica considerada. A disposição dos pontos no diagrama se dá em função das propriedades estatísticas das séries temporais de saída dos modelos (pontos enumerados) relativas à série temporal observada (ponto rotulado como “obs”). Note que as distâncias radiais com respeito à origem são proporcionais ao desvio padrão das séries temporais. A posição azimutal indica o coeficiente de correlação entre uma estimativa e as observações, indicado no gráfico pelo cosseno do ângulo

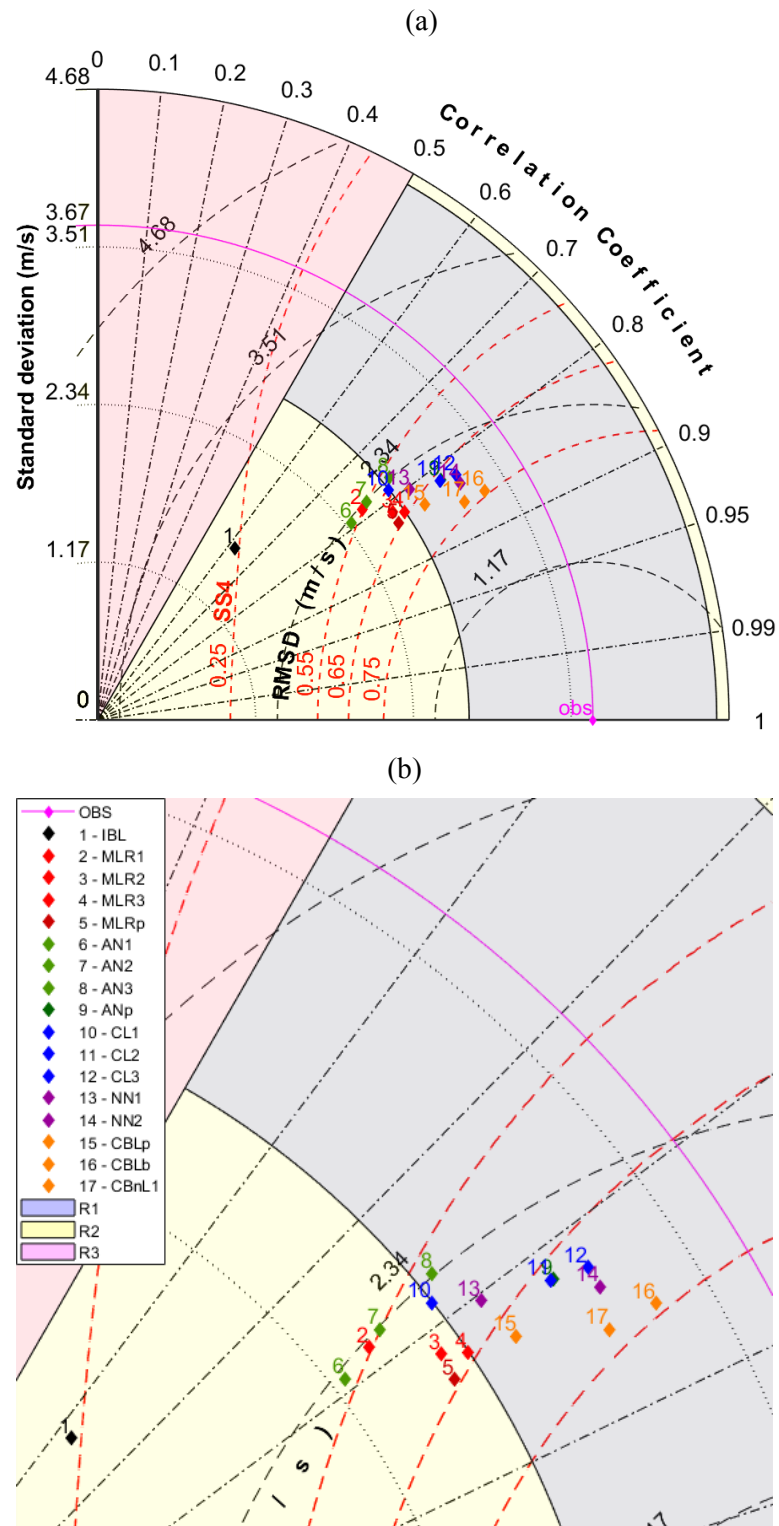
formado entre uma reta contendo a origem e a abcissa. O RMSD é proporcional às distâncias radiais centradas no ponto de referência (observações).

De forma similar ao RMSD, o SS4 quantifica a qualidade geral das estimativas levando em consideração o desvio padrão reproduzido normalizado pelo desvio padrão observado, bem como a correlação entre os sinais (estimado e observado). Note, porém, que o valor resultante do cálculo do SS4 é adimensional, em contraste com o RMSD, que possui a mesma unidade dos sinais intercomparados. Note ainda que as isolinhas de SS4 e RMSD possuem sensibilidades diferentes com relação às características mencionadas. Comparando as curvas, é possível observar que o SS4 possui maior sensibilidade à razão entre os desvios sobre a correlação, quando comparada à sensibilidade circular do RMSD.

As saídas dos modelos baseados em domínios com malhas de 4, 16 e 36 pontos referem-se ao nível que atinge o maior SS4 para cada modelo (ver Tabela 2). Além disso, três regiões são definidas para avaliar a habilidade dos modelos (R1, R2 e R3). A primeira região (R1) limita os resultados de correlação acima de 0,5 e os desvios padrão reproduzidos entre 75% e 125% do desvio padrão observado. A segunda região (R2) inclui todas as saídas com correlação acima de 0,5 e desvios padrão inferiores a 75% ou superiores a 125% do desvio padrão observado. A terceira região (R3) demarca resultados com coeficiente de correlação inferior a 0,5.

A Figura 5.2 mostra que todas as saídas de modelo produziram valores de correlação similares (entre 0,7 e 0,9), exceto pelo IBL, que apresenta um SS4 da ordem de 0,25 (correlação aproximadamente 0,6 e razão entre os desvios em um pouco menos que 0,5). É possível observar que, de forma geral, todos os modelos implementados apresentam resultados superiores aos do IBL tanto em termos de variância reproduzida, como em termos de correlação. Isso significa que, nesse caso, há uma melhora significativa com respeito à correlação e à variância reproduzida (por exemplo, entre IBL e CBLb). Resumidamente, é possível observar que as técnicas baseadas em agrupamento, redes neurais e combinação estão situadas na região R1 do diagrama (CL2, CL3, NN1, NN2, CBLp, CBLb, CBnL1, CBnL2), região que comporta as saídas que melhor descrevem as observações. Os modelos “MLR” estão apresentando resultados ligeiramente inferiores. Os demais modelos estão situados na zona R2, indicando uma perda maior na variância reproduzida.

Figura 5.2. Caso 1: Diagrama de Taylor (escala horária).



(a) intercomparação entre todos os modelos na Tabela 4.1 e combinações, a figura ao lado fornece uma legenda;
(b) zoom em modelos com maior SS4. Fonte: Própria.

A Figura 5.2 também mostra que os modelos baseados em análogos, AN1, AN2 e AN3 estão sendo menos capazes de reproduzir a variância de observações, em seguida, os modelos baseados em MLR sobre domínios equivalentes (Seção 4.1.2; MLR1, MLR2 e MLR3). Sabe-

se que os modelos baseados em análogos normalmente exigem longas séries temporais para garantir que análogos sejam bem definidos (LORENTZ, 1969; RUOSTEENOJA, 1988; VAN DEN DOOL, 1994). Nesse quesito, é importante notar que a curta extensão temporal do conjunto de treinamento utilizado (cerca de um ano) pode estar relacionada, de forma geral, à baixa variância reproduzida pelos modelos AN. Apesar disso, o ANp está mostrando uma pequena melhora em relação à variância reproduzida do MLR3, porém, com menor fator de correlação.

É importante ressaltar que a escolha dos períodos de calibração e validação em uma estratégia de validação cruzada pode impactar significativamente os resultados (especialmente no caso de modelos menos robustos). Portanto, um estudo voltado para a escolha de períodos para estratégias de validação cruzada é fundamental. A título de exemplo, verificou-se uma variância ligeiramente superior para os dados observacionais e 85% de todos os preditores considerados na validação quando comparados aos dados de calibração correspondentes, fator que pode estar relacionado à menor habilidade dos modelos baseados em PCA (MLRp, ANp e CBLp).

No que concerne os modelos baseados em agrupamento, a Figura 5.2 mostra que os modelos CL1, CL2 e CL3 obtiveram maiores valores de SS4 quando comparados aos modelos de regressão correspondentes (MLR1, MLR2 e MLR3). A esse respeito, resultados similares foram obtidos em escala diária (Figura 5.5). Isso é esperado, considerando que o método de agrupamento especializa parâmetros para períodos específicos, o que, em teoria, reduz a influência de eventos díspares quando comparado com a aplicação de parâmetros globais unicamente. A Tabela 5.2 fornece as configurações resultantes do procedimento de otimização proposto na Seção 4.1.5, para o Caso 1, em escala horária (Tabela 5.2a) e diária (Tabela 5.2b). Como já mencionado (Seção 4), as configurações finais são estabelecidas através do maior valor de SS4 para cada nível.

Em geral, os resultados foram similares com respeito ao número de classes formados, ao limiar T , e aos níveis do GCM que produzem os melhores resultados, exceto pelo CL3 (em regime diário). Entretanto, o número de classes aplicadas é similar entre o CL2 e CL3. O CL1 apresenta os maiores números de classes aplicáveis, o que pode estar relacionado com a extensão do domínio aplicado. Isso significa dizer que um número reduzido de variáveis próximas ao local de interesse possui maior potencial de classificação, produzindo números maiores de classes aplicáveis. Note, no entanto, que o CL2 possui um número de clusters aplicados menor que o CL3, cujo domínio é mais amplo. Assim, faz-se necessário um estudo mais aprofundado a esse respeito. Por outro lado, de fato os resultados mostram que a escolha

de domínios mais amplos, com um número de classes especializadas reduzido tende a produzir melhores resultados.

Tabela 5.2. Caso 1: configurações resultantes para CL1, CL2 e CL3.

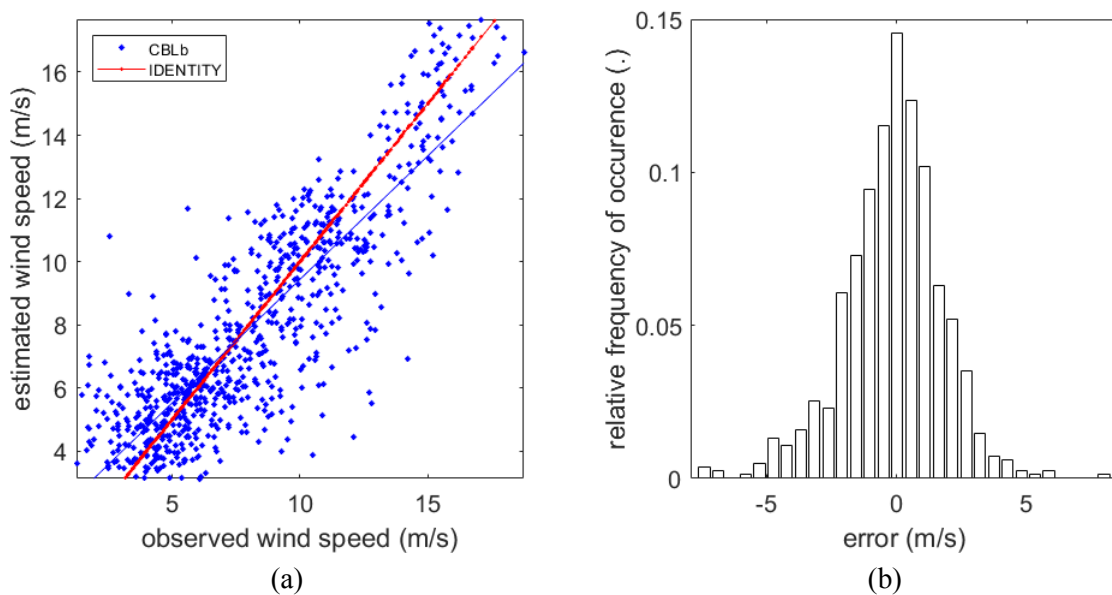
a) Escala Horária	CL1	CL2	CL3
Distância (AHC)	Correlação	Correlação	Cosseno
Algoritmo (AHC)	VP	Vínculo Médio	VD
Distância (NHC)	<i>Manhattan</i>	Cosseno	-
Classes Formadas	74	86	80
Classes Aplicadas	15	3	5
Limiar (T)	2%	4%	4%
Nível do GCM	51	53	53
b) Escala Diária	CL1	CL2	CL3
Distância (AHC)	Correlação	Euclideana	Correlação
Algoritmo (AHC)	VP	VD	Ward
Distância (NHC)	Euclideana	-	-
Clusters Formados	94	80	6
Clusters Aplicados	8	1	3
Limiar (T)	2%	4%	18%
Nível do GCM	10m	53	57

(a) escala horária; (b) escala diária. Mais detalhes sobre as distâncias e algoritmos estão disponíveis na Tabela 3.
Fonte: Própria.

Ainda acerca da Figura 5.2, mostra-se que os modelos de regressão não linear NN1 e NN2 superam os modelos MLR correspondentes (MLR1 e MLR2) em termos da variância reproduzida e da correlação, sendo NN2 o que obteve o melhor desempenho dentre de todos os modelos não combinados. Em relação aos modelos baseados em combinação linear, CBLp está mostrando desempenho semelhante ao MLR3 (Figura 5.2b). Por sua vez, CBLb fornece o melhor resultado geral com referência à correlação, à razão entre os desvios e, conseqüentemente, ao SS4, atingindo um valor de 0,7412 para esse último. Esse valor representa uma melhoria de 159,61% sobre a pontuação de SS4 do modelo de interpolação (IBL). Estes resultados da CBLb estão sendo obtidos combinando MLR1, MLRp, ANp, CL1, CL2, CL3 e NN2 (ver Seção 4.2).

As Figuras mostradas em seguida, nesta subseção, detalham os resultados do modelo CBLb no tocante às possíveis tendências do valor esperado das estimativas. A Figura 5.3a apresenta os resultados em diagramas de dispersão (“*scatter plot*”) entre as saídas do CBLb e as observações. Observe que os resultados e observações dos modelos estão fortemente correlacionados, como pode ser observado no ângulo entre as retas de identidade em vermelho (a ordenada contém os valores observados). Também é possível observar na figura uma tendência equilibrada entre subestimação e superestimação do valor esperado, apesar de haver uma dispersão expressiva na faixa de valores estimados entre 4 e 8 m/s. Nessa faixa, observa-se uma tendência de subestimação das observações. Por outro lado, no intervalo entre 8 e 12 m/s, alguns valores foram significativamente superestimados. O equilíbrio entre valores sub e superestimados é mais evidente na Figura 5.3b ao lado. Nessa, o histograma de erro¹⁷ apresenta uma distribuição concentrada de forma aproximadamente homogênea em torno do valor nulo. Em outras palavras, apesar de apresentar variância reduzida, CBLb reproduz satisfatoriamente o valor esperado das observações.

Figura 5.3. Caso 1: Diagramas de Dispersão e Histograma de Erros (escala horária).



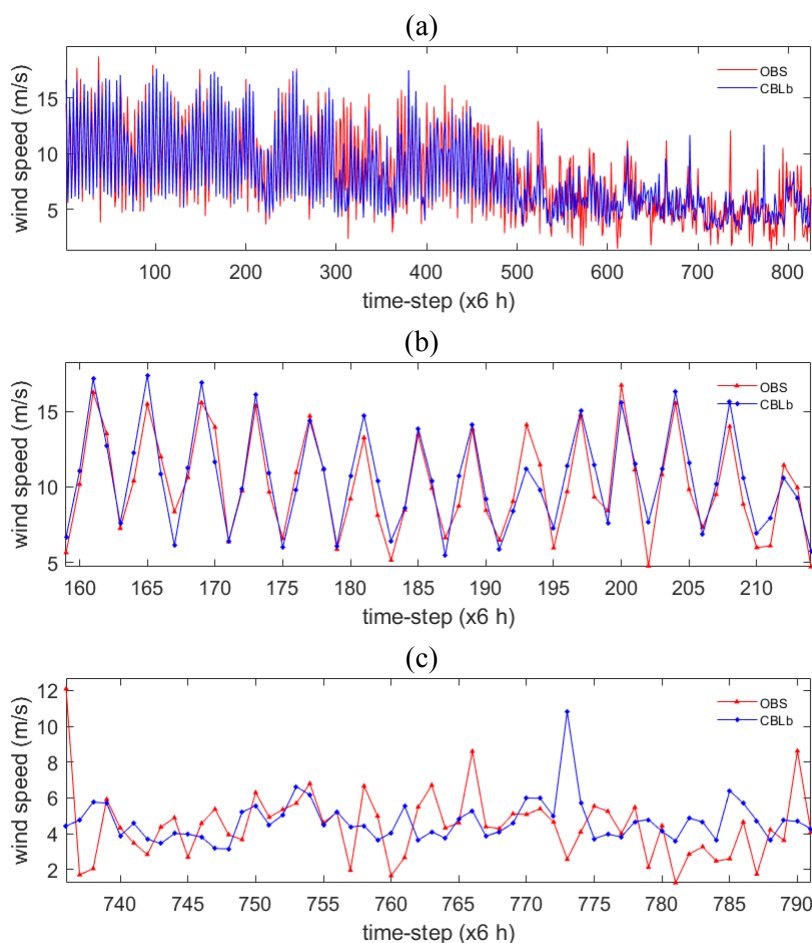
(a) dispersão entre saídas de modelo com maior SS4 (CBLb, na ordenada) e observações (abscissa); (b) histogramas de erro de CBLb (frequência relativa de ocorrência versus erro). Fonte: Própria.

A Figura 5.4 mostra a série temporal derivada do modelo CBLb em três janelas temporais. A primeira, Figura 5.4a, abrange toda a série temporal (sob o período de validação).

¹⁷ O erro está sendo definido como $e_t = \hat{y}_t - y_t$, onde $\hat{y}_t$ – estimativa no instante t ; y_t – observação no mesmo instante.

Essa figura mostra que, em geral, as estimativas descrevem satisfatoriamente a periodicidade de médio e longo prazo do vento local observado. As outras janelas (Figuras 5.4b e 5.4c) mostram exemplos de momentos em que o CBLb está bem ajustado e mal ajustado às observações, respectivamente. Em síntese, CBLb é adequado para aplicações como a avaliação de recursos eólicos, mas ainda há limitações relacionadas com as aplicações em alta frequência, como previsão de curto prazo.

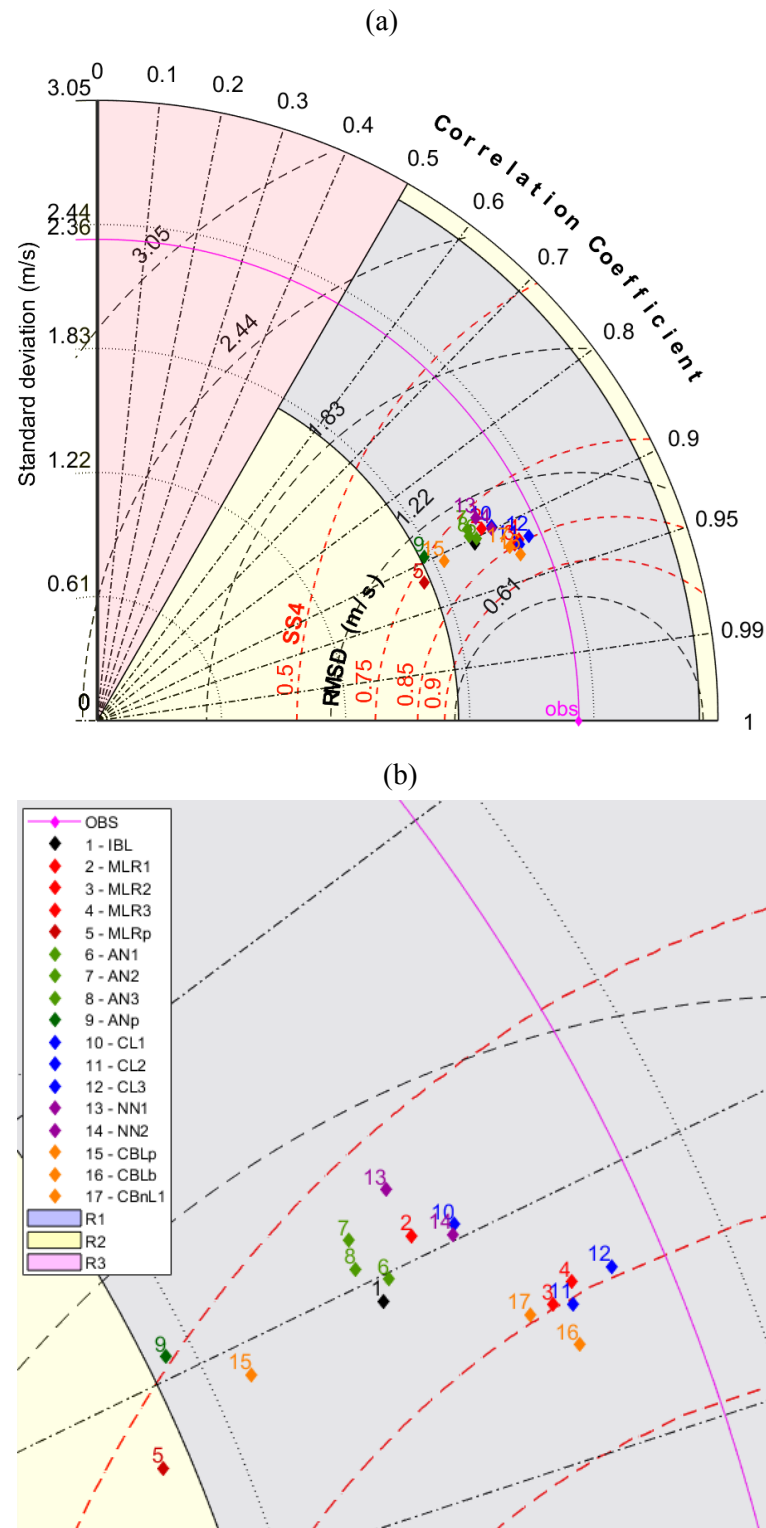
Figura 5.4. Caso 1: Superposição de Séries Temporais (escala horária).



(a) séries temporais do modelo com maior SS4 (CBLb) e observações; (b) zoom em uma janela com um bom ajuste entre modelo e observação; (c) zoom em uma janela com um ajuste ruim entre modelo e observações.

Fonte: Própria.

Figura 5.5. Caso 1: Diagrama de Taylor (escala diária).



(a) intercomparação entre todos os modelos na Tabela 4.1 e combinações, a figura ao lado fornece uma legenda;
(b) zoom em modelos com maior SS4. Fonte: Própria.

A Figura 5.5 mostra os resultados obtidos para o Caso 1 em regime diário (médias tomadas a cada 24 horas). No geral, os modelos estão atingindo níveis mais altos de SS4 e valores mais baixos de RMSD. Todos os modelos estão na região do diagrama R1, exceto o

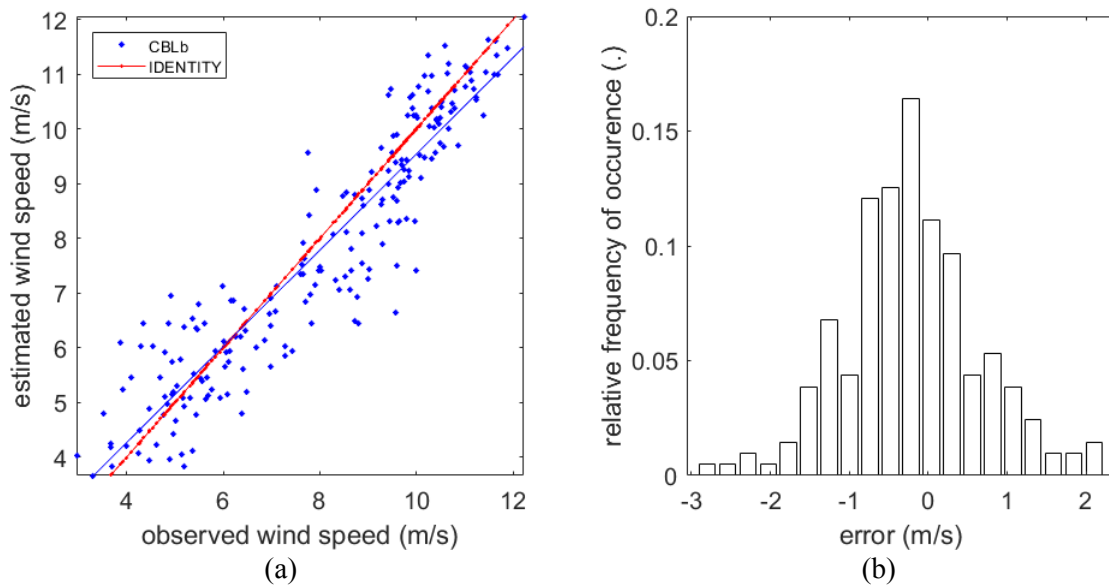
MLRp, que está situado zona de fronteira entre R1 e R2, o que indica uma subestimação em aproximadamente 25% do desvio padrão das observações (também próximo ao ANp). Um aumento na previsibilidade do sinal ocorre devido ao filtro de altas frequências (passa-baixa) aplicado durante o processo de conversão do intervalo de integração para a base diária (e.g., média móvel). Em regime diário, a interpolação apresenta uma performance superior a alguns modelos mais complexos em termos de correlação com as observações (e.g., AN1, AN2, AN2, ANp, MLR1, NN1 e NN2) e de variância reproduzida (CBnL, CBLp, ANp e MLRp).

Com base na mesma discussão apresentada para modelos baseados em PCA (Seção 4.1.3) em escala horária, CBLp, MLRp e ANp estão apresentando uma redução significativa na habilidade em reproduzir a variância observada, o que pode ocorrer devido à maior sensibilidade às características estatísticas do conjunto de calibração (por exemplo, variância e valor esperado). A esse respeito, a base diária apresenta maiores valores de variância para período de validação (em contraste com o período de treinamento) em 90,04% dos preditores, bem como na observação, o que indica uma mudança no comportamento de médio e longo-prazo durante o período de validação. Mudanças nas características estatísticas nas séries temporais podem comprometer o processo de assimilação modelos estatísticos simples, bem como invalidar parametrizações pouco robustas. Por outro lado, a interpolação possui dependência apenas com as distâncias geográficas e com os preditores no instante para o qual se deseja estimar o comportamento do vento.

Também em base diária, o melhor resultado geral está sendo obtido pelo CBLb, cujo SS4 atinge a marca de 0,8653. Esse valor representa uma melhoria de 6,29% em relação à pontuação de habilidade IBL. Os resultados estão sendo obtidos através da combinação dos seguintes modelos: MLR1, MLR2, MLR3 e CL2.

A Figura 5.6 ilustra as estimativas diárias de CBLb em contraste com as observações em um diagrama de dispersão (Fig. 5.6a) e um histograma de erro (Fig. 5.6b) diárias produzidas pelo CBLb. Como no caso da escala horária, essas figuras sugerem que a perda de variância, na prática, ocorre com estabilidade no valor esperado, o qual é bem representado. Além disso, apesar das limitações do ponto de vista local, CBLb exibe bom desempenho em relação à estrutura de frequência/fase e amplitude de variação (ver Figuras 5.7a, 5.7b e 5.7c).

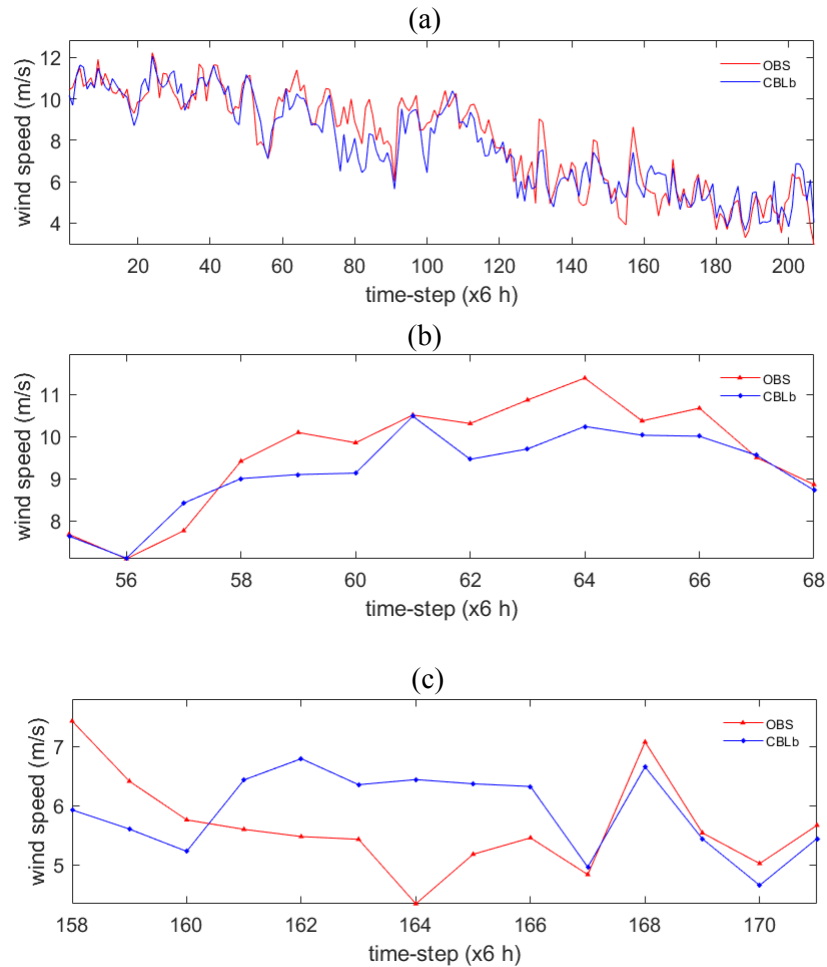
Figura 5.6. Caso 1: Diagramas de Dispersão e Histograma de Erros (escala diária).



(a) dispersão entre saídas de modelo com maior SS4 (CBLb, na ordenada) e observações (abscissa); (b) histogramas de erro de CBLb (frequência relativa de ocorrência versus erro). Fonte: Própria.

A Figura 5.8 apresenta os resultados da análise de complementaridade espectral para o Estudo de Caso 1, considerando as três abordagens propostas para a análise no domínio do tempo (baseadas em: TF, TFCP e TWC). O estudo foi realizado sobre as séries temporais em regime horário, tendo em vista que múltiplas escalas de variação temporal, inclusive a escala diária, pode ser avaliada considerando um período de amostragem menor (e.g., a cada 6 horas). Em face da discussão apresentada na Seção 4.3, note que a frequência máxima nesse caso é de duas ocorrências por dia, como mostra a figura (2 dia^{-1}). Para cada transformada implementada, são mostrados os Erros Espectrais Absolutos (EA; Eq. 38) e Erros Espectrais Relativos (ER; Eq. 39). Mais detalhes sobre esses parâmetros estão disponíveis na Seção 4.3.4. Note que, tal como mencionado anteriormente, E_A e E_R são adimensionais, em decorrência da normalização realizada sobre os coeficientes das transformadas.

A Figura 5.8 mostra que há uma correspondência entre as três abordagens propostas (TF, TFCP e TWC) com respeito às características do espectro global reproduzidas em cada saída de modelo. A abordagem baseada na TF produz uma imagem com aspecto mais ruidoso, o que dificulta a análise visual (Fig. 5.8a e 5.8b). É importante ressaltar que a aplicação de médias em torno de faixas de frequência pode comprometer a representatividade do espectro. A suavização obtida por meio da TFCP e TWC, por outro lado, apresenta um espectro com aspecto mais contínuo, sendo assim mais apropriados para análises semi-objetivas.

Figura 5.7. Caso 1: Superposição de Séries Temporais (escala diária).

(a) séries temporais do modelo com maior SS4 (CBLb) e observações; (b) zoom em uma janela com um bom ajuste entre modelo e observação; (c) zoom em uma janela com um ajuste ruim entre modelo e observações.

Fonte: Própria.

No caso da TF, considerou-se os valores médios e intervalos de frequências (“bins”) de $0,025 \text{ dia}^{-1}$, tendo em vista que as magnitudes do espectro global obtidos diretamente através da TF costumam apresentar uma dispersão elevada. O comprimento do intervalo foi estabelecido de forma pragmática no intuito de simplificar a representação visual, mantendo a correspondência com os demais resultados (TFCP e TWC). A respeito da TWC, por sua vez, pode ser visualizado nas Figuras 5.8e e 5.8f que valores de frequência próximos a zero foram desconsiderados. Isso se deve aos coeficientes da TWC fora da zona do cone de influência, os quais foram desconsiderados (Seção 4.3.3).

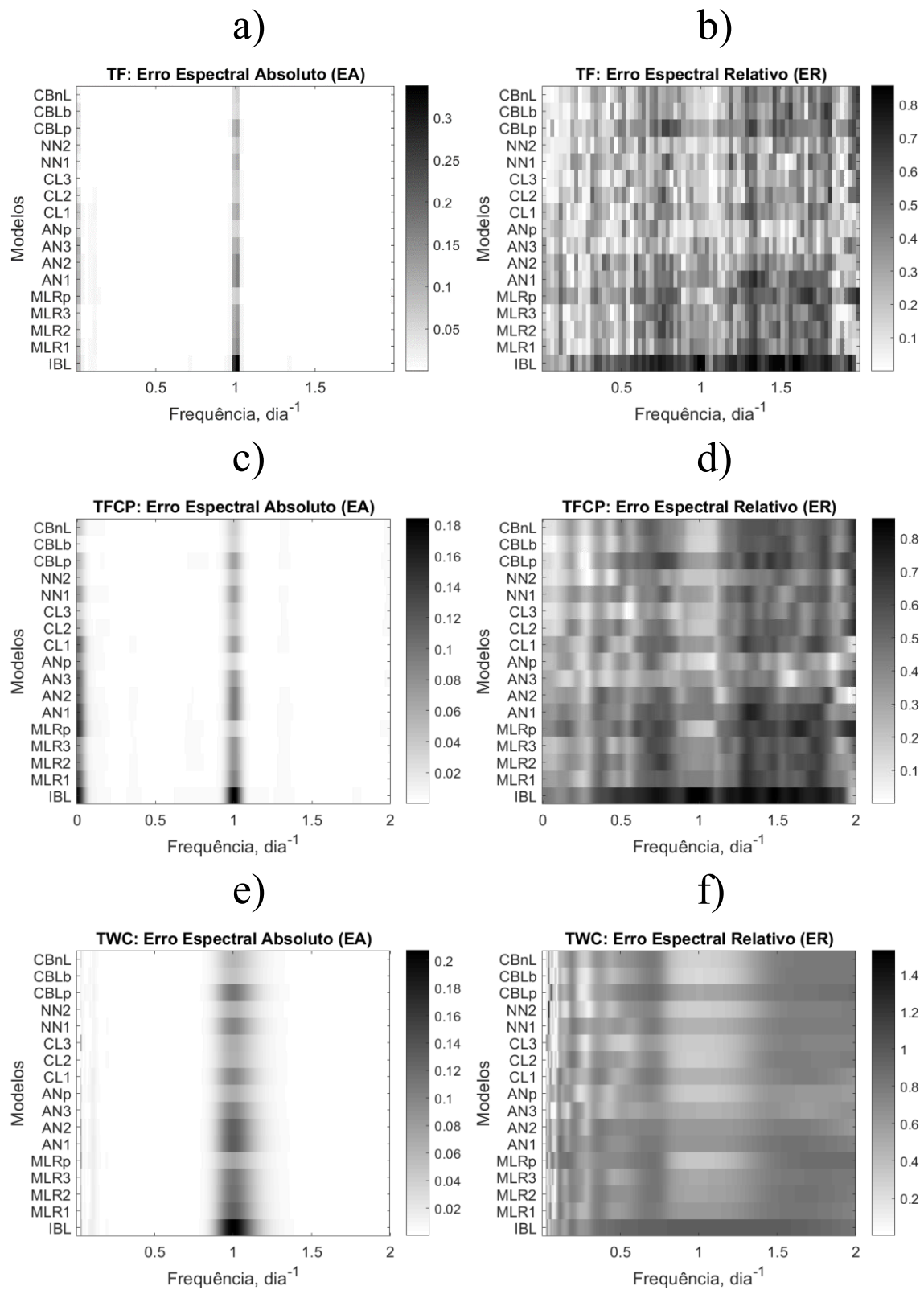
As principais diferenças entre a TFCP e TWC ficam mais evidentes comparando as Figuras 5.8c e 5.8e ou 5.8d e 5.8f. É possível observar que a TWC apresenta um aspecto mais alongado (i.e., baixa resolução) em faixas de alta frequência. Isto ocorre em virtude das

propriedades da função *wavelet*, em que uma variação da janela temporal de projeção aplica intervalos temporais menores para inspecionar altas frequências (Figs. 5.8e e 5.8f). Por esse mesmo motivo, a TWC discrimina com maior resolução as baixas frequências em relação à TFCP. Essa última, por outro lado, mantém a mesma resolução de tempo-frequência em toda análise (Figs. 5.8c e 5.8d). Também se observou na disposição temporal dos coeficientes da TWC que esta é mais apropriada para discriminar efeitos transientes quando comparada à TFCP (resultado não apresentado no texto), o que pode ser útil na elaboração de modelos de classificação ou adaptativos (com parâmetros variáveis).

Comparando as Figuras 5.8a, 5.8c e 5.8e, observa-se que o EA está, em geral, distribuído em torno da frequência 1 dia^{-1} (aqui referido como faixa de médias frequências, entre $0,5$ e $1,5 \text{ dia}^{-1}$, aproximadamente). Esta componente está associada a fenômenos mesoescalares cujo período típico de ocorrência é de ordem diária (e.g., brisa marítima; MAGNAGO et al., 2010). A TF apresenta valores de energia espectral mais concentrados em torno desse valor, havendo uma dispersão maior em torno desse valor no caso da TWC. A comparação dessas figuras também sugere que alguns dos modelos incluídos no conjunto de entrada do CBLb em escala horária, MLRp, ANp, CL2, CL3 e NN2, se destacam com respeito à amplitude espectral diária reproduzida. Em geral, esses foram os modelos que melhor se ajustaram aos dados observacionais (Fig. 5.2). Como analisado anteriormente, sob a perspectiva do tempo, o CBLb (que apresentou os melhores resultados com respeito ao SS4) apresenta boa performance considerando as baixas e médias frequências, o que é confirmado pela Figura 5.8. O CBnL apresentou resultados similares ao do CBLb, em acordo com a Figura 5.2 Outro resultado importante confirmado entre as Figuras 5.2 e 5.8 é a incapacidade do IBL ao reproduzir o comportamento observado.

Como mencionado anteriormente, a complementaridade é considerada nesse estudo como a possibilidade de aproveitamento de diferentes aptidões de modelos. Nesse sentido, em contrapartida com o aspecto absoluto, as Figuras 5.8b, 5.8d e 5.8f revelam que alguns modelos possuem contribuições complementares em outras faixas de frequência (e.g., AN2, AN3, ANp). Embora sejam pouco expressivas do ponto de vista absoluto, essas podem ajudar a descrever variabilidades associadas com faixas de frequências importantes como, por exemplo, as de alta frequência, aproximadamente entre $1,5 \text{ dia}^{-1}$ e 2 dia^{-1} (e.g., AN2, AN3, ANp e CL1), o que, em teoria, aumentaria a acurácia do modelo de combinação, que atualmente possui limitações nesse intervalo.

Figura 5.8. Caso 1: Erros Espectrais Absolutos e Relativos.



(a) e (b) apresentam, respectivamente, os erros absoluto e relativo considerando a TF; (c) e (d) apresentam os erros (absoluto e relativo) considerando a TFCP; (e) e (f) apresentam os erros obtidos com o emprego da TWC.
Fonte: Própria.

Em razão do que foi mencionado no parágrafo anterior, podem ser propostas estratégias de recombinação de modelos (combinação em dois estágios), para melhorar ainda mais as estimativas. Por exemplo, o modelo CBLp e outros modelos podem ser combinados ao CBLb¹⁸. Por esse motivo, procedimentos de seleção de variáveis que permitam aproveitar as informações sutis de complementaridade espectral podem ser propostos para aproveitar aptidões de saídas que se ocultam em procedimentos convencionais, fazendo com que essas sejam subutilizadas (e.g., aqueles baseados em correlação linear, SS4, RMSE, etc.).

5.2 Estudo de Caso 2: Melhores Performances Relativas à Interpolação

Bilinear

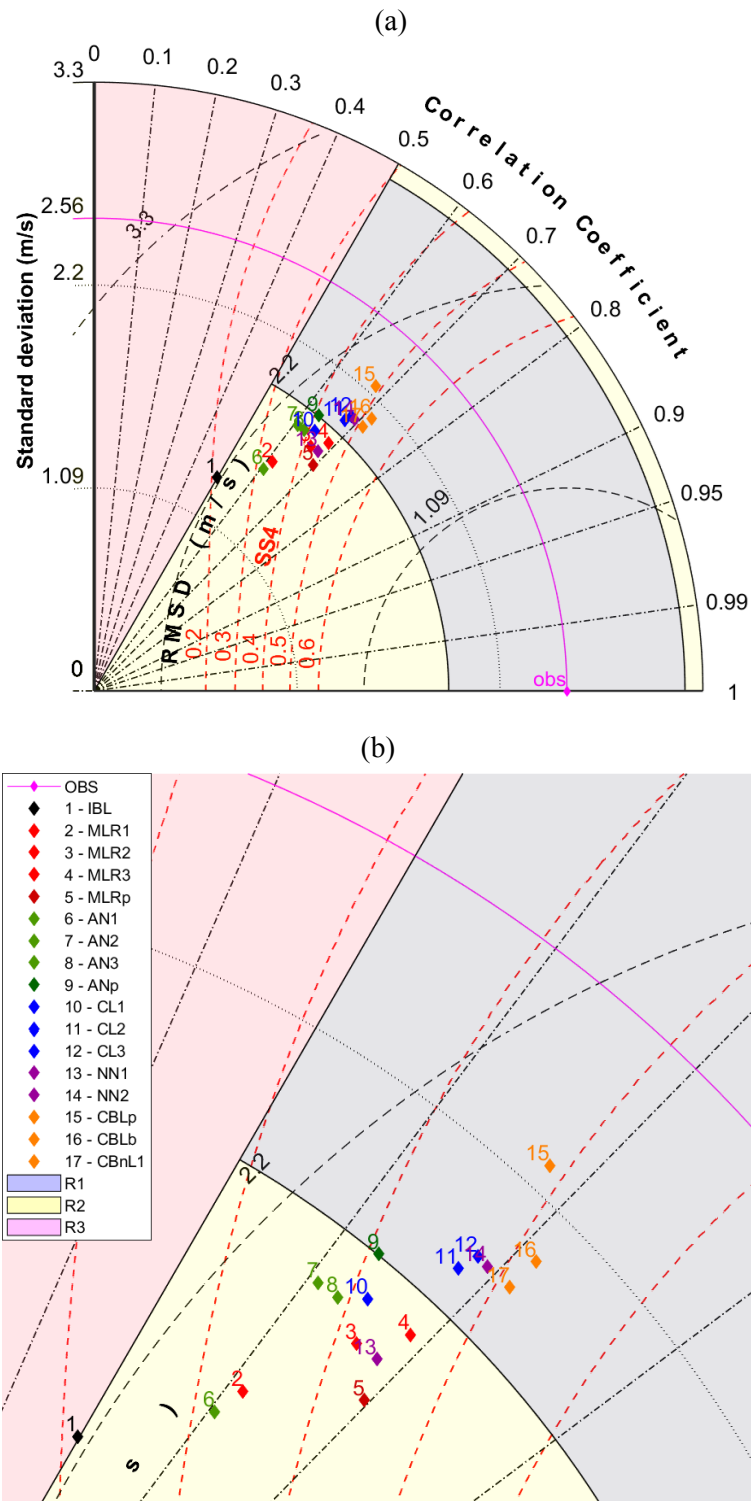
Os resultados apresentados nesta subseção dizem correspondem à parametrização realizada para a estação anemométrica de índice #3 (Figura 5.1b). Tal escolha se dá em virtude do aprimoramento relativo ao IBL, mensurado através do SS4, tanto em escala horária como em escala diária. O conjunto de treinamento se estende desde 1º de setembro de 2013 até 14 de dezembro de 2014; O conjunto de validação abrange o período de 15 de dezembro de 2014 até 6 de agosto de 2015 (observe que todos os resultados nesta subseção se referem apenas ao período de validação).

A Figura 5.9 mostra que a correlação e a variação reproduzida pelos modelos variam significativamente, por exemplo, de IBL até CBLb e CBLp, respectivamente. Os resultados também mostram que todos os modelos foram capazes de superar o IBL em termos da variância reproduzida e correlação, e, conseqüentemente, do SS4.

Em geral, os modelos de regressão linear pura, MLR1, MLR2, MLR3 e MLRp, são capazes de produzir estimativas mais precisas (maior SS4) do que os modelos baseados em análogos correspondentes, AN1, AN2, AN3 e ANp, respectivamente. Entretanto, com exceção do ANp, todos estão localizados na região R2 do gráfico, região que delimita os modelos com a menor variância reproduzida. O maior ganho dos modelos MLR se dá, nesse, com respeito à correlação.

¹⁸ Isso também aumenta a complexidade associada à estimativa final, o que pode não necessariamente implicar em maior robustez ao modelo.

Figura 5.9. Caso 2: Diagrama de Taylor (escala horária).



(a) intercomparação entre todos os modelos na Tabela 4.1 e combinações, a figura ao lado fornece uma legenda;
(b) zoom em modelos com maior SS4. Fonte: Própria.

Os modelos baseados em redes neurais, NN1 e NN2 apresentam melhorias expressivas quando comparados aos valores. O NN1 apresenta uma variância reproduzida reduzida quando compara a alguns modelos, estando na zona R2 do diagrama (saída similar ao MLR2). Apesar

disso, note que esse modelo, em termos do SS4 alcançado, apresenta uma melhora significativa relativa aos modelos que empregam o domínio composto por malhas que contém os quatro pontos mais próximos à posição de interesse (IBL, MLR1, AN1, CL1). O NN2, por outro lado, apresentou os melhores resultados individuais (maiores SS4, correlação e variância reproduzida dentre todos as saídas não combinadas).

O modelos CL1, CL2 e CL3 apresentam melhorias significativas com respeito aos modelos MLR1, MLR2 e MLR3, respectivamente, o que demonstra que também nesse caso, as técnicas de agrupamento foram capazes de detectar classes de padrões sinóticos significativas, o que aumentou a acurácia das estimativas em função da variação dos parâmetros (i.e., ajuste local).

A Tabela 5.3 apresenta em mais detalhes os resultados obtidos com os modelos baseados em agrupamento (CL1, CL2 e CL3) para o Caso 2, em regime diário e horário. Resultados semelhantes foram obtidos para CL1 e CL2 com respeito ao número de classes aplicadas e limiar de tolerância (T). Observe que, em escala horária, esses dois modelos produzem um maior número de classes aplicáveis. Por outro lado, a configuração ótima do CL3 emprega um número reduzido de classes e um maior valor para o fator de tolerância. Assim como observado para o Caso 1, os valores de T ótimos tendem a aumentar para domínios mais amplos, o que torna o processo de classificação mais restritivo.

No Caso 2, em regime horário, as técnicas de combinação CBLb e CBLp alcançam a maior correlação e variância explicada, respectivamente (ver Figura 5.9b). O que demonstra que os dois procedimentos de seleção de variáveis propostos, i.e., PCA e seleção por SS4 (Seção 4.2), foram adequados para descrever características específicas das observações. Deve-se atentar ao fato de que CBLb também alcança o melhor resultado geral em relação ao índice de habilidade SS4, com um valor de 0,5180. Esse valor representa uma melhora de 142,03% em relação ao SS4 obtido pelo IBL. Esses resultados de CBLb estão sendo obtidos combinando os seguintes modelos: MLR1, MLR2, CL2, CL3, NN1 e NN2. O CBnL apresentou um resultado similar ao do CBLb, porém, um pouco inferior, indicando que não houve ganho com o ajuste não linear na combinação nesse caso.

A Figura 5.10 mostra os resultados obtidos para o regime diário. Assim como no caso anterior, um aumento no valor do SS4 obtido e uma diminuição no RMSD também são mostrados, decorrentes à previsibilidade intrínseca ao sistema na base diária. Apesar disso, uma redução no SS4 de todos os modelos pode ser observada quando comparada com os resultados do Estudo de Caso 1 (o que também pode ser visto comparando Estudo do Caso 2 em regime horário). Um dentre os inúmeros motivos possíveis para isso está associado à qualidade geral

dos dados observacionais empregados. Note que, como mencionado no início da Seção 5, o Estudo de Caso 2 dispõe de uma base de dados com um percentual de dados ausentes ou anômalos significativamente maior que no Caso 1.

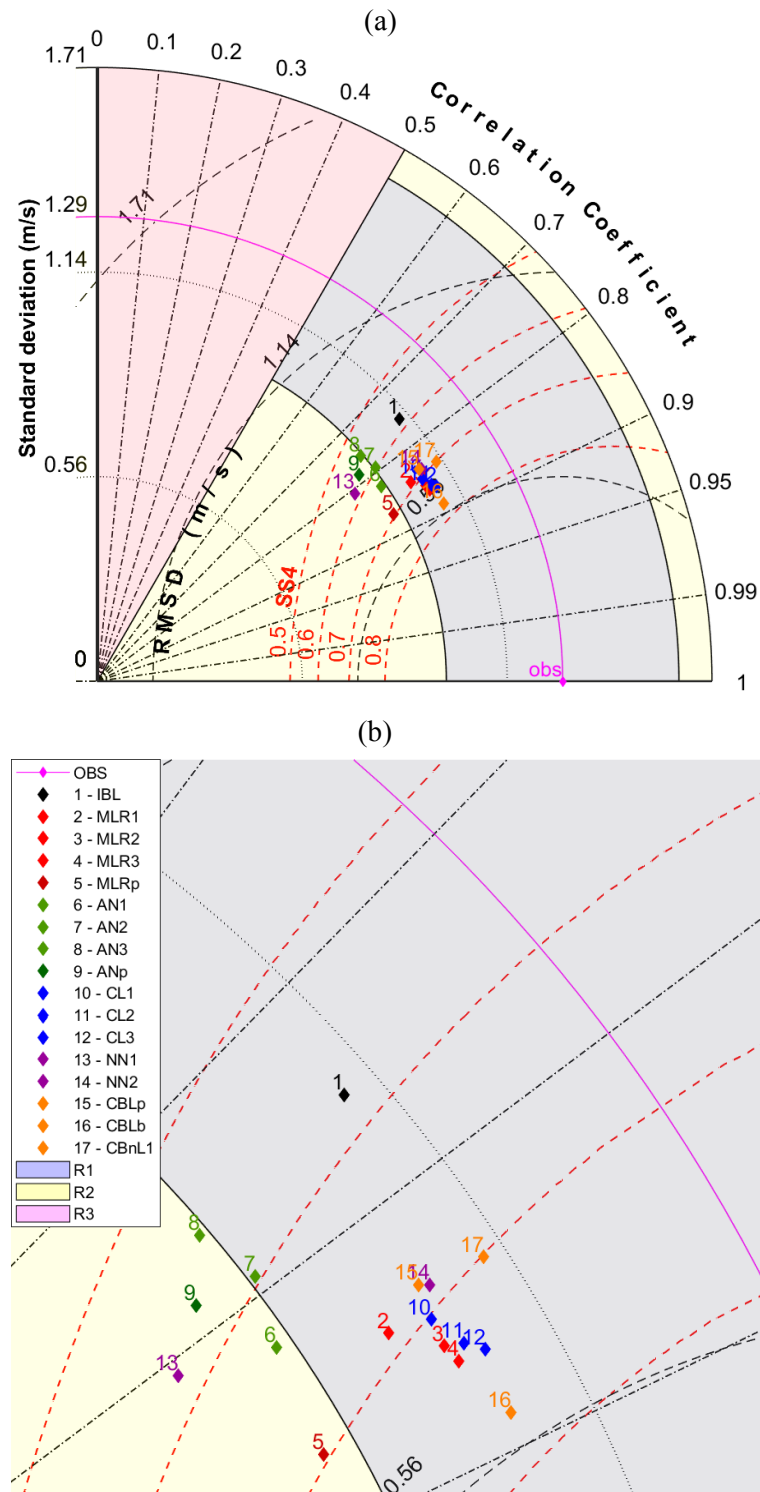
Tabela 5.3. Caso 2: configurações resultantes para CL1, CL2 e CL3.

c) Escala Horária	CL1	CL2	CL3
Distância (AHC)	Cosseno	Cosseno	<i>Manhattan</i>
Algoritmo (AHC)	VP	Vínculo Médio	VP
Distância (NHC)	Euclideana	<i>Manhattan</i>	Cosseno
Classes Formadas	76	94	80
Classes Aplicadas	20	20	7
Limiar (<i>T</i>)	2%	2%	4%
Nível do GCM	10m	60	59
d) Escala Diária	CL1	CL2	CL3
Distância (AHC)	Euclideana	Correlação	Cosseno
Algoritmo (AHC)	VP	VP	Vínculo Médio
Distância (NHC)	Euclideana	Euclideana	-
Clusters Formados	76	70	88
Clusters Aplicados	5	1	2
Limiar (<i>T</i>)	4%	6%	20%
Nível do GCM	52	55	54

(a) escala horária; (b) escala diária. Mais detalhes sobre as distâncias e algoritmos estão disponíveis na Tabela 3.
Fonte: Própria.

É possível observar que, em geral, há uma melhora da maior parte dos modelos com respeito ao IBL. Diferentemente do Caso 1 (regime diário), os modelos AN apresentam uma tendência por subestimar a variabilidade observada, porém, exceto pelo AN3, esse apresentam uma melhora sensível ao fator de correlação (em relação ao obtido pelo IBL). Note que, em conformidade com os resultados apresentados até então, os modelos baseados em regressão (MLR e NN, agrupamento de padrões sinóticos (CL) e combinação (CBL e CBnL) maior habilidade em reproduzir as características do vento local observado. O melhor resultado geral está sendo obtido pela CBLb, que atinge uma pontuação de SS4 de 0,7714. Esse valor representa uma melhora de 32,14% em relação SS4 do IBL. Estes resultados da CBLb estão sendo obtidos através da combinação dos modelos: MLRp, CL2, CL3 e NN2.

Figura 5.10. Caso 2: Diagrama de Taylor (escala diária).



(a) intercomparação entre todos os modelos na Tabela 4.1 e combinações, a figura ao lado fornece uma legenda;
(b) zoom em modelos com maior SS4. Fonte: Própria.

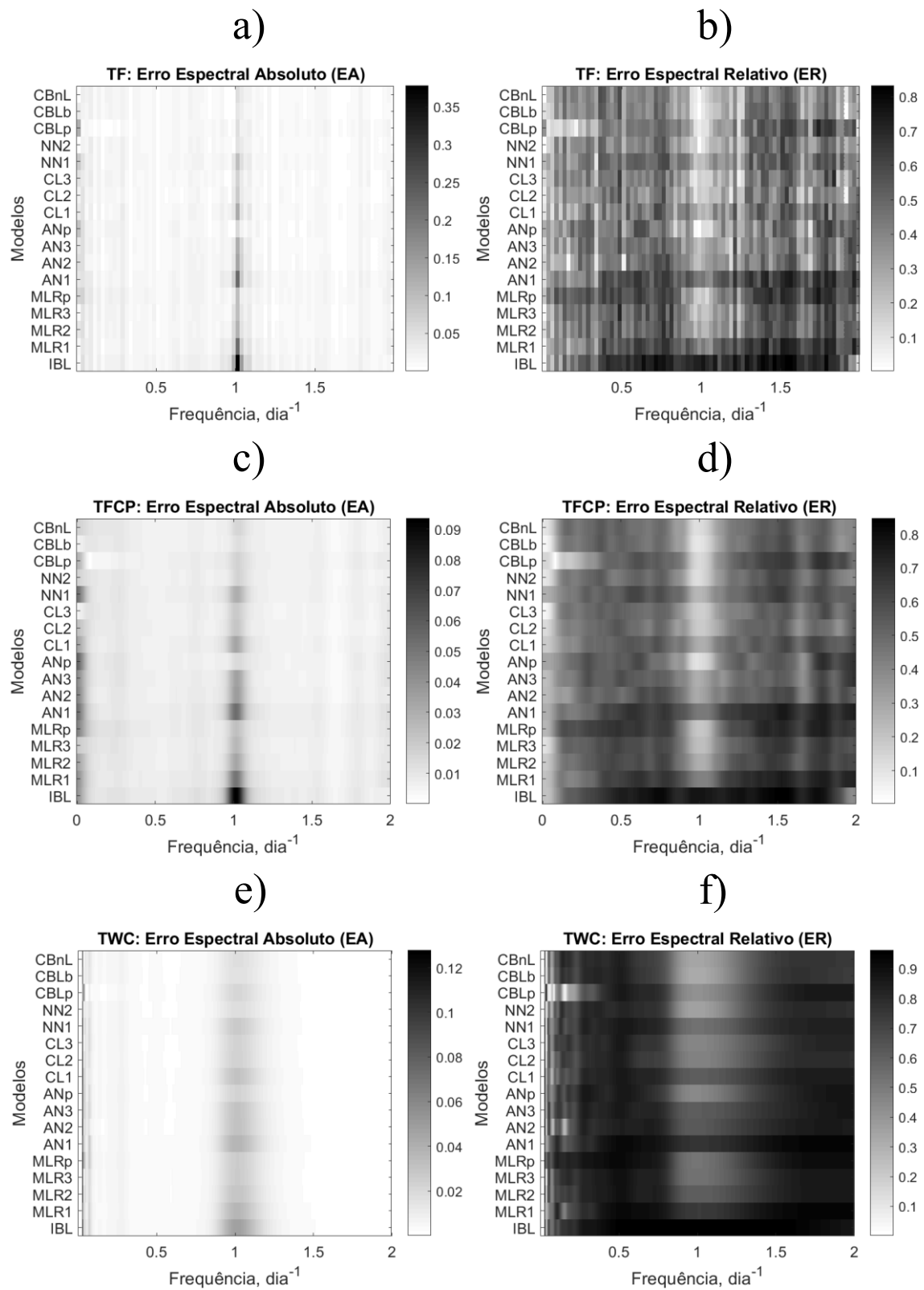
Na Figura 5.11 os resultados da análise de complementaridade espectral são apresentados para o Caso 2. Assim como no Estudo de Caso 1, apenas os dados em escala horária foram avaliados. Também nesse caso se apresentou uma importante componente

espectral em torno da frequência 1 dia^{-1} (Fig. 5.11a, 5.11c e 5.11e). Assim como o local escolhido para o Estudo de Caso 1, o local estudado nessa seção está situado na costa, sendo assim, tal componente espectral pode estar associada a fenômenos como a brisa marítima, que possui periodicidade diurna (ver Figura 5.1, índices de local #1 e #3, para o Caso 1 e 2, respectivamente). Com respeito ao EA, as Figuras 5.11a, 5.11c e 5.11e mostram que o CBLb, modelo que apresentou os melhores resultados também no Estudo de Caso 2 (regime horário), se destaca quanto à habilidade em reproduzir o espectro global das observações. Outros modelos como o MLR2, MLR3, MLRp, ANp, CL2, CL3 e NN2 também reproduzem o espectro em torno da frequência de 1 dia^{-1} . Isso também pode ser verificado nas Figuras 5.11b, 5.11d e 5.11f, que mostram o ER obtidos para cada saída.

As figuras mostram que, em geral, a interpolação apresenta os maiores erros (EA e ER), reafirmando os resultados da Figura 5.2. Também pode ser visto na Figura 5.8 que os modelos MLR3, MLRp e ANp, que não foram incluídos no conjunto de saídas selecionados para o ajuste do CBLb e CBnL, reproduzem satisfatoriamente as faixas de médias frequências (aproximadamente entre $0,5$ e $1,5 \text{ dia}^{-1}$), entretanto, não possuem informações complementares significativas. Todavia, pode ser observado através do EA e ER que o CBLp apresenta uma informação significativa complementar ao CBLb e aos demais modelos em baixas frequências (menor que $0,5 \text{ dia}^{-1}$). Essa aptidão do CBLp pode ser devidamente explorada em estratégias de recombinação, por meio de procedimentos objetivos apropriados, uma vez que a performance reduzida desse modelo em zonas de alta frequência podem impactar a qualidade da estimativa combinada. Nas faixas de alta frequência (entre $1,5$ e 2 dia^{-1}), os modelos AN2, AN3, ANp, CL2, NN2, CBLb e CBnL apresentam uma representatividade sensivelmente superior às demais, havendo uma relação de complementaridade sutil com o CBLp, por exemplo.

É importante ressaltar que as conclusões obtidas nesse estudo estão atreladas aos pressupostos realizados. Entre eles está a aplicabilidade de transformadas nas séries temporais utilizadas. Os testes são realizados sobre os sinais sob o pressuposto de estacionaridade do sinal, condição que, em geral, é fortemente impactada pela qualidade dos dados. Todo o pré-processamento realizado sobre os dados brutos também pode ter forte influência sobre o comportamento espectral dos sinais.

Figura 5.11. Caso 2: Erros Espectrais Absoluto e Relativo.



(a) e (b) apresentam, respectivamente, os erros absoluto e relativo considerando a TF; (c) e (d) apresentam os erros (absoluto e relativo) considerando a TFCP; (e) e (f) apresentam os erros obtidos com o emprego da TWC.

Fonte: Própria.

6 CONCLUSÕES E PERSPECTIVAS

Os resultados da modelagem estatística do vento local revelam que técnicas simples de *downscaling*, associadas entre si através de técnicas de combinação, fornecem estimativas confiáveis para o comportamento do vento observado (em diferentes locais da Região Nordeste do Brasil) em baixas e médias frequências. Assim, a metodologia proposta é adequada para reproduzir o comportamento médio a longo prazo do vento local, sendo útil em aplicações como avaliação do recurso eólico com vistas ao estudo de viabilidade técnico-econômica para a implantação de centrais eólicas. No entanto, ainda há limitações a serem superadas com respeito à simulação de eventos associados a variações rápidas (alta frequência). Apesar de tais limitações, em geral, os modelos são capazes de reproduzir satisfatoriamente as estruturas de fase/frequência e amplitude de variação das observações em escalas intradiária e diária, havendo um maior aprimoramento com respeito ao modelo de referência (a interpolação) nos casos de regime horário.

Outro importante aspecto sobre as técnicas de *downscaling* implementadas neste estudo é o de que elas produzem perdas consideráveis da variabilidade observada, o que é esperado em abordagens baseadas em regressão (WILBY et al., 2004). Contudo, os resultados mostram que os modelos não apresentaram, em geral, tendências significativas de sub ou superestimação do valor médio das observações. Por outra parte, de forma contrária ao que se observou para médias e baixas frequências, os modelos não apresentaram habilidade satisfatória para a reprodução do comportamento das observações em altas frequências. Ainda assim, pode-se dizer que essas técnicas são apropriadas para aplicações como previsão de curto prazo, desde que associadas a outras técnicas que ajudem a prever o comportamento de alta frequência, isto é, em horizontes de minutos até horas à frente (e.g., análises de complementaridade espectral objetivas ou técnicas baseadas em séries temporais).

Na maioria dos casos estudados, técnicas baseadas em regressão, agrupamento e combinação superaram o modelo de referência (interpolação), bem como outros modelos mais simples considerados. Além disso, o aumento da complexidade dos modelos foi justificado, em geral, pelo consequente aumento na acurácia dos resultados. Em outras palavras, o emprego de parâmetros lineares fixos, em geral, é superado por técnicas de parâmetros variáveis (locais) e técnicas não paramétricas (e.g., redes neurais).

Buscando analisar a complementaridade entre as saídas dos modelos individuais e as combinações entre modelos, realizou-se uma análise no domínio da frequência tendo-se em

conta que a relação inversa entre a frequência e escalas de variação temporal permite uma visualização da distribuição da variância (ou energia espectral) em função de faixas de frequência (i.e., escalas de variação temporal ou períodos). Tal análise revela importantes aspectos associados ao comportamento dos modelos, aspectos ocultos sob a perspectiva do domínio do tempo. A análise de complementaridade espectral foi aplicada por meio de inspeção em erros de escala absoluta e relativa entre as energias espectrais dos sinais avaliados. A escala absoluta revelou as escalas de variação temporal que contêm a maior parte da variância do sinal, mostrando conformidade com os resultados apresentados para as saídas individuais e combinadas no domínio do tempo. Por sua vez, a escala relativa mostrou que há informações de complementaridade (entre os modelos) ocultas nas séries temporais. Tais informações podem ser aproveitadas de forma mais ampla, não só para melhor entender o comportamento do sinal, mas também para aumentar a confiabilidade nas saídas individuais e combinadas (e.g., através de modelos de parâmetros variáveis).

Por fim, como proposta para trabalhos futuros, sugerem-se:

a) com relação às técnicas de *downscaling* estatístico: novas abordagens para a definição de diferentes conjuntos de treinamento e validação (validação cruzada), com vistas a avaliar a robustez dos modelos, i.e., sua capacidade de generalizar os resultados obtidos independentemente da escolha realizada para os conjuntos de treinamento e validação;

b) com relação à seleção das variáveis regressoras: aplicação de métodos objetivos (automáticos) para a definição *a priori* dos domínios sobre os quais serão aplicados os modelos lineares e não-lineares. Esses domínios poderão conter, além de dados de vento, variáveis exógenas tais como pressão atmosférica e temperatura;

c) com relação aos modelos de classificação do padrão sinóptico: novas abordagens como, por exemplo, a aplicação de regressão não linear especializada sobre períodos contidos em uma determinada classe;

d) no que se refere aos modelos de combinação e complementaridade espectral: desenvolvimento de métodos automáticos (e.g., baseados em teoria da informação mútua e seleção de atributos) para a avaliação (anterior à calibração dos modelos) da complementaridade entre diferentes subconjuntos de preditores candidatos, de forma a reduzir a necessidade de esforço computacional em algoritmos de combinação, bem como aumentar a confiabilidade das estimativas.

REFERÊNCIAS

ABEEólica. Dados mensais. Associação Brasileira de Energia Eólica. Abril de 2017.

ACCADIA, C.; MARIANI, S.; CASAIOLI, M. e LAVAGNINI, A. Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Weather and forecasting*, v. 18, n. 5, p. 918-932, 2003.

ALLEN, J. Short-term spectral analysis, and modification by discrete Fourier transform. *IEEE Transactions on Acoustics Speech and Signal Processing*, v. 25, n. 3, p. 235-238, 1977.

ALMEIDA, L. B. The fractional Fourier transform and time-frequency representations. *IEEE Transactions on signal processing*, v. 42, n. 11, p. 3084-3091, 1994.

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*, v. 2, n. 3, p. 265-276, 2011.

AMS. Meteorology Glossary. American Meteorology Society, 2012. Disponível em: <<http://glossary.ametsoc.org/wiki/Wiki/History>>. Acesso em: Janeiro 2016.

ANEEL. Resolução normativa nº 482, de 17 de abril de 2012.

ANEEL. Resolução normativa nº 687, de 24 de novembro de 2015.

BADGER, J., FRANK, H., HAHMANN, A. N., GIEBEL, G. Wind-climate estimation based on mesoscale and microscale modeling: statistical–dynamical downscaling for wind energy applications. *Journal of Applied Meteorology and Climatology*, v. 53, n. 8, p. 1901-1919, 2014.

BARDOSSY, A.; DUCKSTEIN, L.; BOGARDI, I. Fuzzy rule-based classification of atmospheric circulation patterns. *International Journal of Climatology*, v. 15, n. 10, p. 1087-1097, 1995.

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

BARNETT, T. P.; PREISENDORFER, R. W. Multifield analog prediction of short-term climate fluctuations using a climate state vector. *Journal of the Atmospheric Sciences*, v. 35, n. 10, p. 1771-1787, 1978.

BARTHELMIE, R. J., MURRAY, F., PRYOR, S. C. The economic benefit of short-term forecasting for wind energy in the UK electricity market. *Energy Policy*, v. 36, n. 5, p. 1687-1696, 2008.

BATHURST, G. N., WEATHERILL, J., STRBAC, G. Trading Wind Generation In Short Term energy markets. *IEEE Transactions on Power Systems*, v. 17, n. 3, p. 782-789, 2002.

BERGLAND, G. D. A guided tour of the fast Fourier transform. *IEEE spectrum*, v. 6, n. 7, p. 41-52, 1969.

BERNARDIN F.; BOSSY, M.; CHAUVIN, C.; DROBINSKI, P.; ROUSSEAU, A.; SALAMEH, T. Stochastic downscaling method: application to wind refinement. *Stochastic Environmental Research and Risk Assessment*, v. 23, n. 6, p. 851-859, 2009.

BERNHARDT, M.; LISTON, G. E.; STRASSER, U.; ZÄNGL, G.; SCHULZ, K. High resolution modelling of snow transport in complex terrain using downscaled MM5 wind fields. *The Cryosphere*, v. 4, n. 1, p. 99-113, 2010.

BIAU, G.; ZORITA, E.; VON STORCH, H.; WACKERNAGEL, H. Estimation of precipitation by kriging in the EOF space of the sea level pressure field. *Journal of Climate*, v. 12, n. 4, p. 1070-1085, 1999.

BOÉ, J.; TERRAY, L.; HABETS, F.; MARTIN E. A simple statistical-dynamical downscaling scheme based on weather types and conditional resampling. *Journal of Geophysical Research: Atmospheres*, v. 111, n. D23, 2006.

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

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*, v. 28, n. 4, p. 449-464, 2008.

BÜRGER, G. Expanded downscaling for generating local weather scenarios. *Climate Research*, p. 111-128, 1996.

CANNON, A. J.; WHITFIELD, P. H. Downscaling recent streamflow conditions in British Columbia, Canada using ensemble neural network models. *Journal of Hydrology*, v. 259, n. 1, p. 136-151, 2002.

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*, v. 36, n. 4, p. 1245-1251, 2011.

COSTA, A.; CRESPO, A.; NAVARRO, J.; LIZCANO, G.; MADSEN, H.; FEITOSA, E. A review on the young history of the wind power short-term prediction. *Renewable and Sustainable Energy Reviews*, v. 12, n. 6, p. 1725-1744, 2008.

COULIBALY, P.; DIBIKE, Y. B.; ANCTIL, F. Downscaling precipitation and temperature with temporal neural networks. *Journal of Hydrometeorology*, v. 6, n. 4, p. 483-496, 2005.

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

DAWSON, C. W.; WILBY, R. L. Hydrological modelling using artificial neural networks. *Progress in physical Geography*, v. 25, n. 1, p. 80-108, 2001.

DEE, D. P.; UPPALA, S. M.; SIMMONS, A. J.; BERRISFORD, P.; POLI, P.; KOBAYASHI, S.; ANDRAE, U.; BALMASEDA, M. A.; BALSAMO, G.; BAUER, P.; BECHTOLD, P.; BELJAARS, A. C. M.; VAN DE BERG, L.; BIDLOT, J.; BORMANN, N.; DELSOL, C.; DRAGANI, R.; FUENTES, M.; GEER, A. J.; HAIMBERGER, L.; HEALY, S. B.; HERSBACH, H.; HÓLM, E. V.; ISAKSEN, I.; KÄLLBERG, P.; KÖHLER, M.; MATRICARDI, M.; MCNALLY, A. P.; MONGE-SANZ, B. M.; MORCRETTE, J.-J.; PARK, B.-K.; PEUBEY, C.; de ROSNAY, P.; TAVOLATO, C.; THÉPAUT, J.-N.; VITART, F. The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the royal meteorological society*, v. 137, n. 656, p. 553-597, 2011.

DEGAETANO, A. T.; BELCHER, B. N. Spatial interpolation of daily maximum and minimum air temperature based on meteorological model analyses and independent observations. *Journal of Applied Meteorology and Climatology*, v. 46, n. 11, p. 1981-1992, 2007.

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

DUBOIS, D.; PRADE, H. What are fuzzy rules and how to use them. *Fuzzy sets and systems*, v. 84, n. 2, p. 169-185, 1996.

ECMWF. What is the direction convention for the U and V components of winds? 2018. Disponível em: <<https://www.ecmwf.int/en/faq/what-direction-convention-u-and-v-components-winds>>. Acesso em Janeiro de 2018.

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

GELADI, P.; KOWALSKI, B. R. Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, v. 185, p. 1-17, 1986.

GRAPS, A. An introduction to wavelets. *IEEE computational science and engineering*, v. 2, n. 2, p. 50-61, 1995.

GRAYBEAL, D. Y., DEGAETANO, A. T., EGGLESTON, K. L. Complex quality assurance of historical hourly surface airways meteorological data. *Journal of Applied Meteorology*, v. 21, p. 1156-1169, 2004.

GRELL, G. A.; DUDHIA, J.; STAUFFER D. R.; A description of the fifth-generation Penn State/NCAR mesoscale model (MM5). 1994.

GRIFFIN, D.; LIM, J. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 32, n. 2, p. 236-243, 1984.

GUO, Y.; LI, J.; LI, Y. A time-scale decomposition approach to statistically downscale summer rainfall over North China. *Journal of Climate*, v. 25, n. 2, p. 572-591, 2012.

GURLEY, K.; KAREEM, A. Applications of wavelet transforms in earthquake, wind and ocean engineering. *Engineering structures*, v. 21, n. 2, p. 149-167, 1999.

GUTIÉRREZ, J. M.; COFIÑO A. S.; CANO R.; RODRÍGUEZ M. A. Clustering methods for statistical downscaling in short-range weather forecasts. *Monthly Weather Review*, v. 132, n. 9, p. 2169-2183, 2004.

GYALISTRAS, D.; VON STORCH, H.; FISCHLIN, A.; BENISTON M. Linking GCM-simulated climatic changes to ecosystem models: case studies of statistical downscaling in the Alps. *Climate Research*, v. 4, n. 3, p. 167-189, 1994.

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*, v. 143, n. 9, p. 3518-3531, 2015.

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

HIBON, M.; EVGENIOU, T. To combine or not to combine: selecting among forecasts and their combinations. *International Journal of Forecasting*, v. 21, n. 1, p. 15-24, 2005.

HIDALGO, H. G.; DETTINGER, Michael D.; CAYAN, Daniel R. Downscaling with constructed analogues: Daily precipitation and temperature fields over the United States. California Energy Commission PIER Final Project Report CEC-500-2007-123, 2008.

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

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

IPCC. Intergovernmental Panel on Climate Change. Data Distribution Centre, 2013. Disponível em: <http://www.ipcc-data.org/guidelines/pages/weather_generators.html>. Acesso em: 29 Outubro 2016.

JIMÉNEZ, P. A., GONZÁLEZ-ROUCO, J. F., NAVARRO, J., MONTÁVEZ, J. P., GARCÍA-BUSTAMANTE, E. Quality assurance of surface wind observations from automated weather stations. *Journal of atmospheric and Oceanic Technology*, v. 27, n. 7, p. 1101-1122, 2010.

JOLLIFFE, I. T. Principle component analysis. Springer Verlag, New York, 1986.

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

JUNG, J.; BROADWATER, R. P. Current status and future advances for wind speed and power forecasting. *Renewable and Sustainable Energy Reviews*, v. 31, p. 762-777, 2014.

JUNG, J.; TAM, K-S. A frequency domain approach to characterize and analyze wind speed patterns. *Applied energy*, v. 103, p. 435-443, 2013.

KALNAY, E.; KANAMITSU, M.; KISTLER, R.; COLLINS, W.; DEAVEN, D.; GANDIN, L.; IREDELL, M.; SAHA, S.; WHITE, G.; J. WOOLLEN; ZHU, Y., LEETMAA A.; REYNOLDS, R.; CHELLIAH, M.; EBISUZAKI, W., HIGGINS, W.; JANOWIAK, J.; MO, K. C.; ROPELEWSKI, C.; WANG J. ;JENNE R.; JOSEPH D. The NCEP/NCAR 40-Year Reanalysis Project. *Bulletin of the American Meteorological Society*, v. 77, n. 3, 1996.

KANG, H.; PARK, C. -K.; HAMEED, S. N.; ASHOK, K. Statistical downscaling of precipitation in Korea using multimodel output variables as predictors. *Monthly weather review*, v. 137, n. 6, p. 1928-1938, 2008.

LAM, N. S.-N. Spatial interpolation methods: a review. *The American Cartographer*, v. 10, n. 2, p. 129-150, 1983.

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

LANDBERG, L. Short-term prediction of local wind conditions. *Journal of Wind Engineering and Industrial Aerodynamics*, v. 89, n. 3-4, p. 235-245, 2001.

LEEDS, W. B.; STEIN, M. L.; MOYER, E. J. Conditional Simulation of Future Climate Under Changing Temporal Covariance Structures. 2013. Disponível em: <http://geosci.uchicago.edu/~moyer/MoyerWebsite/PrivateFiles_unlinked/Leeds_etal_CondSim.pdf>. acesso em janeiro de 2018.

LI, J.; HEAP, A. D. A review of spatial interpolation methods for environmental scientists/Jin Li and Andrew D. Heap. *Record (Australia. Geoscience Australia)*, v. 23, 2008.

LIU, D.; NIU, D.; WANG, H.; FAN, L. Short-term wind speed forecasting using wavelet transform and support vector machines optimized by genetic algorithm. *Renewable Energy*, v. 62, p. 592-597, 2014.

LIU, H.; CHEN, C.; TIAN, H-Q.; LI, Y-F. A hybrid model for wind speed prediction using empirical mode decomposition and artificial neural networks. *Renewable Energy*, v. 48, p. 545-556, 2012.

LORENZ, E. N. Atmospheric predictability as revealed by naturally occurring analogues. *Journal of the Atmospheric sciences*, v. 26, n. 4, p. 636-646, 1969.

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*, v. 28, n. 5, p. 702-708, 1971.

MAGNAGO, R. O.; FISCH, G.; DE MORAES, O. L. L. Análise espectral do vento no Centro de Lançamento de Alcântara (CLA). *Revista Brasileira de Meteorologia*, v. 25, n. 2, 2010.

MEDSKER, L. R.; JAIN, L. C. Recurrent neural networks. *Design and Applications*, v. 5, 2001.

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*, v. 100, n. 3-4, p. 413-421, 2010.

MORAES, C. F. W. C. B. Objective quality assurance procedure to assess local surface wind observational data in the coast of Rio Grande do Norte. *Dissertação de mestrado*. Universidade Federal de Pernambuco (UFPE), 2015.

MURPHY, J. An evaluation of statistical and dynamical techniques for downscaling local climate. *Journal of Climate*, v. 12, n. 8, p. 2256-2284, 1999.

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

OLIVER, M. A.; WEBSTER, R. Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information System*, v. 4, n. 3, p. 313-332, 1990.

PALOMINO, I.; MARTIN, F. A simple method for spatial interpolation of the wind in complex terrain. *Journal of Applied Meteorology*, v. 34, n. 7, p. 1678-1693, 1995.

PEDLOSKY, J. The equations for geostrophic motion in the ocean. *Journal of physical oceanography*, v. 14, n. 2, p. 448-455, 1984.

PROJETO PILACAS. Piloto de um Laboratório em Campo Aberto para Certificação de Aerogeradores de Pequeno Porte. Relatório Final. Universidade Federal de Pernambuco, 2014. Disponível em:

<https://www3.ufpe.br/mecfluamb/index.php?option=com_content&view=article&id=369&Itemid=266> Acesso em Dezembro de 2017.

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*, v. 29, n. 3, p. 183-198, 2005.

RACSKO, P.; SZEIDL, L.; SEMENOV, M. A serial approach to local stochastic weather models. *Ecological modelling*, v. 57, n. 1-2, p. 27-41, 1991.

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

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

RUOSTEENOJA, K. Factors affecting the occurrence and lifetime of 500 mb height analogues: a study based on a large amount of data. *Monthly Weather Review*, v. 116, n. 2, p. 368-376, 1988.

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

SHAFER, M. A., FIEBRICH, C. A., ARNDT, D. S. Quality assurance procedures in the Oklahoma mesonetwork. *Journal of Atmospheric and Oceanic Technology*, v. 7, p. 474-494, 2000.

SCHOOF, J. T.; PRYOR, S. C. Downscaling temperature and precipitation: A comparison of regression-based methods and artificial neural networks. *International Journal of climatology*, v. 21, n. 7, p. 773-790, 2001.

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

SIFUZZAMAN, M.; ISLAM, M. R.; ALI, M. Z. Application of wavelet transform and its advantages compared to Fourier transform. *Journal of Physical Sciences*, v. 13, p. 121-134, 2009.

SILVA, G. R. Características de Vento da Região Nordeste: Análise, Modelagem e Aplicações para Projetos de Centrais Eólicas. Dissertação de mestrado, Curso de Pós-Graduação em Engenharia Mecânica/CTG/EEP/UFPE, 2003.

STOICA, P.; MOSES R.L. Spectral analysis of signals. Upper Saddle River, NJ: Pearson Prentice Hall, 2005.

SNELL, S. E.; GOPAL, S.; KAUFMANN, R. K. Spatial interpolation of surface air temperatures using artificial neural networks: Evaluating their use for downscaling GCMs. *Journal of Climate*, v. 13, n. 5, p. 886-895, 2000.

TAYLOR, K. E. Summarizing multiple aspects of model performance in a single diagram. Program for Climate Model Diagnosis and Intercomparison, Lawrence Livermore National Laboratory, University of California, 2001.

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

TIMBAL, B.; DUFOUR, A.; MCAVANEY, B. An estimate of future climate change for western France using a statistical downscaling technique. *Climate Dynamics*, v. 20, n. 7-8, p. 807-823, 2003.

TIMMERMANN, A. Forecast combinations. *Handbook of economic forecasting*, v. 1, p. 135-196, 2006.

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

TORRENCE, C.; COMPO, G. P. A practical guide to wavelet analysis. *Bulletin of the American Meteorological society*, v. 79, n. 1, p. 61-78, 1998.

VACCARO A.; MERCOGLIANO P.; SCHIANO P.; VILLACCI D. An adaptive framework based on multi-model data fusion for one-day-ahead wind power forecasting. *Electric Power Systems Research*, v. 81, n. 3, p. 775-782, 2011.

VAN DEN DOOL, H. M. Searching for analogues, how long must we wait? *Tellus A*, v. 46, n. 3, p. 314-324, 1994.

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

VERKAIK, J. W.; JACOBS, A. J. M.; TIJM, A. B. C.. Local Wind Speed Estimation by Physical Downscaling of Weather model forecasts. Submitted to *J. Wind Engineering and Industrial Aerodynamics*, 2006.

VON STORCH, H.; HEWITSON, B.; MEARN, L. Review of empirical downscaling techniques. Regional climate development under global warming. General Technical Report, v. 4, 2000.

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

WANG, N.; BAO, J-W.; LEE, J-L.; MOENG, F.; MATSUMOTO, C. Wavelet compression technique for high-resolution global model data on an icosahedral grid. *Journal of Atmospheric and Oceanic Technology*, v. 32, n. 9, p. 1650-1667, 2015.

WARD JR, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, v. 58, n. 301, p. 236-244, 1963.

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*, v. 15, n. 2, p. 70-73, 1967.

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

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

WILBY, R. L.; WIGLEY, T. M. L.; CONWAY, D.; JONES, P. D.; HEWITSON, B.C.; MAIN, J.; WILKS, D. S. Statistical downscaling of general circulation model output: a comparison of methods. *Water resources research*, v. 34, n. 11, p. 2995-3008, 1998.

WILBY, R. L.; CHARLES, S. P.; ZORITA, E.; TIMBAL, B.; WHETTON, P.; MEARN, L. O. Guidelines for use of climate scenarios developed from statistical downscaling methods. 2004.

WILKS, D. S. Multisite downscaling of daily precipitation with a stochastic weather generator. *Climate Research*, v. 11, n. 2, p. 125-136, 1999.

WILKS, D. S. Use of stochastic weather generators for precipitation downscaling. *Wiley Interdisciplinary Reviews: Climate Change*, v. 1, n. 6, p. 898-907, 2010.

WILKS, D. S. *Statistical methods in the atmospheric sciences*. Academic press, 2011.

XU, C.-Y. From GCMs to river flow: a review of downscaling methods and hydrologic modelling approaches. *Progress in physical Geography*, v. 23, n. 2, p. 229-249, 1999.

YANG, Y. Combining forecasting procedures: some theoretical results. *Econometric Theory*, v. 20, n. 01, p. 176-222, 2004.

YU, L.; LIU, H. Feature selection for high-dimensional data: A fast correlation-based filter solution. *ICML*. p. 856-863, 2003.

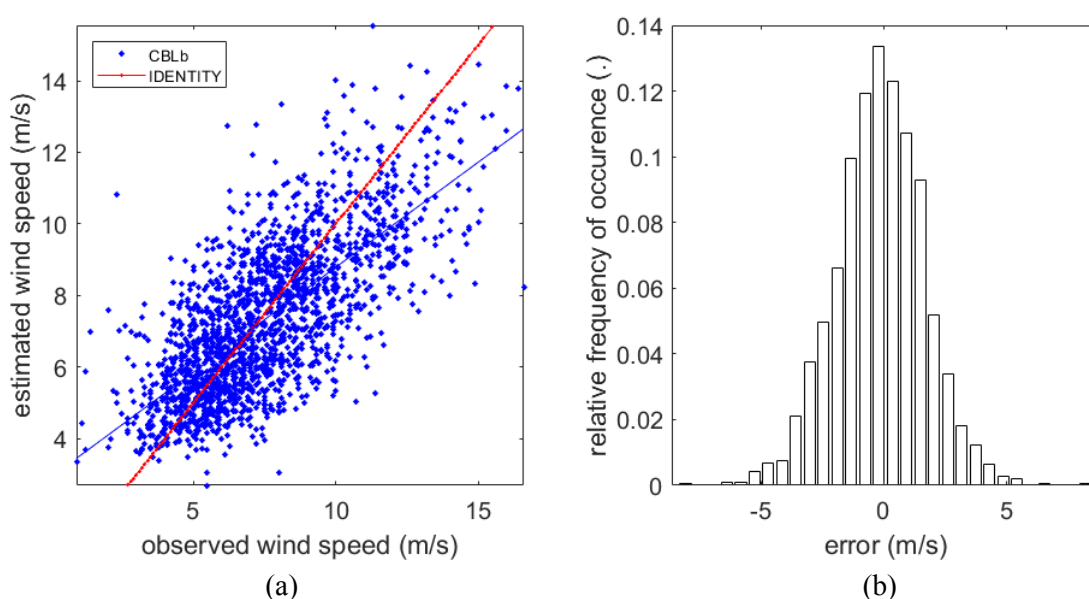
ZORITA, E.; HUGHES, J. P.; LETTEMAIER, D. P.; VON STORCH, H. Stochastic characterization of regional circulation patterns for climate model diagnosis and estimation of local precipitation. *Journal of Climate*, v. 8, n. 5, p. 1023-1042, 1995.

ZORITA, E.; VON STORCH, H. The analog method as a simple statistical downscaling technique: comparison with more complicated methods. *Journal of Climate*, v. 12, n. 8, p. 2474-2489, 1999.

APÊNDICE A – CASO 2: RESULTADO DO MODELO CBLb

A Figura A.1a mostra o resultado de CBLb em um gráfico de dispersão, para a escala horária, em comparação com as observações (abscissas). Note que, em geral, os modelos atingem menores valores de correlação em comparação com o Caso 1 (veja a Figura 5.9a, Seção 5.1). Como no caso anterior, o histograma de erro (Figura A.1b) sugere que o viés na estimativa do valor esperado do período de validação é pequeno, apesar de suavização (perda de variância). Note, porém, que uma ligeira tendência de subestimação do valor esperado das observações ocorre, o que fica mais evidente nos resultados para o regime diário (discutidos mais adiante).

Figura A.1. Caso 2: Diagramas de Dispersão e Histograma de Erros (escala horária).

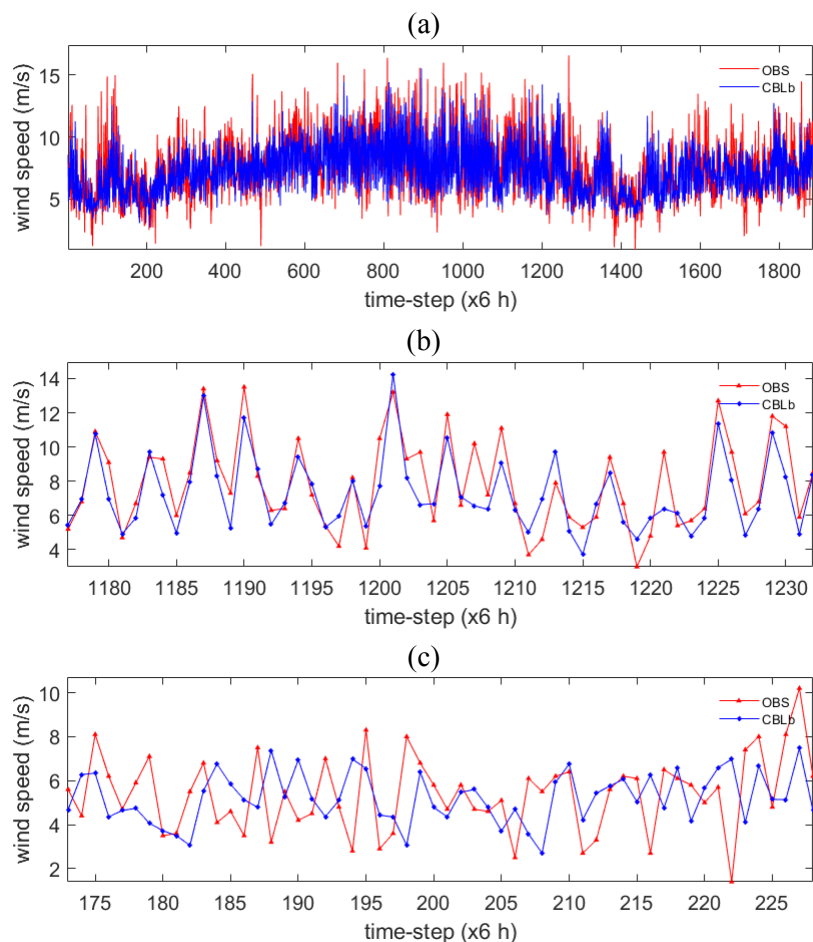


(a) dispersão entre saídas de modelo com maior SS4 (CBLb, na ordenada) e observações (abscissa); (b) histogramas de erro de CBLb (frequência relativa de ocorrência versus erro). Fonte: Própria.

A Figura A.2 mostra três períodos diferentes de séries temporais CBLb. Como no Caso 1, aqui a primeira janela (Figura A.2a) abrange toda a série temporal (período de validação), mostrando um bom acordo global entre o modelo e as observações no tocante às baixas frequências. As outras janelas, nas Figuras A.2b e A.2c, mostram exemplos de momentos em que o CBLb está bem ajustado e mal ajustado às observações, respectivamente. Em relação às aplicações CBLb, o resultado mostrado sugere uma conclusão semelhante à do Caso 1, ou seja, o comportamento típico das frequências baixas e médias está sendo bem reproduzido, embora

algumas melhorias sejam necessárias para aplicações de alta frequência (como em modelos operacionais de previsão).

Figura A.2. Caso 2: Superposição de Séries Temporais (escala horária).

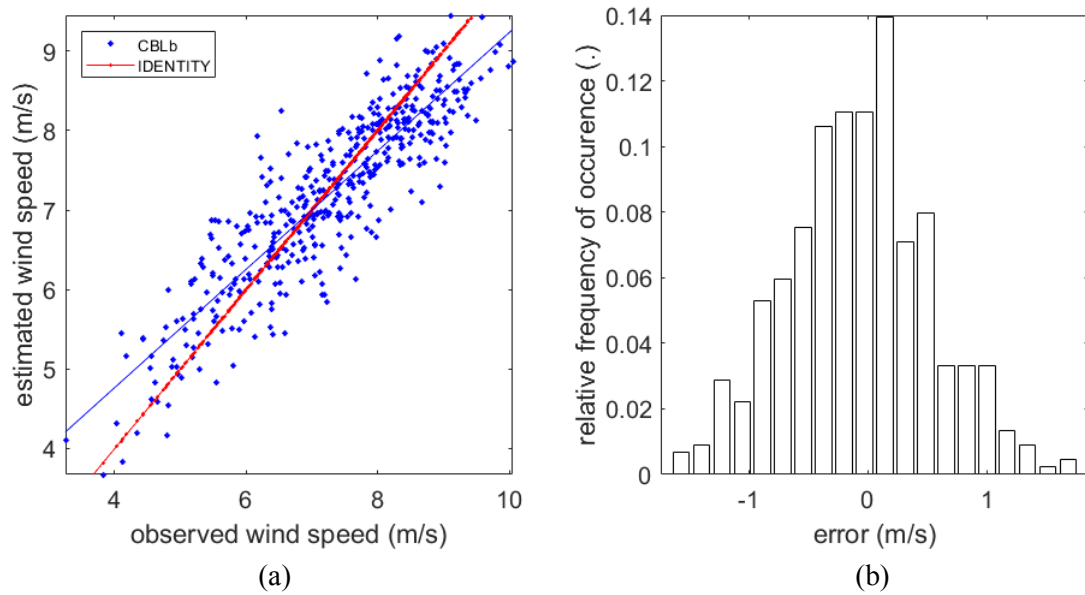


(a) séries temporais do modelo com maior SS4 (CBLb) e observações; (b) zoom em uma janela com um bom ajuste entre modelo e observação; (c) zoom em uma janela com um ajuste ruim entre modelo e observações.

Fonte: Própria.

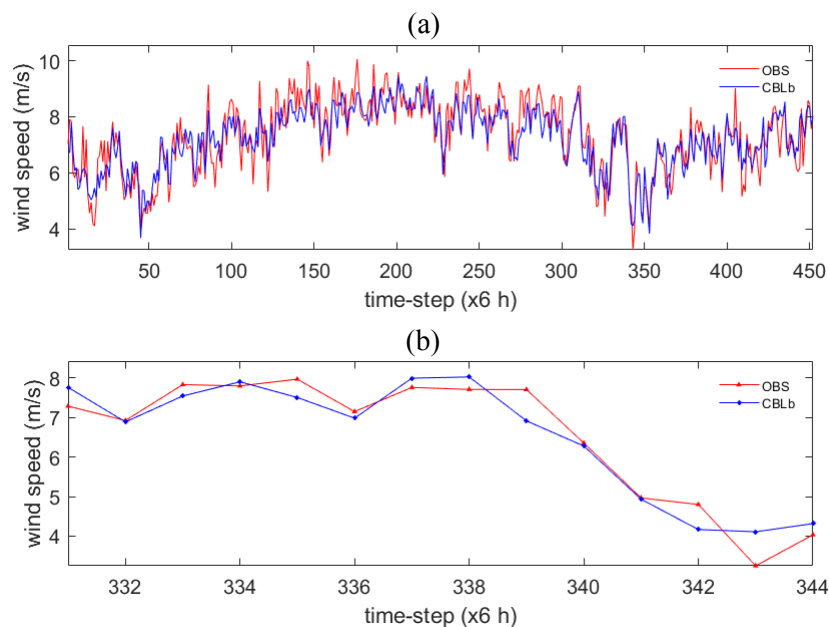
As Figuras A.3 e A.4 mostram os resultados obtidos pelo CBLb relativos ao erro, tal como definido anteriormente. Aqui, note que, novamente, a estimativa derivada de CBLb está altamente correlacionada às observações (Figura A.3a). O histograma de erro mostra que o valor esperado do erro introduz uma ligeira tendência de subestimação do valor observado (Figura A.3b). Tal como na escala horária (Seção 5.1), apesar das limitações do ponto de vista local (ver Figura A.4c), CBLb exibe um bom desempenho em relação à amplitude de variação e médias e baixas frequências (Figuras A.4a e A.4b).

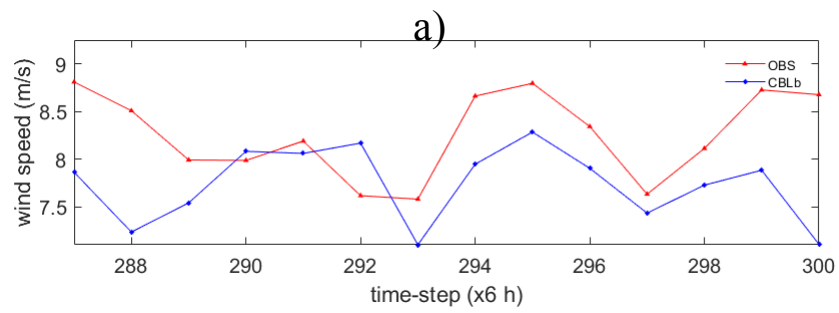
Figura A.3. Caso 2: Diagramas de Dispersão e Histograma de Erros (escala diária).



(a) dispersão entre saídas de modelo com maior SS4 (CBLb, na ordenada) e observações (abscissa); (b) histogramas de erro de CBLb (frequência relativa de ocorrência versus erro). Fonte: Própria.

Figura A.4. Caso 2: Superposição de Séries Temporais (escala diária).





(a) séries temporais do modelo com maior SS4 (CBLb) e observações; (b) zoom em uma janela com um bom ajuste entre modelo e observação; (c) zoom em uma janela com um ajuste ruim entre modelo e observações.

Fonte: Própria.

APÊNDICE B – ESTUDOS DE CASO: RESULTADOS GERAIS

Este apêndice apresenta, de forma resumida, os resultados obtidos por meio das estratégias de *downscaling* estatístico (incluindo combinação de modelos) avaliadas nesse estudo. As informações referentes às estações anemométricas apresentadas nessa seção, bem como os dados de simulação atmosférica proveniente do GCM empregados na análise estão disponíveis na Figura 5.1 e Tabela 1 (Seção 4, Metodologia). Além disso, todas as técnicas aplicadas à modelagem do vento local estão descritas nas Seções 4.1 e 4.2. Diagramas de Taylor serão adicionados ao Apêndice em etapas futuras para uma discussão mais detalhada sobre os resultados obtidos para as demais localizações.

A Tabela B.1 abaixo mostra que na grande maioria dos casos, o modelo CBLb produziu os melhores resultados. Na maior parte dos casos, os resultados são obtidos a partir da combinação das variantes de modelo MLR, CL e NN, as quais, em geral, produzem os melhores resultados individuais. Apenas em dois casos, dados pelos índices 6 e 10, em escala horária, o CBLp foi capaz de superar o CBLb, assim como o CBnL para a estação de índice 12, em regime diário. Tal como mencionado na seção de metodologia, um possível motivo para a eficácia mostrada pelo CBLb é o procedimento exaustivo na qual são parametrizados todos os subconjuntos possíveis, e, em seguida, selecionado o que apresenta a melhor performance. Esse tipo de busca é proibitivo no caso de técnicas que demandam um maior esforço computacional, como o CBnL.

Em um único caso (índice 8, regime horário), os resultados da interpolação produziram os melhores resultados gerais. Um estudo mais aprofundado é necessário para que possa ser determinada uma possível causa.

Tabela B.1. Melhores Resultados em Todas as Estações Consideradas.

Índice da Estação Anemométrica	Escala Horária			Escala Diária		
	Melhor Modelo	SS4	Aprimoramento de SS4 (sobre IBL) (%)	Melhor Modelo	SS4	Aprimoramento de SS4 (sobre IBL) (%)
#1	CBLb	0,7412	159,6107	CBLb	0,8653	6,2899
#2	CBLb	0,7755	13,9921	CBLb	0,9321	6,0718
#3	CBLb	0,5180	142,0265	CBLb	0,7714	32,1369
#4	CBLb	0,6141	57,9471	CBLb	0,7720	1,9611

#5	CBLb	0,4511	17,8002	CBLb	0,5932	24,4075
#6	CBLp	0,4957	0,1320	CBLb	0,7705	2,9588
#7	CBLb	0,5544	58,21	CBLb	0,5539	6,4667
#8	IBL	0,4210	0	CBLb	0,6030	28,2492
#9	CBLb	0,5283	0,9769	CBLb	0,8702	12,9831
#10	CBLp	0,4180	16,1541	CBLb	0,8220	4,0779
#11	CBLb	0,5521	10,7825	CBLb	0,7584	0,1857
#12	CBLb	0,6429	12,2984	CBnL	0,8496	6,4167
#13	CBLb	0,6952	8,0254	CBLb	0,8857	4,4172

Mais detalhes sobre as estações anemométricas estão disponíveis na Figura 5.1 e na Tabela 1. Fonte: Própria.