sid.inpe.br/mtc-m21c/2021/02.17.21.55-TDI

# DOWNSCALING DA PRECIPITAÇÃO DIÁRIA SOBRE O BRASIL PARA SUB-DIÁRIA BASEADO EM MÚLTIPLOS CONJUNTOS DE DADOS USANDO REDES NEURAIS ARTIFICIAIS

Rogério da Silva Batista

Dissertação de Mestrado do Curso de Pós-Graduação em Computação Aplicada, orientada pelo Dr. Alan James Peixoto Calheiros, aprovada em 25 de março de 2021.

URL do documento original:
<http://urlib.net/8JMKD3MGP3W34R/447AG38>

INPE
São José dos Campos
2021

**PUBLICADO POR:**

Instituto Nacional de Pesquisas Espaciais - INPE
Coordenação de Ensino, Pesquisa e Extensão (COEPE)
Divisão de Biblioteca (DIBIB)
CEP 12.227-010
São José dos Campos - SP - Brasil
Tel.:(012) 3208-6923/7348
E-mail: pubtc@inpe.br

sid.inpe.br/mtc-m21c/2021/02.17.21.55-TDI

# DOWNSCALING DA PRECIPITAÇÃO DIÁRIA SOBRE O BRASIL PARA SUB-DIÁRIA BASEADO EM MÚLTIPLOS CONJUNTOS DE DADOS USANDO REDES NEURAIS ARTIFICIAIS

Rogério da Silva Batista

Dissertação de Mestrado do Curso de Pós-Graduação em Computação Aplicada, orientada pelo Dr. Alan James Peixoto Calheiros, aprovada em 25 de março de 2021.

URL do documento original:
<http://urlib.net/8JMKD3MGP3W34R/447AG38>

INPE
São José dos Campos
2021

Dados Internacionais de Catalogação na Publicação (CIP)

Batista, Rogério da Silva.

Ba32d Downscaling da precipitação diária sobre o Brasil para sub-diária baseado em múltiplos conjuntos de dados usando redes neurais artificiais / Rogério da Silva Batista. – São José dos Campos : INPE, 2021.

xviii + 118 p. ; (sid.inpe.br/mtc-m21c/2021/02.17.21.55-TDI)

Dissertação (Mestrado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2021.

Orientador : Dr. Alan James Peixoto Calheiros.

1. Downscaling. 2. Precipitation. 3. Satellite. 4. ANN. 5. MERGE. I.Título.

CDU 004.82:551.577

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÕES

**INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS**

# DEFESA FINAL DE DISSERTAÇÃO DE ROGERIO DA SILVA BATISTA BANCA Nº 042/2021, REG 115339/2018

No dia 25 de março de 2021, as 14h, por teleconferência, o(a) aluno(a) mencionado(a) acima defendeu seu trabalho final (apresentação oral seguida de arguição) perante uma Banca Examinadora, cujos membros estão listados abaixo. O(A) aluno(a) foi APROVADO(A) pela Banca Examinadora, por unanimidade, em cumprimento ao requisito exigido para obtenção do Título de Mestre em Computação Aplicada. O trabalho precisa da incorporação das correções sugeridas pela Banca Examinadora e revisão final pelo(s) orientador(es).

**Novo Título: "DOWNSCALING DA PRECIPITAÇÃO DIÁRIA SOBRE O BRASIL PARA SUB-DIÁRIA BASEADO EM MÚLTIPLOS CONJUNTOS DE DADOS USANDO REDES NEURAIS ARTIFICIAIS"**

Eu, Rafael Duarte Coelho dos Santos, como Presidente da Banca Examinadora, assino esta ATA em nome de todos os membros, com o consentimento dos mesmos

Dr. Rafael Duarte Coelho dos Santos - Presidente - INPE
Dr. Alan James Peixoto Calheiros - Orientador - INPE
Dr. Daniel Alejandro Vila - Membro Interno - INPE
Dr. Samuellson Lopes Cabral - Membro Externo - CEMADEN
Dr. Álvaro Luiz Fazenda - Membro Externo - UNIFESP

Documento assinado eletronicamente por **Rafael Duarte Coelho dos Santos**, **Tecnologista**, em 29/03/2021, às 13:32 (horário oficial de Brasília), com fundamento no art. 6º do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.mctic.gov.br/verifica.html, informando o código verificador **6816408** e o código CRC **633E9463**.

## AGRADECIMENTOS

Agradeço a Deus e a minha família em especial meus filhos Isabela e Pedro, minha principal motivação, aos quais me cederam preciosos momentos de sua infância para elaboração deste trabalho. Também a minha esposa Elaine por toda paciência em meio ao caos ocasionado pela pandemia do COVID19, o que nos levou a reinventar nossas vidas em frente aos novos desafios. Aos meus pais, Marcílio e Leotildes que me apoiaram durante todo o período de mestrado, e a todos meus amigos e colegas de trabalho que me incentivaram nesse período em especial ao meu orientador Dr. Alan que perseverou comigo durante o percurso.

## RESUMO

A precipitação é uma variável de extrema importância para a sociedade, sua intensidade e persistência são responsáveis por causar enchentes, deslizamentos e até mesmo mortes. Para um monitoramento eficiente da precipitação em uma determinada região é necessário conhecer seu ciclo diurno. No Brasil, devido à baixa densidade de dados observacionais em superfície, tanto por redes de pluviômetros (que possui mais estações diárias do que horárias) quanto por radares meteorológicos, é necessária a utilização de produtos de estimativa de chuva via satélites. No entanto, o erro ainda é alto para este tipo de estimativa. Neste contexto, este estudo analisou técnicas de Inteligência Artificial, especificamente as Redes Neurais Artificiais (RNA), no downscaling de dados diários para uma escala sub-diária utilizando múltiplos conjuntos de dados. As principais informações sobre a precipitação diária vêm da estimativa de satélites corrigida por pluviômetros, técnica denominada MERGE. Para avaliar a representatividade do ciclo diurno e dos processos físicos nas diferentes regiões do país foram aplicados dois tipos de RNA, a Rede Neural Profunda (DNN) e a Rede Neural Recorrente (RNN), sendo o alvo a precipitação sub-diária com resolução temporal de 3 horas. Foram selecionadas variáveis meteorológicas que possuem relação física com a chuva, provenientes de satélites meteorológicos como temperatura de brilho infravermelho do satélite GOES, estimativa de precipitação horária por sensores de micro-ondas (IMERG) e dados ambientais (por exemplo, umidade, vento, etc) do modelo de reanálise ERA5. Além disso, foram utilizadas informações de topografia e localização geográfica. Cada uma das variáveis escolhidas foi analisada quanto à correlação com a chuva observada acumulada no mesmo período. Os resultados foram avaliados para diferentes regiões, estações e horários. Conclui-se que os resultados obtidos pelas RNAs foram melhores quando comparados à estimativa de precipitação do IMERG (referência). Para resultados com menos dados de entrada (sem informação de vento), a DNN foi a RNA que teve melhor desempenho, principalmente quando treinada com dados de todas as regiões, apresentando MSE de 11,33 mm e a RNN de 11,88 mm. A validação cruzada mostrou resultados um pouco melhores para o IMERG, mas com superestimativa da precipitação. Além disso, a DNN apresentou melhores resultados para todas as diferentes regiões do Brasil, bem como para as diferentes estações do ano. Os valores de BIAS para RNN foram melhores para horas com baixa precipitação, enquanto a DNN e o IMERG foram melhores para os períodos chuvosos (18 e 21 GMT). No entanto, as diferenças de BIAS entre DNN e RNN foram muito pequenas e o MSE mostrou valores ligeiramente melhores para DNN em todos os horários. Portanto, a DNN foi escolhida como a melhor RNA, seguindo com testes de sensibilidade para determinar a melhor configuração desconsiderando os custos computacionais. Em sua versão aprimorada com a inclusão de mais variáveis meteorológicas, a DNN apresentou melhor desempenho em todos os aspectos, inclusive na de triagem de chuva, quando comparado ao IMERG.



Palavras-chave: Downscaling. Precipitação. Satélite. RNA. MERGE.

**DAILY TO SUB-DAILY PRECIPITATION DOWNSCALING OVER BRAZIL BASED ON MULTIPLE DATASETS USING ARTIFICIAL NEURAL NETWORKS**

## ABSTRACT

Precipitation is an extremely important variable for society, its intensity and persistence are responsible for causing floods, landslides and even deaths. For an efficient monitoring of the precipitation over a certain region, it is necessary to know its diurnal cycle, and a sub-daily measurement is required. In Brazil, due to the low density of ground observational data, both networks of rain gauges (more daily data than hourly) and weather radars, it is necessary to use satellite rain estimation products. However, the error is still high for this kind of precipitation estimation. In this context, this study analyzed Artificial Intelligence techniques, specifically Artificial Neural Networks (ANN), for the downscaling of the daily data to a sub-daily scale using multiple datasets. The main information about the daily rainfall comes from satellites estimation corrected by rain gauges, a technique called MERGE. In order to represent the characteristics of the diurnal cycle and the physical processes of the different regions of the country we applied two types of ANN, the Deep Neural Network (DNN) and the Recurrent Neural Network (RNN). The target is a sub-daily rainfall with temporal resolution of 3 hours. Meteorological variables with physical relationship with the rain were selected, most coming from meteorological satellites, like infrared brightness temperature from GOES satellite, hourly precipitation estimation from microwave sensors (IMERG), and environmental data (e.g. humidity, wind, etc) from ERA reanalysis. Also, we used topography and location information for the whole area. Each of the chosen variables was pre-processed, producing averages, accumulated and other measures for the 3-hour resolution and it was analyzed their correlation with the accumulated observed rain at the same time. The results were evaluated for different regions, seasons, and times. It is concluded that the results obtained by the ANNs were better when compared to the precipitation estimation from IMERG (the reference). For results with less input data (without wind information), to save computer time, the DNN was the one with the best performance, especially when trained with data from all regions. DNN obtained an MSE of 11.33 mm and RNN shows a value of 11.88 mm. However, the rain screen was slightly better to IMERG, but with superestimation of the precipitation. Also, DNN showed better results for all the different regions of Brazil as well as for the different seasons. The BIAS results for RNN were better for hours with low precipitation, while DNN and IMERG were better for rainy periods (18 and 21 GMT). However, the BIAS differences between DNN and RNN were very small and MSE shows a slightly better values to DNN for all times. Therefore, DNN was chosen as the best ANN. Sensitivity tests were carried out to determine the best DNN configuration without considering computational costs. In its improved version with the inclusion of more meteorological variables, DNN performed better in all aspects, including that of rain screening, when compared to IMERG.

Keywords: Downscaling. Precipitation. Satellite. ANN. MERGE

## LISTA DE FIGURAS

**Pág**.

## LISTA DE TABELAS

**Pág**.

**SUMÁRIO**

**Pág**.

# 1 INTRODUÇÃO

A chuva é a principal fonte de água para abastecimento e manutenção dos rios, lagos, barragens e reservatórios do Brasil, mudanças significativas no ciclo hidrológico impactam diretamente em diversas atividades produtivas da sociedade, incluindo setores que vão desde a geração de energia elétrica e abastecimento de água até a agroindústria e turismo, causando prejuízos à economia do país, como foi o caso da crise hídrica que abalou a região Sudeste nos anos 2014 e 2015, e da crise energética nos anos 2001 e 2002.

Caracterizada pela ausência de planejamento adequado para o gerenciamento do recurso hídrico e a ausência de consciência coletiva dos consumidores brasileiros para o uso racional da água, o ocorrido de 2015 foi uma crise anunciada segundo Marengo et al. (2015), que ocorreu devido aos baixos acumulados pluviométricos, em particular, sobre a região do sistema Cantareira, localizado na divisa entre os estados de São Paulo e Minas Gerais, e do histórico de situações hídricas semelhantes como, por exemplo, a chamada “crise do apagão” ocorrida nos anos 2001 e 2002. Segundo Marengo et al. (2015) um bloqueio atmosférico atípico que atuou durante 45 dias consecutivos impedindo a transferência de umidade entre as regiões, foi responsável pelas falhas nas previsões, segundo eles os anticiclones de bloqueio geralmente tem duração entre 7 a 8 dias e muito raramente ultrapassam os 15 dias, apesar disso as estações de superfície e as estimativas de precipitação via satélite indicavam uma anomalia negativa na chuva, que poderia ter sido utilizada como indicativo de alerta, induzindo o estado a gerir medidas preventivas como racionamento da água e a promoção do reuso, entre outras medidas que chegaram a ser adotadas, porém tardiamente.

As estiagens, assim como a chuva em excesso, são responsáveis por problemas de interesse público, uma vez que setores como a agricultura dependem da chuva para irrigação das lavouras, quando ela não ocorre, recorrem a soluções alternativas para captação e armazenamento de água, como cisternas, poços artesianos, barragens subterrâneas e até mesmo a

compra de água transportada em caminhões pipa, porém muitos agricultores deixam de irrigar suas plantações, o que resulta na diminuição ou na perda do produto. Durante a estiagem também é comum à redução da umidade no ar e consequentemente o aumento do número de queimadas, impactando no número de doenças respiratórias, intoxicações e acidentes causados pela fumaça, segundo Grigoletto et al. (2016), a seca prolongada afeta milhões de pessoas e contribui para a fome, pobreza e desnutrição, e é causador de surtos de doenças infectocontagiosas e respiratórias além de outros agravantes.

O excesso de chuva, além de prejudicar as plantações, também traz dificuldades na colheita e no transporte, além disso, nessas situações, é comum às chuvas transportarem as camadas superficiais do solo, ricas em matéria orgânica e fertilizantes, causando erosão e empobrecimento do solo, muitas vezes transportando material até áreas de água próximas como rios e lagos causando também alagamentos e assoreamentos. Outros setores da economia como turismo e construção civil também são afetados devido a problemas em estradas como erosões, queda de barreiras, enchentes, entre outros, além de causar problemas com saneamento básico e infraestrutura de drenagem.

É certo que a falta ou o excesso de chuva centralizado em uma determinada região causará algum prejuízo diretamente ou indiretamente aos órgãos públicos e ao setor privado, e por isso há uma necessidade de conhecer os ciclos hidrológicos de áreas de interesse, para monitorar variações em seus índices pluviométricos a fim de prever alterações significativas e prover ações para minimizar os prejuízos.

Segundo a Agência Nacional de Águas e Saneamento Básico (ANA), 49,2% dos municípios brasileiros decretaram Situação de Emergência ou Estado de Calamidade Pública devido a cheias e 51,1% devido à seca ou estiagem, pelo menos uma vez no período entre 2003 a 2019, ainda segundo ela somente em 2019, um total de dois milhões de brasileiros foram afetados por cheias, considerando alagamentos, enxurradas e inundações, e 22 milhões por secas e estiagem (ANA, 2020).

Quando a crise hídrica ocorreu em 2014-2015 havia diversos dados de precipitação disponíveis publicamente em tempo quase real, tanto por órgãos nacionais quanto internacionais, e que poderiam ter sido utilizados no monitoramento da situação, contudo, unicamente essa informação não é suficiente para diagnosticar uma condição anômala. Para isso, tão importante quanto conhecer o valor instantâneo da chuva é necessário conhecer o Ciclo Diário de Precipitação (CDP), ou a curva moduladora representativa do comportamento da chuva em 24 horas, onde com base nessas duas informações, é possível identificar situações onde a chuva se distancia da normalidade, calcular a intensidade e a persistência dos padrões, classificá-los e com isso prever as situações que podem vir a causar prejuízos à sociedade.

Segundo dados do Instituto Brasileiro de Geografia e Estatística (IBGE), em dezembro de 2020 o país tinha 5.570 municípios, onde alguns mesmo que próximos geograficamente, podem apresentar características distintas, como topografia, propriedades de solo, vegetação, uso da terra, hidrografia, qualidade da água, entre outras, o que pode influenciar no comportamento local do CDP.

Tradicionalmente o CDP é estimado utilizando uma série climatológica de precipitação sub-diária, que geralmente é obtida através de modelos numéricos de previsão de tempo e clima, estimativas provenientes de satélites e/ou radares meteorológicos, extrapolação de medidas diretas provenientes de pluviômetros e/ou pluviógrafos, ou ainda da interpolação entre as diversas estimativas, contudo, cada um desses métodos possui limitações. Os valores aferidos diretamente em pluviômetros, por exemplo, são os mais próximos da realidade, inclusive são comumente utilizados em validações de outras técnicas de estimativa, porém, devido à alta variabilidade da chuva essas medidas são válidas apenas para um pequeno espaço geográfico, logo, para projeção em uma área maior é necessário um número de estações proporcional à extensão da área de interesse.

Ao estudar determinados fenômenos meteorológicos, como o CDP, muitas vezes faz-se necessário a análise de séries temporais longas, contudo a

crescente evolução tecnológica aumentou significativamente a capacidade humana de produzir, armazenar e processar informação, e dentro do contexto meteorológico não foi diferente, atualmente os satélites, radares e outros instrumentos meteorológicos são capazes de produzir informações com resoluções espaciais e temporais cada vez maiores, porém como consequência disso, os dados mais antigos das séries temporais têm resoluções espaciais e temporais menores. Neste sentido, com a necessidade de obter séries temporais longas e com resoluções maiores, comumente os dados de épocas passadas são estimados ou interpolados.

Segundo Kumar et al. (2012) existem basicamente duas soluções para geração de dados meteorológicos em uma resolução maior, uma delas é o *downscaling* dinâmico que utiliza modelos numéricos em mais alta resolução (e.g., WRF) para simular os efeitos locais em uma escala sub-diária, e a segunda é o *downscaling* estatístico, utilizado para extrapolar dados para uma frequência de amostras mais refinada com base em um cenário climático. De acordo com Khan et al. (2006), os métodos de *downscaling* estatísticos como regressão linear, regressão não linear e estimador estocástico são mais simples e mais "baratos" computacionalmente de serem implementados do que as técnicas de *downscaling* dinâmicas.

Com base nisso a proposta deste trabalho é aplicar os conceitos de Rede Neural Artificial (RNA) no *downscaling* temporal de dados de precipitação diária, do produto denominado MERGE do Instituto Nacional de Pesquisas Espaciais (INPE), em valores de precipitação sub-diária, utilizando para isso múltiplos conjuntos de dados provindos de satélites e modelos de reanálise, buscando obter estimativas sub-diárias qualitativamente próximas às medidas diretas de estação de superfície para toda extensão do território Brasileiro.

Os resultados neste estudo foram validados espacialmente e temporalmente, produzindo métricas estatisticamente boas que foram analisadas e comparadas ao *Integrated Multi-Satellite Retrievals for GPM* (IMERG), uma estimativa de precipitação amplamente utilizada dentro do contexto meteorológico, onde foram avaliadas os erros das estimativas por horário, região e estação do ano.

Os resultados foram significativamente melhores que o IMERG em todas as avaliações propostas mostrando que com as RNAs é possível produzir uma série de dados de estimativa de precipitação sub-diária longa e com qualidade para o cálculo do CDP sobre o Brasil, e que eventualmente pode ser utilizada no monitoramento hidrometeorológico.

# 2 OBJETIVOS

Este trabalho tem como objetivo principal avaliar a eficácia das RNAs no *downscaling* de dados de precipitação diária, para produção de um conjunto de dados de estimativas de precipitação sub-diária, para o cálculo do CDP sobre regiões estratégicas do Brasil. Para alcançar este objetivo principal foi necessária a execução de diversos objetivos específicos, que são listados abaixo.

## 2.1 Objetivos específicos

Os objetivos específicos deste trabalho foram:

a) realizar o *downscaling* de dados diários de precipitação, calibrados com dados de estações de superfície, para uma escala sub-diária (3 horas). Para isso foi aplicada uma metodologia de *downscaling* temporal utilizando uma RNA, treinada para produzir dados com as mesmas características de dados aferidos em estações de superfície;

b) verificar as variáveis já utilizadas na literatura para estimativa da precipitação e sua importância para caracterizar as chuvas sobre o Brasil a partir de análises exploratórias dos dados. Avaliou-se a correlação síncrona das variáveis com a chuva, o espalhamento e a distribuição dos dados, entre outros;

c) definir uma RNA compatível com os dados e capacidade computacional para este estudo. No processamento foi utilizado um computador com processador Intel® Core™ i7-6700 com 3.41GHz, 16,0 GB de memória RAM, e disco rígido de 10,0 TB;

d) definir o desempenho da RNA para diferentes condições, aspectos climáticos e regionais. Neste caso foram utilizadas informações de precipitação horária proveniente de estações de superfície automáticas para as diferentes regiões do Brasil e estações do ano, além de testes de sensibilidade para os diferentes dados de entrada.

# 3 REFERENCIAL TEÓRICO

## 3.1 Estimativa dos campos de precipitação instantânea

A precipitação é a variável principal desta pesquisa, segundo Ahrens (2019) trata-se da água precipitada na superfície terrestre em qualquer forma como chuva, neve, granizo, orvalho, entre outros. Sua formação consiste de partículas de água ascendida à atmosfera através da evapotranspiração e absorvida por núcleos higroscópicos, que se condensam, devido ao resfriamento do ar, e crescem, devido ao choque com outras partículas, até atingir uma condição de saturação e consequentemente sua queda. O resfriamento do ar úmido pode ocorrer através da colisão frontal de uma massa de ar frio com uma massa de ar quente, resultando em uma chuva frontal ou ciclônica, ou através de uma convecção térmica, resultado da ascensão da umidade que é resfriada pela própria altitude, causando as chamadas chuvas convectivas, ou ainda por condições de relevo, onde as nuvens colidem com terrenos montanhosos e ascendem, causando chuvas denominadas orográficas ou de relevo.

Os três tipos de chuva, frontal, convectiva e orográfica tem características microfísicas bem definidas e podem ser detectadas por equipamentos como sensores micro-ondas e infravermelhos, contudo existem outros tipos de sistemas difíceis de serem identificados através dessas tecnologias como é o caso das nuvens quentes, típicas de regiões tropicais costeiras como o nordeste brasileiro, onde não se observa partículas de gelo em sua formação. Ambos os tipos de chuva podem ser observados através de estações pluviométricas, que são equipamentos de superfície que medem a precipitação diretamente, contudo esses equipamentos são escassos para produzir estimativas em grandes áreas.

Existem ainda diferentes maneiras e métodos de definir as taxas de chuva associadas a sistemas meteorológicos. As medidas de superfície podem ser interpoladas e as estimativas podem ser calculadas através de diferentes algoritmos e técnicas de processamento. Nesta seção serão apresentadas

algumas metodologias de estimativa dos campos de chuva para diferentes plataformas.

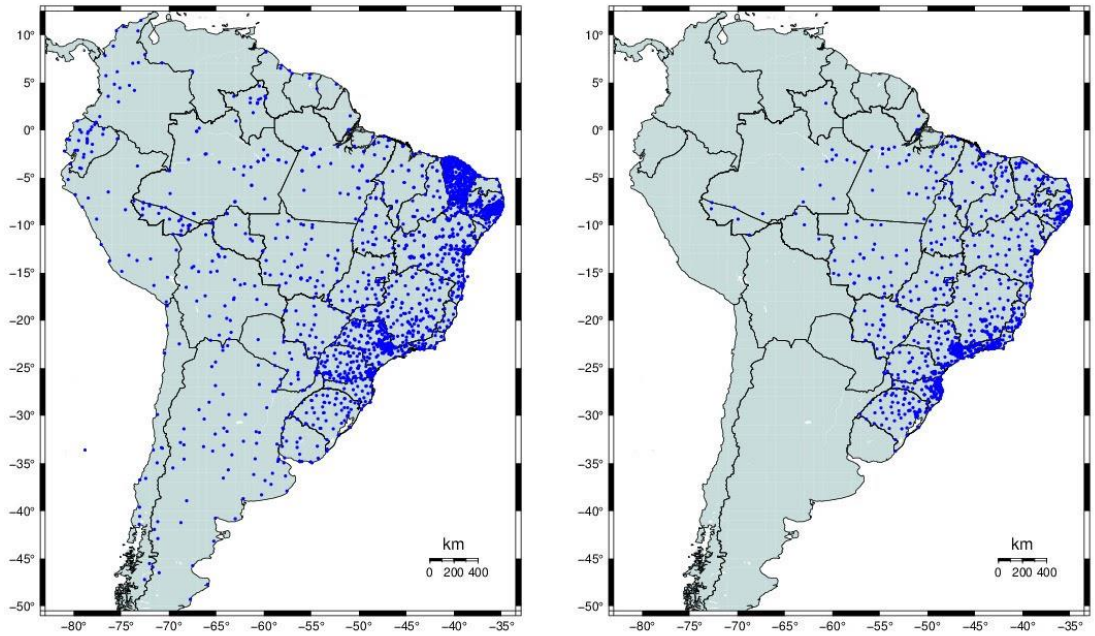
### 3.1.1 Estações pluviométricas

A precipitação pode assumir várias formas, porém no Brasil a que ocorre com mais frequência e consequentemente causa maiores problemas é a precipitação em forma de chuva, que tradicionalmente é medida através de equipamento chamado estação pluviométrica ou pluviômetro. Este equipamento possui as medidas mais precisas devido ao fato de medir diretamente o valor através de sensores posicionados em superfície logo abaixo da ocorrência. Contudo, devido a grande variabilidade da chuva, a medida abrange uma área limitada com cerca de 1 a 10 m², sendo necessária uma grande densidade de equipamentos para produzir uma informação de área representativa. Além disso, apesar de terem um custo relativamente baixo comparado aos demais equipamentos (satélites e radares), esses equipamentos precisam estar instalados ao ar livre e precisam rotineiramente de ajustes e calibrações, além de que, tanto os equipamentos quanto o local onde estão instalados necessitam de manutenções constantes para impedir interferências, como a obstrução do funil que captura a água da chuva, crescimento de vegetação próximo aos equipamentos impedindo que a chuva precipite diretamente sobre a área de coleta, depredação dos sensores, entre outros. Devido a esses fatores, e o Brasil sendo um país de proporções continentais, as redes pluviométricas disponíveis são esparsas e muitos dos equipamentos instalados estão inoperantes ou descalibrados.

Segundo Kidder e Haar (1995) essa característica pontual das estações pluviométricas combinado a uma distribuição esparsa e irregular são fatores que causam incertezas no entendimento dos processos evolutivos dos sistemas precipitantes e de sua variabilidade espaço-temporal. Com relação aos valores aferidos, apesar de se tratar de uma medida direta, alguns fatores naturais como o vento e a evaporação interferem no processo de aferição fazendo com que os dados por vezes sejam subestimados.

Quanto à transmissão das informações coletadas, alguns equipamentos mais modernos possuem transmissão automática via satélite ou internet, sendo comum ocorrer em uma escala sub-diária, contudo a maioria dos equipamentos brasileiros é do tipo convencional e não possuem esta funcionalidade, sendo a transmissão das informações realizada em escala diária, geralmente às 12 horas GMT, sendo em algumas ocasiões realizada manualmente através de interferência humana, o que também pode vir a ser um fator de erro. A Figura 3.1 mostra um exemplo da diferença significante entre a quantidade de estações convencionais e automáticas no Brasil.

Figura 3.1: Densidade de estações de superfície utilizadas pelo INPE.




Distribuição das estações pluviométricas utilizadas pelo INPE na geração da estimativa de precipitação MERGE de diferentes fontes (i.e. *Global Telecommunication System* (GTS), INMET, INPE e centros regionais) para o dia 16 de maio de 2019: (a) dados diários de estações convencionais e automáticas, e (b) dados horários de estações automáticas.

Fonte: Produção do autor.

Para validar valores suspeitos e filtrar ruídos, alguns métodos baseados em estações pluviométricas de localidade próximas, podem ser aplicados para melhorar a qualidade das informações ou estimar o valor quando o equipamento apresenta falhas (ANA, 2012).

Alguns desses métodos utilizados pela ANA são:

a) **método da ponderação regional:** Este método segundo Tucci (2004) é um método simples baseado em regressão linear que permite o preenchimento em séries mensais ou anuais, onde, são utilizadas pelo menos três estações com no mínimo dez anos de dados e em uma região climatológica semelhante ao local a ser preenchido;

b) **método da regressão linear:** Segundo Tucci (2004), este método consiste na aplicação de regressões lineares simples ou múltiplas, utilizando informações pluviométricas de estações vizinhas correlacionadas;

c) **método da dupla massa:** Desenvolvido pelo Serviço Geológico dos Estados Unidos (USGS) em 1966, segundo a ANA (2012) é o método mais adotado no Brasil para séries mensais e anuais. Consiste basicamente em plotar em um gráfico cartesiano com os valores acumulados da estação a serem validados no eixo das ordenadas e de outra estação confiável no eixo das abcissas, observando erros sistemáticos, erros de transição ou estações com diferentes regimes pluviométricos.

Além dos métodos de preenchimento de falhas é possível a utilização de interpoladores para estimar o comportamento da precipitação em áreas onde não se verifica a presença de estações pluviométricas, Camaro et al. (2004) destaca três abordagens principais:

a) **modelos determinísticos de efeitos locais:** Cada ponto da superfície é estimado a partir da interpolação das medidas mais próximas;

b) **modelos determinísticos de efeitos globais:** Para a caracterização do fenômeno em estudo, predomina a variação em larga escala, e a variabilidade local não é relevante;
c) **modelos estatísticos de efeitos locais e globais (*krigagem*):** Cada ponto da superfície é estimado apenas a partir da interpolação das amostras mais próximas, utilizando um estimador estatístico que apresentam propriedades de não serem tendenciosos e minimizarem os erros inferenciais.

Existem atualmente vários interpoladores baseados nessas três abordagens, dentre eles o Inverso da Distância Ponderada (IDW), *Krigagem*, *Spline*, *Trend*, *Barnes*, entre outros, porém, apesar da eficiência comprovada desses métodos, a modificação de uma medida direta infere valores, o que a torna menos confiável.

### 3.1.2 Satélites meteorológicos

Outra maneira de estimar a precipitação é através de satélites meteorológicos, diferentemente das estações de superfície as medidas realizadas pelos sensores são indiretas e consequentemente possuem uma menor acurácia, contudo podem cobrir uma grande área geográfica e com maior resolução temporal.

Segundo a Divisão de Satélites e Sensores Meteorológicos (DISSM) do INPE duas órbitas são comumente utilizadas no monitoramento meteorológico, a equatorial, na qual o satélite, denominado satélite geoestacionário, fica estacionado em um ponto sobre a linha do equador, há cerca de 36.000 km de altitude deslocando-se a mesma velocidade radial da Terra, fazendo com que as observações sejam sempre de uma mesma área com aproximadamente 70° de raio. E a órbita polar, na qual os satélites movimentam-se perpendicularmente em relação ao sentido de rotação da Terra, passando pelos polos ou se aproximando deles aproximadamente a cada hora, sendo posicionados em distâncias muito mais curtas que os geoestacionários, geralmente entre 300 e 2.000 km de altitude, o que permite aferir dados com uma maior resolução espacial.

Os sensores a bordo dos satélites e as técnicas utilizadas para estimativas variam muito de acordo com fenômenos meteorológicos aos quais se deseja estimar, como temperatura, descargas elétricas, vento, umidade, entre outros. Segundo Kidder a Haar (1995), especificamente para a estimativa de precipitação por satélite, duas técnicas são utilizadas, as que utilizam dados de radiância na faixa do visível e/ou infravermelho do espectro eletromagnético e aquelas que utilizam dados de radiância na faixa do micro-ondas. No caso a radiação infravermelha emitida pela Terra sofre forte atenuação por nuvens fazendo com que o valor aferido pelo satélite seja o do topo das nuvens, enquanto a radiação emitida na faixa do micro-ondas consegue penetrar e interagir com os hidrometeoros presentes no interior das nuvens, tornando possível inferir propriedades físicas, como o tipo da precipitação, como gelo, neve, água líquida, poeira entre outros (CALHEIROS, 2013).

Desde o advento dos satélites geoestacionários nos anos 1960, muitas técnicas para estimar precipitação utilizando dados de radiação na faixa do visível e infravermelho foram desenvolvidas e aprimoradas, entre elas estão o auto-estimador (VICENTE et al. 1998), e sua versão atualizada, o hydro-estimator (SCOFIELD et al. 2003), onde em ambos o canal infravermelho, é utilizado para determinar a temperatura no topo das nuvens. Estudos como o de Vicente (1998) mostraram que nuvens com a temperatura do topo abaixo de 210° K são características de tempestades convectivas e que através de relações empíricas é possível estimar sua taxa de precipitação. Já o canal visível é comumente utilizado para observar a estrutura do topo das nuvens e classificá-las, eliminando valores suspeitos, como por exemplo, a detecção de chuva sob uma nuvem do tipo *cirrus*, contudo, este canal espectral é dependente da luz solar sendo disponibilizado apenas em períodos diurnos. Apesar de ser uma solução empírica, e consequentemente a que apresenta o maior erro quantitativo, esse tipo de estimativa tem como vantagem o fato de ter uma alta resolução temporal e abranger uma grande área de cobertura. Por exemplo, o satélite GOES-16, mais recente modelo de satélites geoestacionário da série *Geostationary Operational Environmental Satellite* (GOES), atualmente localizado em -75.0° de longitude, consegue observar todo

o continente americano com uma resolução temporal de até 5 minutos e resolução espacial de 2 km no canal infravermelho e 500 metros no canal visível, produzindo informações com as características essenciais para o monitoramento e previsão de tempestades em curto prazo.

Quanto aos sensores de micro-ondas, estes são amplamente utilizados a bordo de satélites de órbita polar devido à sua proximidade com a Terra. Em dezembro de 1972 foi lançado o satélite Nimbus-5, carregando o radiômetro *Electrically Scanning Microwave Radiometer* (ESMR) que detectava radiação polarizada horizontalmente na frequência de 19,35 GHz. Wilheit et al. (1977) utilizaram esses dados para estimar precipitação sobre o oceano através de um modelo de transferência radiativa. Seu sucessor, o Nimbus-6 lançado em junho de 1975 teve várias mudanças como a inclusão da polarização vertical e um canal na frequência de 37 GHz, que possibilitou a estimativa da precipitação também sobre o continente (RODGERS et al. 1979). Em 1978, o sensor ESMR foi substituído pelo *Scanning Multichannel Microwave Radiometer* (SMMR) a bordo do Nimbus-7, este por sua vez realizava medidas através de cinco canais do micro-ondas com diferentes frequências: 6,63; 10,69; 18; 21; e 37 GHz.

Em 1997, a NASA em parceria com a *Japan Aerospace Exploration Agency* (JAXA) criaram uma missão conjunta denominada *Tropical Rainfall Measuring Mission* (TRMM) voltada ao estudo da precipitação principalmente nos trópicos. Nessa missão foi lançado ao espaço um satélite de órbita baixa, que carregava pela primeira vez um radiômetro de micro-ondas ativo denominado *Precipitation Radar* (PR, radar de precipitação em português). Segundo Kummerov et al. (2000) o PR foi projetado para fornecer mapas tridimensionais da estrutura das chuvas, com resolução horizontal de cerca de 5 km e uma largura de faixa de 247 km, possuía capacidade de capturar dados de intensidade, distribuição, tipo e profundidade da precipitação. Além do PR, o satélite TRMM possuía ainda outros quatro sensores, o imageador *Visible and Infrared Scanner* (VIRS), que possui dois canais visíveis centrados nos comprimentos de onda 0,63 e 1,6 μm, mais três canais infravermelhos centrados em 3,75, 10,8 e 12 μm; o radiômetro de micro-ondas passivo *TRMM*

*Microwave Imager* (TMI), com multicanais nas frequências 10,65, 19,35, 22,235, 37 e 85,5 GHz; o *Cloud and Earth Radiant Energy Sensor* (CERES), que funcionou apenas por dois anos após o lançamento, e que foi utilizado no estudo da troca de energia entre o Sol, a atmosfera e a superfície terrestre; e o *Lightning Imaging Sensor* (LIS), sensor óptico utilizado para detectar descargas elétricas. Apesar da previsão inicial de três anos de vida, o satélite TRMM resistiu por 17 anos, gerando inúmeras e significativas pesquisas, assim como, novos e melhores métodos para estimativa de precipitação, como Huffman et al. (2007), que buscando criar a melhor estimativa de precipitação por satélite, desenvolveram o *TRMM Multisatellite Precipitation Analysis* (TMPA), uma estimativa criada a partir de um conjunto de algoritmos que utiliza tanto os dados de micro-ondas, quanto os dados do infravermelho do satélite TRMM combinados a estimativas de precipitação de outros satélites e dados de estações de superfície quando disponíveis.

Para atestar a eficácia dos dados do TRMM, muitas pesquisas foram realizadas ao longo dos anos em que esteve operacional. No Brasil Pereira et al. (2013) observaram que a qualidade das estimativas está relacionada a variabilidade espacial, comparando 13 anos de dados mensais do TMPA com dados de 183 estações meteorológicas espalhadas pelo território Brasileiro, o estudo apresentou índice de concordância média de 96%, porém os autores observaram valores mensais superestimados de 9% no Nordeste e 13% para o Sudeste.

Em um trabalho semelhante, Melo et al. (2015) compararam uma série temporal de 14 anos do TMPA a dados em grade de precipitação gerados com aproximadamente 3.625 pluviômetros e 735 estações meteorológicas, distribuídos por todo o Brasil, que mostraram uma grande variação na qualidade da estimativa, conforme a variação temporal (i.e. mensal ou diária) e ainda conforme a variação regional (Norte, Nordeste, Centro Oeste, Sul e Sudeste).

Outras pesquisas também foram realizadas em regiões menores como para o estado do Amazonas, realizada por Almeida et al. (2015) que compararam

dados do TMPA a estações meteorológicas convencionais para a região durante os anos de 2004 a 2008, e obtiveram uma alta correlação linear 83% e alto índice de concordância 85%.

Para os estados de Mato Grosso, Mato Grosso do Sul, Goiás e Distrito Federal Danelichen et al. (2013) compararam dados do TMPA a séries de precipitação aferidas pelo Instituto de Controle de Espaço Aéreo (ICEA) do Comando da Força Aérea, onde neste estudo a precipitação anual estimada pelo satélite TRMM foi superestimada entre 0,6 e 37,4%, no entanto apresentou alta correlação com a série medida de 88%.

A variação sazonal também foi muito questionada em estudos como Soares et al. (2016) que avaliaram a qualidade dos dados para o estado da Paraíba no Nordeste Brasileiro e concluíram que nos períodos mais secos as estimativas se correlacionam melhor com as observações, assim como Pereira et al. (2013), que observaram o mesmo comportamento para as regiões Centro-Oeste e Norte do país.

Apesar dos satélites de órbita polar possuir cobertura ainda maior que os geoestacionários, chegando a cobrir todo o globo terrestre, O TRMM não possuía uma boa resolução temporal, o que é um fator necessário para o monitoramento do tempo, neste sentido, em 2014, foi lançada a missão sucessora denominada *Global Precipitation Measurement* (GPM), constituída de um satélite principal denominado GPM-Core, com dois sensores chamados *Dual-frequency Precipitation Radar* (DPR) e *GPM Microwave Imager* (GMI), e uma constelação de outros satélites que fornecem seus dados de micro-ondas, infravermelho e demais sensores, para o cálculo da estimativa de precipitação. Esse arranjo possibilitou a extensão da área de cobertura para latitudes mais elevadas, além dos trópicos, e uma melhora significativa na resolução temporal, assim como, o aumento da precisão dos dados. Além da JAXA e da NASA, o programa GPM se tornou uma cooperação internacional incluindo outras agências espaciais como a *Indian Space Research Organization* (ISRO), a *European Organization for the Exploitation of Meteorological Satellites*

(EUMETSAT), o *Centre National d'Études Spatiales* (CNES), a *National Oceanic and Atmospheric Administration* (NOAA), e outros.

Da mesma forma que o antecessor PR/TRMM o sensor DPR a bordo do GPM-Core foi projetado para fornecer mapas tridimensionais da estrutura das chuvas e informações como tamanho e diâmetro médio das gotas, densidade e número de partículas, além disso, também foi melhorado para fornecer informações mais precisas ao combinar as bandas Ku e Ka, e ficou mais sensível a chuvas de baixa intensidade e queda de neve, além disso se mostrou eficaz para estimar chuvas com até 19 km de altitude. Já o GMI também baseado no antecessor TMI/TRMM foi acrescido de mais oito canais totalizando 13 canais micro-ondas entre as frequências 10 GHz até 183 GHz.

Correspondente ao TMPA a missão GPM disponibilizou uma série de algoritmos e um acervo de dados chamado *Integrated Multi-Satellite Retrievals for GPM* (IMERG) que durante os primeiros anos de operação apresentou sobreposição com os dados do TMPA o que pode ser utilizado para validar e comparar as estimativas de ambas às missões.

Wu et al. (2018) compararam os dados sobrepostos de ambas as missões e concluíram como era esperado, que devido ao aperfeiçoamento dos sensores e do algoritmo, os dados do IMERG mostraram-se melhor que o antecessor no período sobreposto, apesar disso, e do satélite TRMM ter sido descomissionado em 2014, quando seu combustível foi totalmente esgotado, o TMPA continuou a ser produzido utilizando os dados do programa GPM, garantindo assim a continuidade do conjunto de dados iniciado em 1998 mantendo as mesmas características de resolução temporal e espacial. Apenas em 2019 a NASA anunciou que na sexta versão do IMERG, foi implementado adaptações no algoritmo que possibilitaram o reprocessamento da estimativa utilizando os satélites e sensores anteriores aos do projeto GPM, produzindo assim estimativas com resolução espacial e temporal melhores que o TMPA para o mesmo período (desde 1998), sendo este produzido com resolução espacial de 0.1° e temporal de 30 minutos, substituindo oficialmente o TMPA, que encerrou sua operação no mesmo ano.

### 3.1.3 MERGE

Conforme a Tabela 3.1, apresentada por Valisoff et al. (2007), todas as principais metodologias para se estimar a precipitação instantânea apresentam limitações que geram incertezas em suas medidas. Para minimizar essas incertezas, existem algumas técnicas de fusão ("*merge*" em inglês) das estimativas de precipitação com dados observados em superfície. A missão GPM, por exemplo, disponibiliza uma versão do IMERG denominada IMERG-Final, que é considerada a versão mais refinada do conjunto de dados porque tem suas estimativas calibradas com valores observados mensais, porém sua disponibilização ocorre em média até três meses após o dado observado mensal ter sido disponibilizada, o que inviabiliza a utilização do produto em monitoramento.

Tabela 3.1: Vantagens e desvantagens e escala espaço-temporal dos principais equipamentos de medição/estimativa da precipitação.

| Sensor | Vantagens | Desvantagens |
|---|---|---|
| Estações meteorológicas | Medição direta da precipitação | Distribuição espacial não uniforme. Latência na transferência de dados em tempo real. Problemas com a qualidade das medições. Hidrometeoros congelados. Efeitos do vento. Não calibrado (tipo basculante com alta taxa de chuva). |
| Satélites geoestacionários | Cobertura espacial contínua | Medida indireta da precipitação e dificuldade com nuvens sem precipitação. |
| Satélites de Órbita Polar (micro-ondas passivo) | Cobertura espacial contínua | Resolução espacial e temporal pobre, medida indireta da precipitação e dificuldade com nuvens sem formação de gelo. |

Fonte: Adaptado de Vasiloff et al. (2007).

Na América do Sul duas técnicas de fusão se destacaram, tornando-se produtos operacionais do INPE, são elas o *Combined Scheme* (CoSch) (Vila et al., 2009), técnica baseada na remoção do viés das estimativas diárias satelitais, onde para cada ponto geográfico com valor observado, são calculados os vieses aditivo e multiplicativo que são interpolados utilizando um algoritmo de ponderação pelo inverso da distância, e o MERGE, (Rozante et al., 2010), que combina os dados de precipitação observada em estações de superfície a dados de precipitação em ponto de grade, subtraindo da estimativa os valores, onde dentro dos limites do tamanho do pixel foram realizadas observações, subtraindo também os valores ao redor do ponto de grade em uma distância equivalente a duas vezes a resolução da grade. Os pontos removidos são substituídos pelos valores observados interpolados, cada ponto de grade removido é então recalculado através da interpolação com os valores vizinhos próximos. Inicialmente as técnicas foram propostas para o TMPA e atualmente utilizam o IMERG como dado de entrada.

Existem ainda outras inúmeras avaliações positivas da fusão de estimativas satelitais e dados observados pelo mundo, como Li et al. (2010), que combinou dados da missão TRMM a rede de pluviômetros da Austrália, observando aumento significativo da qualidade de sua estimativa. Almazrou (2011) utilizou modelos de regressão linear com base em pluviômetros para corrigir dados superestimados do satélite TRMM sobre a Arábia Saudita; Zhou et al. (2016) utilizou os dados e as coordenadas geográficas das estações de superfície da Bacia de Qaidam na China para desenvolver um modelo de regressão ponderada e corrigir os dados do satélite TRMM, reduzindo o erro das estimativas, entre outros.

Vários estudos de avaliação das estimativas via satélite concluem que os dados são eficientes, contudo estudos mais profundos em território Brasileiro, como Gadelha et al. (2019) ressaltam uma tendência das estimativas via satélite de subestimar as medidas, e apontam a zona costeira da região Nordeste como um dos grandes responsáveis pelo resultado negativo, devido aos processos precipitantes derivados de nuvens quentes predominantes na região, e que são de difícil detecção aos sensores passivos de micro-ondas e

infravermelho. Contudo, a região possui uma grande quantidade de estações de superfície o que contribui para que as técnicas de fusão de dados de satélites com estações de superfície como o MERGE produzam uma estimativa mais eficiente e por essa razão o MERGE foi utilizado neste trabalho como referência para a precipitação diária.

## 3.2 Ciclo diurno de precipitação

O ciclo diurno de precipitação (CDP) é caracterizado por determinar os horários preferenciais de ocorrência e ausência de chuvas em uma escala diária (24 horas). Segundo Kikuchi e Wang (2008) na região tropical os CDPs podem ser agrupados em quatro regimes, continental, oceânico, costeiro continental e costeiro oceânico, como segue:

a) **continental:** Amplitude alta e máximo no final da tarde;

b) **oceânico:** Amplitude moderada e máximo no início da manhã;

c) **costeiro continental:** Amplitude muito alta com propagação de fase continente adentro;

d) **costeiro oceânico:** Amplitude relativamente alta e possibilidade de propagação de fase em direção ao oceano.

Definir adequadamente o CDP de uma região permite entender e estudar os processos físicos e dinâmicos que atuam nela. Baseando-se nesta hipótese vários estudos buscam relacionar o CDP a fenômenos de escalas global e regional. Oki e Musiake (1994) e Dai et al. (1999) relacionaram a máxima da precipitação à desestabilização da camada limite a tarde causada pelo aquecimento da superfície durante o dia, porém outros estudos argumentam que existem áreas onde os máximos de precipitação ocorrem durante a madrugada e podem ser consequência de fenômenos locais como a circulação orográfica, brisa marítima, ciclo de vida noturno de sistemas convectivos de mesoescala, entre outros. Dessa forma o CDP passa a ser uma variável importante para o desempenho de modelos de previsão do tempo e clima. Yang e Smith (2008) revisaram os quatro regimes moduladores e atestaram que esse comportamento é válido quando observado em grande escala, contudo, em escalas menores, é comum a ocorrência de sinergismo entre diferentes mecanismos, principalmente sobre a América do Sul, podendo alterar significativamente o CDP em uma determinada região, como por exemplo, no interior da bacia Amazônica, que de acordo com Kikuchi e Wang (2008) é descrita com o regime continental, mas frequentemente registra chuvas noturnas que Cohen et al. (1995) associam a mecanismos de brisa

marítima, fontes remotas de calor e jatos em baixos níveis. Santos e Silva et al. (2011) apresentam não quatro, mas seis regimes de ciclos diários e semi-diários para o continente Sul-Americano e outros cinco para o oceano.

Diversos trabalhos foram desenvolvidos para o estudo do CDP usando dados de estimativas de precipitação por satélite sobre as diversas regiões do Brasil (BRITO; OYAMA 2014); (SANTOS; SILVA et al. 2011), (OLIVEIRA et al. 2016) e (SANTOS; SILVA 2013), mas eles estão limitados a áreas delimitadas, a certos períodos de tempo, e as limitações dos satélites que não permitem uma generalização dos resultados para todo território brasileiro. Deste modo este trabalho vem para melhor definir essas características nas mais diversas regiões, a partir de um novo conjunto de dados, que pode ser utilizado para obter o CDP não regionalmente, mas localmente, e assim poder avaliar anomalias em regimes de ciclos diurnos em pontos específicos, como regiões de reservatórios e bacias hidrográficas, e também auxiliar os meteorologistas e hidrólogos a determinarem os efeitos inerentes do CDP e suas anomalias nos sistemas meteorológicos atuantes sobre o território brasileiro.

## 3.3 Downscaling da precipitação

A principal dificuldade para se estudar o CDP é justamente a falta de informações pluviométricas em escalas temporais altas (e.g. até 3 horas) e com alta representatividade espacial (dezenas de quilômetros) em uma série longa de dados. Atualmente, devido à evolução tecnológica ocorrida principalmente nos últimos anos, os equipamentos (satélites, radares e estações meteorológicas) conseguem satisfazer esses requisitos em boas condições para algumas regiões do Brasil, mas não para todas. Conforme retrocedemos a linha do tempo, as escalas espaciais e temporais também são reduzidas, o que torna difícil a realização de estudos sobre a variação espaço-temporal de uma situação passada, e uma solução seria estimar os valores a partir de processos de interpolação ou *downscaling*.

Para a geração da série de dados proposta nesta dissertação, será utilizado o *downscaling* estatístico temporal, que segundo Kumar et al. (2012) é o método mais comumente aplicado para aumentar a frequência das amostras consistindo em dividir cada passo de tempo em frações de tamanho aleatório na frequência da amostra desejada. Os primeiros modelos probabilísticos de ocorrência de precipitação utilizavam a série histórica para simular estatisticamente os dias com e sem chuva e a persistência de cada um e através de simulações de Monte-Carlo estimar a quantidade de chuva. Gabriel e Neumann (1962) foram um dos primeiros a propor e aplicar a cadeia de Markov de primeira ordem em um modelo estatístico, que se tornou referência para vários outros estudos futuros devido a sua praticidade e eficiência. Em meados de 1980 impulsionado pelo crescente uso de computadores por universidades e agências de pesquisas os modelos hidrológicos foram aprimorados e motivados a produzir dados de entrada para os locais e épocas onde não havia disponibilidade de série histórica, vários estudos sobre modelos estatísticos surgiram inclusive simulando além da precipitação, outras variáveis meteorológicas como Bruhn et al. (1980) que simulou além da precipitação, a temperatura máxima, temperatura mínima, humidade relativa mínima e radiação solar total para Geneva em Nova York e para Fort Collins no Colorado obtendo resultados precisos para ambos; Larsen e Pense (1982) simularam

precipitação, temperatura máxima, temperatura mínima e radiação solar total para outras seis localidades também obtendo um bom desempenho. Richardson (1981) propôs um modelo estocástico que utilizou para produzir quatro variáveis meteorológicas, precipitação, temperatura máxima, temperatura mínima e radiação solar, no qual ficou conhecido mais tarde como modelo *Weather Generation* (WGEN), o modelo foi aplicado inicialmente em cinco diferentes regiões que produziram séries temporais de 30 anos, em resolução temporal diária para cada uma delas, em seguida as médias mensais e anuais foram avaliadas e não apresentaram diferenças significativas, também foi avaliado o número médio de dias com chuva no mês e sua persistência comparados aos dados observados e os resultados também foram satisfatórios (RICHARDSON; WRIGHT, 1984). Apesar do WGEN ser um dos primeiros modelos propostos, atualmente ainda é reconhecido e muito utilizado, além dele vários outros geradores foram desenvolvidos ao longo dos anos, sendo outras das principais motivações, preencher lacunas de séries temporais meteorológicas e prover possíveis cenários do impacto de mudanças climáticas, dentre eles destacam-se o CLIGEN (NICKS; HARP, 1980), WXGEN (WILLIAMS et al., 1985), LARS-WGEN (RACSKO et al., 1991) e LARS-WG (SEMENOV; BARROW, 1997). No Brasil destacam-se os modelos CLIMABR (OLIVEIRA et al., 2005) e PGECLIMA_R (VIRGENS FILHO et al., 2011).

Alternativamente ao modelo WGEN e suas variações Hewitson e Crane (1992) e Hewitson e Crane (1996) utilizaram uma Rede Neural Artificial (RNA) para calcular a precipitação local baseados em trabalhos anteriores que a correlacionavam a certos eventos em larga escala. Em 1998 Wilby e Wigley (1997) publicou um estudo comparando seis diferentes modelos de *downscaling*, incluindo dois modelos estocásticos de simulação, entre eles o WGEN, dois métodos baseados em vorticidade e duas variações de RNAs. Naquela ocasião os modelos de simulação obtiveram resultados muito próximos aos observados, porém as RNAs obtiveram resultados insatisfatórios. Segundo os autores o resultado negativo das RNAs foi devido a uma limitação em simular a ocorrência de dias úmidos adequadamente. Contudo devido ao grande potencial das RNAs em resolver problemas complexos e não lineares,

novos estudos utilizando-as em *downscaling* apresentaram resultados melhores como Schoof e Pryor (2001) que utilizaram a forma mais simples de RNA chamada *MultiLayer Perceptron* (MLP) e obtiveram resultados similares aos métodos de *downscaling* baseados em regressão múltipla e Cannon e Whitfield (2002) que utilizando um modelo de *downscaling* de *ensemble* RNA perceberam que era capaz de estimar mudanças de fluxo apenas utilizando condições atmosféricas de grande escala. Kumar et al. (2012) efetuaram o *downscaling* temporal de diversas variáveis meteorológicas mensais do modelo de reanálises do NCEP para uma escala sub-diária, utilizando um modelo MLP, resultando em bias para a variável precipitação de 0,6. Ibarra-Berastegi et al. (2011) efetuaram o *downscaling* de vários conjuntos de dados de reanálises utilizando Floresta Aleatória (RF, *Random Forests* em inglês), Shi e Song (2015) também utilizaram RFs para o *downscaling* espacial de dados de precipitação da missão TRMM com base no índice de vegetação e outras seis variáveis geoespaciais.

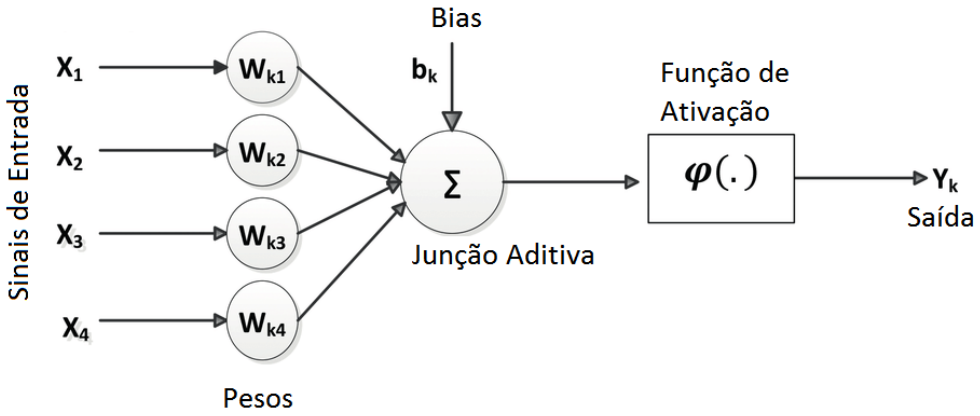
Para He et al. (2016), a popularização das técnicas de aprendizado de máquina (e.g. RNAs) é consequência da popularização de outra inovação tecnológica denominada *big data*. Na hidrologia as fontes de *big data* podem provir de instrumentos de medição, modelos atmosféricos de tempo e clima, satélites, radares, produtos de reanálise, entre outros. Apesar disso, He et al. (2016) argumenta que as técnicas de aprendizado de máquinas aplicadas no *downscaling* estatístico ainda foram pouco estudadas e tem muito a oferecer.

Neste trabalho as variações espaciais dos dados serão levadas em consideração pela própria natureza dos dados utilizados, ou seja, o uso de estimativa de chuva a partir de dados de satélites será aplicado em regiões onde a densidade de pluviômetros é baixa. Maiores detalhes serão dados nas seções posteriores.

## 3.4 Redes Neurais

É possível que uma pessoa sem nenhum conhecimento técnico em meteorologia consiga prever a chuva, apenas observando o ambiente ao seu redor, percebe-se por exemplo o céu escurecer, a temperatura cair, ouve-se alguns trovões e o cérebro automaticamente associa esses sinais a experiências vividas anteriormente e com certa acurácia momentos depois a precipitação ocorre. Computacionalmente é possível criar modelos que imitam o aprendizado humano através de associações e com isso resolver problemas complexos e difíceis de serem resolvidos através de algoritmos convencionais. Em 1956, John McCarthy idealizou o conceito de Inteligência Artificial (IA), que basicamente trata da capacidade de uma máquina raciocinar, aprender, reconhecer padrões e deduzir (MCCARTHY et al., 2006). Dentre as inúmeras técnicas de IA existe um modelo matemático/computacional baseado na estrutura neural de organismos inteligentes, em que as unidades de processamento, similares aos neurônios, são interconectadas formando uma Rede Neural Artificial (RNA) que por sua vez é capaz de adquirir conhecimento através de treinamento.

Figura 3.2: Modelo Neuronal de McCulloch e Pitts.




Fonte: Adaptado de Haykin (2001).

O modelo neuronal apresentado na Figura 3.2, corresponde ao modelo McCulloch e Pitts, referência aos idealizadores do modelo artificial em 1943 (MCCULLOCH; PITTS, 1943), que ilustra o *perceptron* de camada única, trata-se da representação de um único neurônio que possui a capacidade limitada de classificar apenas padrões separáveis linearmente em duas classes.

O sinal de entrada ($x_m$), quando conectado a um neurônio ($k_m$) é multiplicado pelo seu respectivo peso sináptico ($W_m$), criando uma sinapse. Conforme o exemplo inicial, as características observadas como queda na temperatura, som de trovões, entre outros, correspondem aos sinais de entrada, cada um desses sinais é ponderado conforme sua correlação com a saída, em seguida as sinapses são somadas, e da mesma forma que em um modelo de regressão linear, a partir do valor das variáveis correlacionadas é possível estimar o valor de saída. A função de ativação ($\varphi$) pode ser aplicada para restringir a saída do neurônio a uma amplitude finita, geralmente [0,1], que pode também ser interpretada como [falso,verdadeiro], amplamente utilizada em problemas de classificação. É possível representar esta etapa do processamento do neurônio, matematicamente através da Equação 3.1.

$$u_k = \sum_{,}^{j} k_m x_m \tag{3.1}$$

Onde:

$j$: total de sinais de entrada

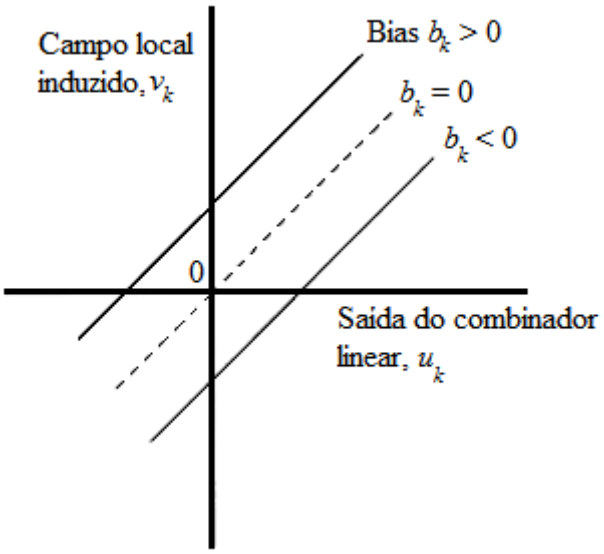
$x_m$ : sinal de entrada;

$k_m$ : peso sináptico do neurônio $k$;

$u_k$ : saída da junção aditiva

O *bias* ($b_k$) é utilizado para induzir uma posição diferente para a saída da junção aditiva, conforme Figura 3.3, alterando o campo local induzido ($v_k$) e a saída da junção aditiva ($u_k$), possibilitando o ajuste dos valores.

Figura 3.3: Ajuste da função de ativação através do bias.




Fonte: Haykin (2001).

Matematicamente o processo é representado pela Equação 3.2:

$$v_k = u_k + b_k \tag{3.2}$$

Onde:

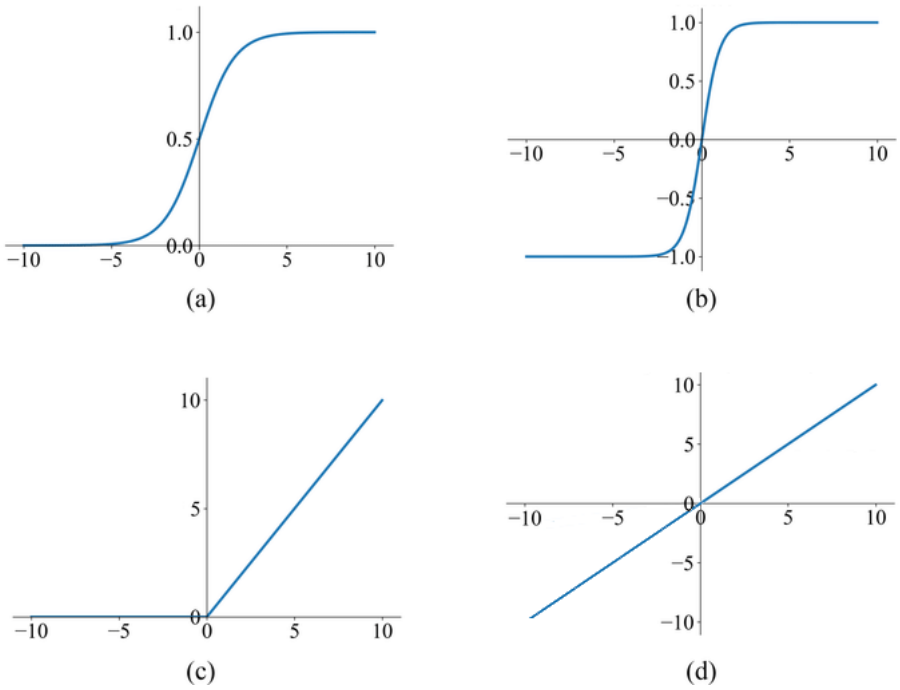
$u_k$: saída do combinador linear;

$b_k$: *bias*;

$v_k$: campo local induzido;

As funções de ativação são utilizadas geralmente para restringir a saída do neurônio a uma amplitude finita, por exemplo, a Figura 3.4(a), exibe a função *Sigmoid* que é utilizada para normalizar a saída do neurônio em valores entre 0 e 1, e a Figura 3.4(b) mostra a função *Tanh* utilizada para normalizar os dados para valores entre -1 e 1. Essas duas funções são comumente utilizadas em problemas de classificação. Já a Figura 3.4(c) mostra a função *ReLU*, que converte as saídas do neurônio para medidas maiores ou iguais a zero e a Figura 3.4(d) mostra a função de ativação linear que não altera a saída do neurônio e são comumente utilizadas em problemas de regressão.

Figura 3.4: Gráfico das principais funções de ativação.

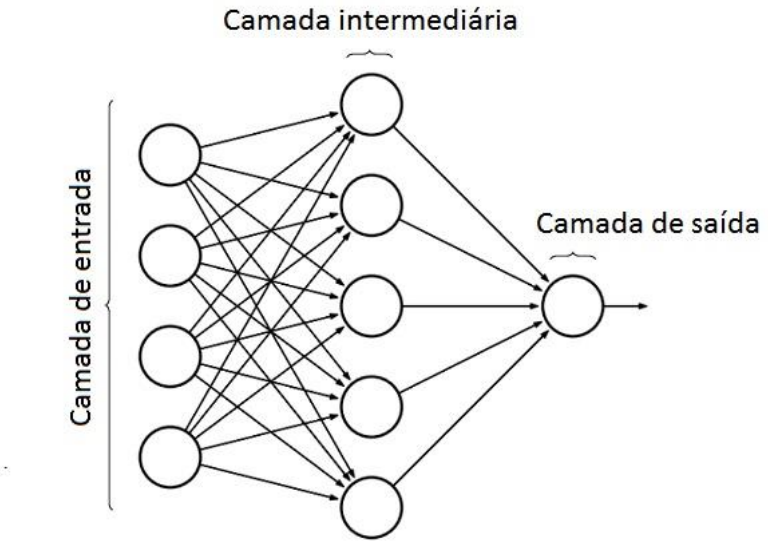



Fonte: Produção do Autor.

O *perceptron* de camada única possui capacidade de resolver apenas problemas lineares. Quando dois ou mais neurônios são combinados uma rede de neurônios artificial é criada chamada *perceptron* multicamadas (como definido a MLP do inglês, *Multi Layer Perceptron*). As RNAs possuem três tipos

de camadas, conforme a Figura 3.5, a camada de entrada é composta pelos neurônios que recebem sinais ou os dados que serão analisados, a de saída, que apresenta os resultados da RNA e uma camada intermediária, onde os neurônios nela pertencentes podem estar organizados em uma ou várias camadas internas, também chamadas de camadas ocultas, que fazem o processamento e extração de padrões associados ao sistema inferido. Cada uma das camadas pode conter um ou vários neurônios.

Figura 3.5: Exemplo de distribuição das camadas em uma RNA do tipo MLP.



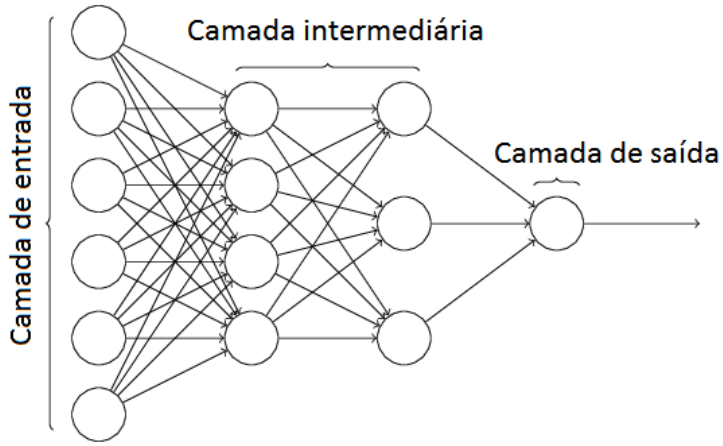

Fonte: Adaptado de Haykin (2001).

Este acréscimo no número de camadas permite resolver problemas e identificar padrões não lineares, o que já possibilitou resolver problemas complexos como reconhecimento de voz, classificação de imagens, previsão do tempo e clima, previsão da evolução da bolsa de valores, entre outros. Nas MLPs o fluxo de processamento se propaga em uma única direção, da camada de entrada para a camada de saída, ou seja, as saídas de uma camada são inputs para os neurônios da camada seguinte, e essa topologia recebe o nome de *feedfoward*.

As RNAs que possuem ciclos, ou seja, possuem conexões entre os neurônios em uma mesma camada ou com neurônios de camadas anteriores são chamadas Redes Neurais Recorrentes (RNN, em inglês *Recurrent Neural Network*), estes ciclos dão a elas a capacidade de armazenar informações de estados e reutiliza-los como novas entradas, permitindo explorar a complexidade de problemas que envolvem variáveis autocorrelacionadas, como em processamento de séries temporais onde o estado anterior de uma variável tem correlação com o estado atual.

Outro arranjo de RNA vastamente utilizado atualmente são as Redes Neurais Profundas (DNN, em inglês *Deep Neural Network*) ou como é popularmente referenciada na literatura, *deep learning,* que consiste no aumento do número de camadas ocultas na camada intermediaria, conforme a Figura 3.6, o que dá a RNA uma capacidade muito maior de processamento.

Figura 3.6: Exemplo de distribuição das camadas em uma RNA do tipo DNN.




Fonte: Adaptado de Haykin (2001).

Parâmetros como o número de camadas ocultas, quantidade de neurônios em cada camada e as funções de ativação são determinantes no desempenho da RNA, porém as metodologias existentes atualmente são inconclusivas e geralmente está modelagem é realizada de forma empírica.

A chamada fase de aprendizado ou treinamento consiste no processo de calibração dos pesos aplicados em cada um dos sinais da RNA, podendo este ser categorizada em dois tipos:

a) **supervisionado:** Quando é apresentada à RNA a saída esperada, também denominado alvo, para cada entrada. Neste caso a RNA altera os pesos sistematicamente até reproduzir o valor ou os padrões do alvo com o menor erro possível;

b) **não supervisionado:** Nesta categoria, apenas o conjunto de entrada é entregue a RNA, que é responsável por buscar singularidades nos sinais de entrada;

No treinamento supervisionado o algoritmo de retropropagação do erro é um dos parâmetros que precisam ser definidos, dentre os mais amplamente utilizados está o *backpropagation*, que se baseia no método do gradiente descendente, e o *levenberg marquardt* que se baseia na aproximação do método de Newton, para aproximação dos pesos computados. No *backpropagation*, utilizado neste estudo, o treinamento supervisionado é executado em duas fases chamadas *forward* e *backward*, onde na fase *forward* os sinais são apresentados aos neurônios da camada de entrada, cada neurônio calcula seus valores de saída aplicando os pesos, e emitem os sinais para a camada posterior até alcançar a camada de saída onde o resultado produzido é comparado com os valores desejados (alvos) e avaliados, e então, o erro associado a cada neurônio da RNA é calculado e inicia-se a fase *backward* ajustando recursivamente os pesos de cada neurônio, em seguida as fases são refeitas até obter o menor valor de erro possível. O erro que os algoritmos de retropropagação buscam minimizar é calculado através da função custo, no qual podem ser aplicadas diversas métricas como, raiz do erro médio quadrático (RMSE, em inglês *Root Mean Squared Error*), erro médio quadrático (MSE, em inglês *Mean Squared Error*), erro médio absoluto (MAE, em inglês *Mean Absolute Error*), entre outros.

Com a descoberta e aperfeiçoamento dos algoritmos de retropropagação do erro, as RNAs tornaram-se mais populares e atualmente existem inúmeras

arquiteturas além da MLP, RNN e DNN, dentre as principais estão, por exemplo, as Redes Neurais Convulacionais (CNN, em inglês *Convolutional Neural Network*), criadas para o processamento e reconhecimento de padrões em imagens digitais, a *Long Short-Term Memory* (LSTM), uma variação da RNN que agrega conceitos de memória de curto prazo, entre inúmeras outras.

Especificamente dentro da área de pesquisa deste trabalho as RNAs têm sido aplicadas para estimar chuva através do treinamento com dados de satélites e radares meteorológicos. Bellerby et al. (2000) utilizaram quatro das cinco bandas espectrais do satélite geoestacionário GOES combinado com o sensor PR do satélite de órbita polar TRMM para treinamento de uma RNA do tipo MLP para estimar chuva com 30 minutos de resolução temporal e 0.12° de resolução espacial, obtendo correlação de aproximadamente 0,47. Tapiador et al. (2004) realizaram um trabalho similar acrescentando dados dos satélites Nimbus e METEOSAT e conseguiu aumentar a correlação para 0,6. Já Tohma e Igata (1994) estimaram chuva com eficácia baseados em imagens de satélite de sensoriamento remoto visível e infravermelho para a região costeira do Japão. Kuligowski e Barros (1998) desenvolveram um modelo de previsão de chuva a curto prazo (0-6 horas) utilizando uma RNA treinada com dados de direção do vento e dados históricos de precipitação. Além do uso de plataformas espaciais, dados de superfície também foram utilizados como nos trabalhos de Xiao e Chandrasekar (1997) e Chiang et al. (2007), que aplicaram dados de radares para treinamento de uma RNA e conseguiram obter desempenho mais preciso e estável do que os modelos de relação de potência ZR vastamente utilizados na literatura.

# 4 DADOS E METODOLOGIA

Este capítulo descreve os formatos, resolução espacial e temporal e a origem dos dados utilizados no desenvolvimento deste trabalho, além de detalhar as técnicas e métodos que serão utilizados.

## 4.1 Dados

### 4.1.1 Pluviômetros

Os dados de estações automáticas foram adquiridos do Instituto Nacional de Meteorologia (INMET) para o período de 2000 a 2020 para todo o território brasileiro. Os dados foram fornecidos em mídia do tipo CD contendo 610 arquivos no formato *eXceL Spreadsheet* (XLS), onde cada um possui armazenado dados de precipitação horária de uma estação de superfície específica, desde seu primeiro registro até o último registro limitado em 31 de dezembro de 2020.

Figura 4.1: Exemplo de arquivo XLS disponibilizado pelo INMET.

| MINISTÉRIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO-MAPA | | | | | | | |
|---|---|---|---|---|---|---|---|
| INSTITUTO NACIONAL DE METEOROLOGIA - INMET | | | | | | | |
| ESTAÇÃO METEOROLÓGICA AUTOMATICA DE BRASILIA/DF | | | | | | | |
| | | | | | | | |
| Alt. | 1159,54m | | | | | | |
| Lat. | 15°47'S | | | | | | |
| Lon. | 47°55'W | | | | | | |
| | | | | | | | |
| | PRECIPITAÇÃO (mm) | PRECIPITAÇÃO (mm) | PRECIPITAÇÃO (mm) | ... | PRECIPITAÇÃO (mm) | PRECIPITAÇÃO (mm) | PRECIPITAÇÃO (mm) |
| HORA UTC | 0000 | 0100 | 0200 | ... | 2100 | 2200 | 2300 |
| 07-mai-2000 | NULL | NULL | NULL | ... | 0,0 | 0,0 | 0,0 |
| 08-mai-2000 | 0,0 | 0,0 | 0,0 | ... | 0,0 | 0,0 | 0,0 |
| 09-mai-2000 | 0,0 | 0,0 | 0,0 | ... | 0,0 | 0,0 | 0,0 |
| 10-mai-2000 | 0,0 | 0,0 | 0,0 | ... | 0,0 | 0,0 | 0,0 |
| 11-mai-2000 | 0,0 | 0,0 | 0,0 | ... | 0,0 | 0,0 | 0,0 |
| 12-mai-2000 | 0,0 | 0,0 | 0,0 | ... | 0,0 | 0,0 | 0,0 |
| 13-mai-2000 | 0,0 | 0,0 | 0,0 | ... | 0,0 | 0,0 | 0,0 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 26-nov-2011 | 0,0 | 0,0 | 0,0 | ... | 0,4 | 0,4 | 0,4 |
| 27-nov-2011 | 4,4 | 14,2 | 6,0 | ... | 0,0 | 0,4 | 0,2 |
| 28-nov-2011 | 0,6 | 1,6 | 0,4 | ... | 0,0 | 1,6 | 0,8 |
| 29-nov-2011 | 0,0 | 0,0 | 0,0 | ... | 0,8 | 0,4 | 0,0 |
| 30-nov-2011 | 0,0 | 0,0 | 0,0 | ... | 0,2 | 0,0 | 0,0 |

Fonte: Produção do Autor.

Conforme a Figura 4.1 cada arquivo é composto de um cabeçalho contendo as coordenadas e altitude referente à localização geográfica da estação e as medidas de precipitação encontram-se no corpo do arquivo organizados em uma tabela com 25 colunas sendo a primeira delas correspondente ao dia em que a medida foi coletada, e as demais referentes ao horário de coleta, no caso a primeira correspondente a 00 UTC, a segunda 01 UTC e assim por diante até a última em 23 UTC.

Cada um dos 610 arquivos possui um número variado de linhas que somam ao todo aproximadamente dois milhões e quinhentos registros, cada uma dessas linhas equivale a uma série diária de medidas horárias, porém várias dessas sequências possuem valores nulos em um ou mais horários decorrentes, segundo o INMET, de falhas nos equipamentos ou na transmissão dos dados, além disso, a maioria das medidas, devido à ausência de chuva no dia ou em grande parte do dia, é composta do valor zero. Cabe ressaltar que não há nenhum controle de qualidade e/ou validação nesses dados por parte do INMET (INMET, 2020).

### 4.1.2 Satélites

#### 4.1.2.1 MERGE

A composição MERGE produzida operacionalmente pelo INPE, com base na estimativa de precipitação via satélite IMERG, combina dados de sensores infravermelhos, micro-ondas ativo e micro-ondas passivo, além de ser calibrada com dados de estações de superfície espalhadas por todo território brasileiro. Considerando que, em composições como essa, o principal motivo de utilizar diversas fontes de dados é minimizar o erro associado a cada uma delas, o MERGE é indicado neste estudo como a melhor estimativa diária para o monitoramento hidro meteorológico em grandes áreas territoriais.

Os dados disponibilizados pelo INPE têm resolução espacial de 0.1° e formato *GRIdded Binary II* (GRIB2). O produto é gerado operacionalmente nas resoluções diária a partir de junho de 2000 e em resolução horária a partir de janeiro de 2010 (ROZANTE et al., 2020).

#### 4.1.2.2 *Merged InfraRed Temperature brightness product* (MERGIR)

Com o surgimento dos satélites meteorológicos, tornou-se possível observar e aferir medidas de energia eletromagnéticas de extensas áreas de superfície terrestre e oceânica, desde então diversas técnicas de estimativa de precipitação foram desenvolvidas e aperfeiçoadas, dentre elas destacam-se aquelas que utilizam os canais infravermelhos termais, principalmente os localizados na janela atmosférica entre 10 a 12 μm, onde a radiação emitida pela Terra é fortemente atenuada por nuvens densas, sendo assim, o valor de radiação aferido para o ponto de grade onde há ocorrência dessas nuvens é correspondente à radiação emitida pelo topo da nuvem, e através de técnicas físico empíricas é possível obter o valor precipitante ou potencialmente precipitável na coordenada geográfica onde a medida foi realizada. A NOAA produz operacionalmente um conjunto de dados referente a esse comprimento de onda e em parceria com a NASA, disponibiliza uma composição global em formato NetCDF-4, com 4km de resolução espacial e 30 minutos de resolução temporal em um banco de dados iniciado em 1998, sendo estes advindos de satélites geoestacionários em operação na época do registro, o qual os satélites GOES-8/9/10/11/12/13/14/15/16 provém dados para o continente Americano, os continentes Europeu e Africano são cobertos pelos satélites METEOSAT-5/7/8/9/10, e os continentes Asiático e a Oceania pelos satélites GMS-5/MTSat-1R/2/Himawari-8. Segundo a NOAA para fazer a fusão dos diferentes satélites, os dados originais foram corrigidos pela dependência de ângulo zenital, calibrando os dados mais afastados em relação ao nadir e suavizando a transição entre as medidas de um satélite para o outro (JANOWIAK et al., 2017).

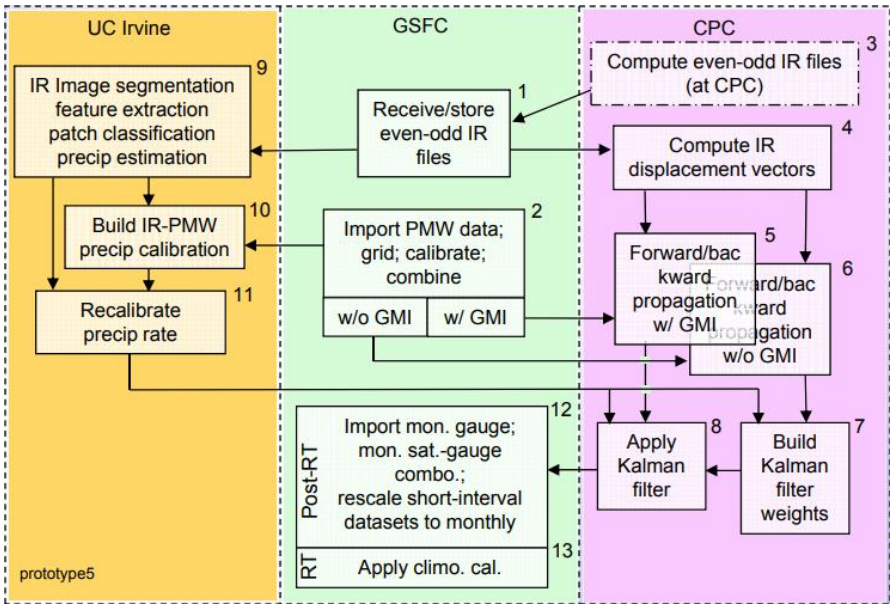
#### 4.1.2.3 *Integrated Multi-satellitE Retrievals for GPM* (IMERG)

Dentre as técnicas de estimativa de precipitação via satélite analisadas neste estudo, o algoritmo IMERG na versão 06B é o que atualmente produz a melhor estimativa sub-diária para toda a extensão do território brasileiro. O conjunto de dados é disponibilizado a partir de 1998. Segundo Huffman et al. (2020), o algoritmo interpola, combina e faz a inter-calibração entre todas as estimativas

de precipitação produzidas com dados de sensores infravermelho e micro-ondas dos diversos satélites que integram os programas GPM e TRMM, conforme a época e disponibilidade de cada um. Possui resolução espacial de 0.1° e resolução temporal de 30 minutos. É produzido em três versões denominadas *Early*, *Late* e *Final*. A versão *Early*, é equivalente ao produto em tempo real (*"NRT" Near Real Time* em inglês), ou seja, é a primeira estimativa gerada logo após a aquisição dos primeiros dados de satélites, ideal para produtos de monitoramento de tempo. Conforme a aquisição de novos dados o algoritmo é reprocessado, podendo ser executado inúmeras vezes após a primeira estimativa ter sido disponibilizada, gerando um produto mais refinado denominado como *Late*. O produto *Late* tem a latência aproximada entre 12 a 14 horas, e ainda segundo Huffman et al. (2020), a principal diferença entre essas duas versões é que devido à quantidade menor de informação para a produção da versão *Early* os dados são apenas extrapolados enquanto na versão *Late* os dados podem ser interpolados. Por fim a versão *Final* é calibrada com dados observados mensais e disponibilizada com aproximadamente 3.5 meses de latência, sendo essa a versão mais indicada para o desenvolvimento de produtos em nível de pesquisa, contudo, como o propósito deste trabalho tem a premissa do monitoramento hidro meteorológico, e o produto que possivelmente alimentará a RNA depois de finalizada será a versão *Early*, está será a versão utilizada neste estudo para treinamento da RNA. Espera-se que a RNA aplicada seja capaz de corrigir possíveis erros na estimativa da chuva, uma vez que o alvo são as chuvas observadas. Logo, este trabalho não só visa o *downscaling* da precipitação, mas também sua correção a partir da rede treinada.

O Processo detalhado, conforme diagrama de bloco apresentado na Figura 4.2, demonstra como o processo de geração do produto é realizado operacionalmente, envolvendo algoritmos de três diferentes instituições Americanas, *University of California - Irvine* (UC Irvine), *Goddard Space Flight Center* (GSFC) da NASA e *Climate Prediction Center* (CPC) da NOAA.

Figura 4.2: Diagrama de bloco da geração do produto IMERG.




Fonte: NASA (2019).

O CPC fornece os dados satelitais de radiância aferidos através de sensores infravermelhos com cobertura global para o GSFC, que os armazena e distribui. Esses dados são processados inicialmente pelo algoritmo de IA da UC-Irvine, *Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks* (PERSIANN) para produzir uma estimativa inicial. Paralelamente o GSFC recebe também dados de micro-ondas passivo (PMW) dos diversos satélites do programa GPM, e através do algoritmo *Goddard Profiling Algorithm* (GPROF) combina e calibram os dados com medidas de sensores micro-ondas ativo a bordo do satélite GPM-Core (DPR) ou do TRMM (PR) dependendo da época, e dados climatológicos. Os perfis calibrados são então utilizados pelo *Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks – Cloud* Classification *System* (PERSIANN-CCS) para recalibrar a estimativa inicial. O resultado juntamente com campos verticais integrados de vapor d'agua dos modelos de reanálise, *Modern-Era Retrospective analysis for Research and Applications* (MERRA) e *Goddard Earth Observing System* (GEOS), mais os perfis produzidos são interpolados com o algoritmo *Climate Prediction Center Morphing-Kalman Filter*

(CMORPH-KF), produzindo a estimativa *Early* do IMERG (HUFFMAN, et al., 2019).

### 4.1.3 Reanálises

Segundo o *Center for Hydrometeorology and Remote Sensing at UCI* (CHRS) da Universidade da Califórnia, responsável pelo algoritmo de IA dentro do processamento do IMERG, o requisito mais importante para o sucesso da RNA na estimativa de precipitação é a escolha das variáveis de entrada. Scofield e Kuligowski (2003) ressaltam que as taxas de precipitação são melhores ajustadas quando combinadas a variáveis como campos de água precipitável, umidade relativa e orografia. A técnica do hidroestimador utiliza, além das variáveis supracitadas, os campos de direção, magnitude do vento e temperatura de brilho do topo das nuvens extraídas de sensores IR a bordo de satélites para estimar chuva. Baseado nisso, os campos de coluna de água total, umidade relativa do ar e a direção e magnitude do vento, serão utilizados neste trabalho e descritos nas seções seguintes. Para obtenção dos campos será utilizado os dados do modelo atmosférico global CY41R2 disponibilizados através do repositório ERA5 da ECMWF, em formato *GRIdded Binary* (GRIB), com resolução espacial de 0.28° e resolução temporal de 1 hora (HERSBACH et al., 2018).

#### 4.1.3.1 Coluna de água total

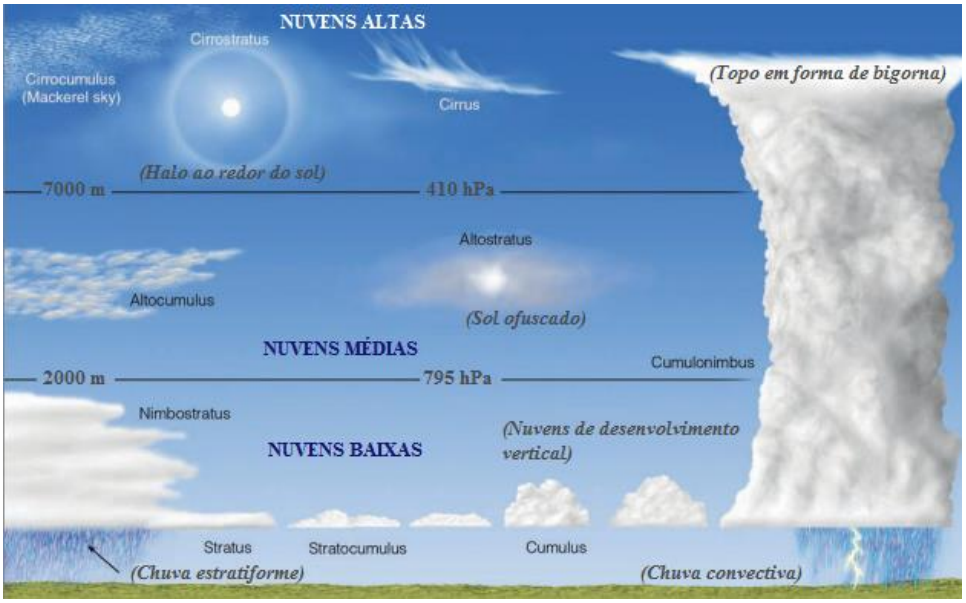
Segundo a descrição do ECMWF, a variável coluna de água total consiste da soma de vapor d'água, água líquida, gelo, chuva e neve em uma coluna vertical que se estende da superfície da Terra ao topo da atmosfera, ou seja, em termos gerais, representa a quantidade de água líquida em uma região sob a forma de precipitação acrescido da quantidade armazenada sob a região com potencial para precipitar-se.

#### 4.1.3.2 Umidade relativa do ar

A umidade relativa é medida em porcentagem e indica o quão próximo o ar atmosférico encontra-se da saturação de vapor d'agua, ou seja, é a razão entre a quantidade de vapor d'agua presente na atmosfera e a quantidade de vapor que a atmosfera suporta. Em condições de saturação o excesso de água se

condensa formando gotículas de água que podem vir a precipitar-se. Neste trabalho será utilizada o valor de umidade relativa presente nos níveis de pressão atmosféricos entre 500 e 1000 hPa, equivalente a umidade relativa localizada entre as altitudes aproximadas de 0 e 6000 m, onde conforme Figura 4.3, encontram-se as principais nuvens associadas a precipitação.

Figura 4.3: Linha de equivalência entre pressão e altitude.



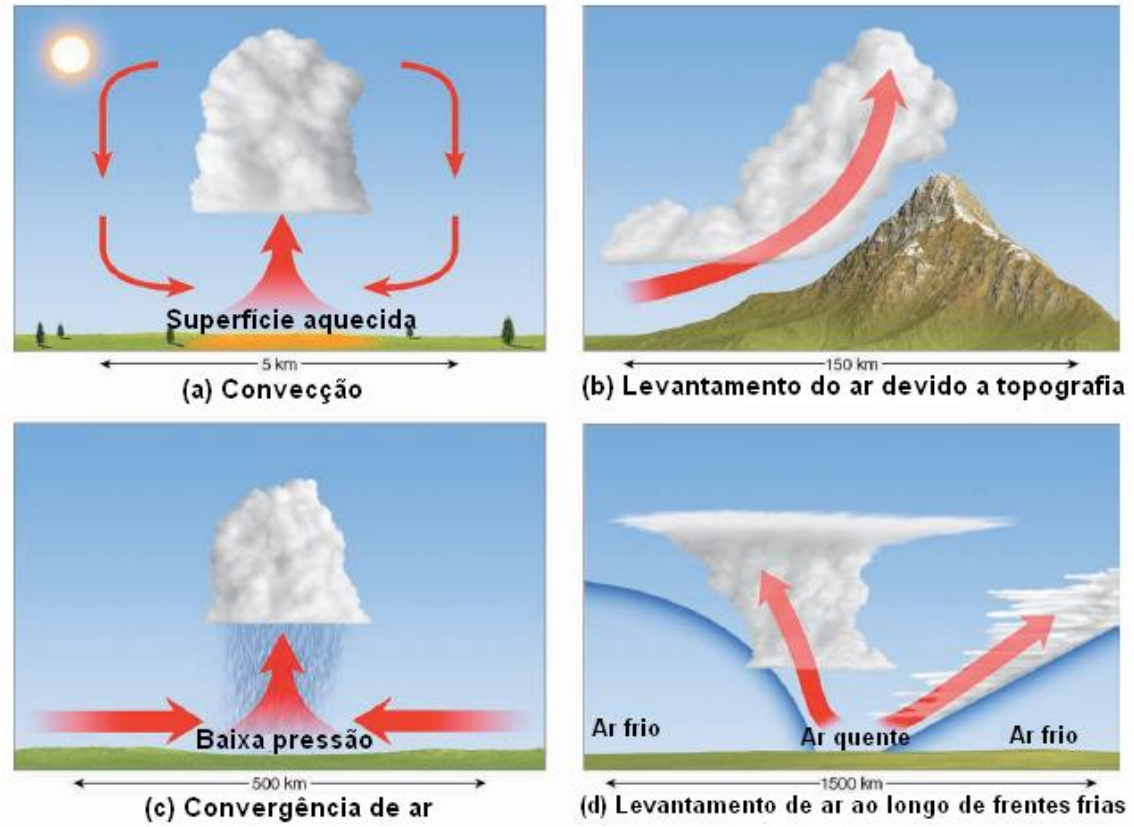

Fonte: Adaptado de Ahrens (2009).

Entre os limites de interesse, dezesseis níveis são disponibilizados no repositório ERA5: 500, 550, 600, 650, 700, 750, 775, 800, 825, 850, 875, 900, 925, 950, 975 e 1000 hPa.

#### 4.1.3.3 Direção e magnitude do vento

A direção e magnitude do vento tem grande influência na formação das nuvens conforme pode ser observado na Figura 4.4, além disso, também é uma componente que influência no erro dos vários equipamentos utilizados para aferir e estimar a chuva, como satélites, radares e pluviômetros. Para uso neste estudo será utilizada a variável em 850 hPa, das componentes U e V disponibilizadas pelo ECMWF.

Figura 4.4: Mecanismos de formação de nuvens.




Fonte: Ahrens (2009).

### 4.1.4 Geolocalização

Os dados de geolocalização são medidas praticamente constantes, alterados apenas por grandes desastres como terremotos e desmoronamentos, portanto, não são comuns medidas como essa ter uma resolução temporal muito grande, sendo assim a mesma grade de valores das variáveis desta seção será utilizada para toda a série temporal neste estudo.

#### 4.1.4.1 *Shuttle Radar Topography Mission* (SRTM) V2.1

A NASA juntamente com a instituição *National Imagery and Mapping Agency* (NIMA) organizaram no mês de fevereiro de 2000 a missão SRTM com o propósito de elaborar um Modelo Digital de Elevação (MDE) para todo o planeta através de radar de abertura sintética interferométrico (InSAR) instalado no ônibus espacial *Space Shuttle Endeavour*. A missão resultou em um MDE de 30 a 90 metros de resolução vertical. Segundo Ahrens (2009) em

terrenos montanhosos a circulação atmosférica interage com a topografia, sendo este fator determinístico para ocorrência e inibição da precipitação, esse processo pode ser visualizado na Figura 4.4(b) (JPL, 2021). Para utilização neste estudo a grade de 30 metros foi degradada para a mesma grade da estimativa diária MERGE (0.1°).

#### 4.1.4.2 Latitude e longitude

Espera-se que com a inserção da localização geográfica da precipitação na RNA, as diferentes condições climáticas características de algumas regiões do Brasil sejam observadas, como por exemplo, o clima seco da região Nordeste ou o clima úmido da região Amazônica, entre outros. Os valores utilizados serão o do ponto de grade do MERGE mais próximo da estação de superfície.

#### 4.1.4.3 Dia juliano

Assim como a inserção da localização geográfica na RNA, a inserção da variável em questão é justificada visando que as diferentes condições atmosféricas ocorridas em diferentes épocas do ano possam ser assimiladas.

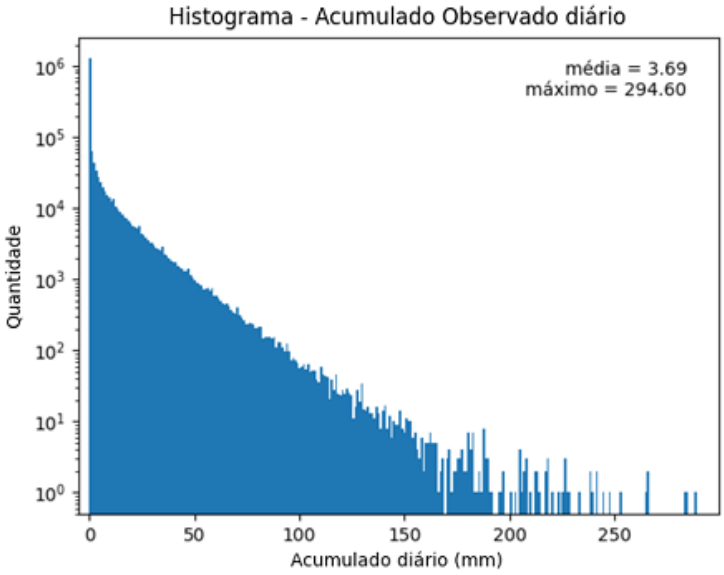
## 4.2 Metodologia

Nesta seção serão apresentadas as metodologias que foram utilizadas para produzir o *downscaling* da precipitação diária em horária a partir da RNA.

### 4.2.1 Pré-processamento dos dados

Para este estudo o acumulado diário foi calculado a partir da soma das 24 medidas horárias iniciando às 14 UTC do dia anterior ao de referência, até as 13 UTC do dia de referência. Na Figura 4.5 é apresentado o histograma desses acumulados de chuva observada pelos pluviômetros, onde é possível observar que os dias sem chuva representam a grande maioria das medidas, o que pode prejudicar o aprendizado da RNA.

Figura 4.5: Distribuição dos dados observados com os dias sem chuva.

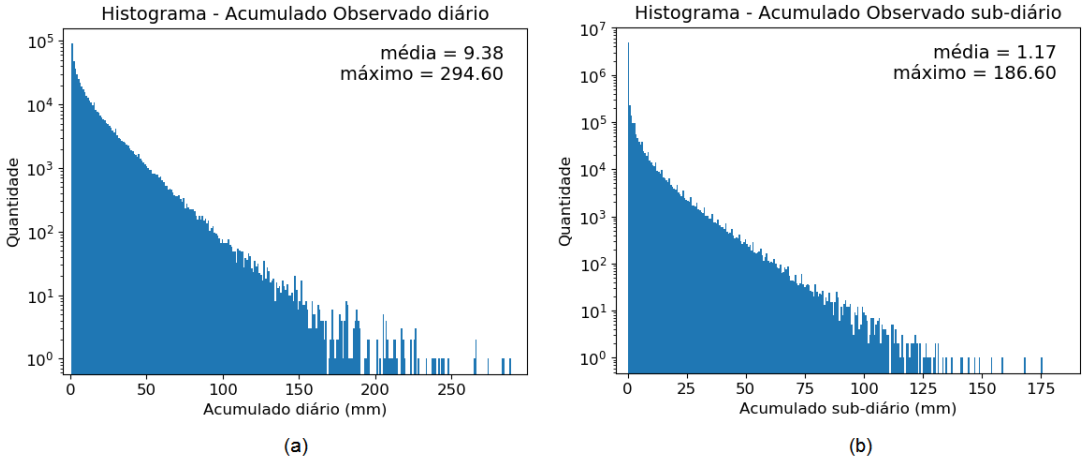



Fonte: Produção do Autor.

O conjunto de dados de estações de superfície automáticas, fornecido pelo INMET, não foi submetido a um controle de qualidade por parte da instituição, segundo eles, ruídos e falhas ocorridas durante a transmissão dos dados são representados, no conjunto de dados, com o valor nulo (não numérico). Realizar um acumulado ignorando os horários com valores nulos ou tentar estima-los produziria incertezas no processo de aprendizado da RNA, portanto, os dias que apresentam esses valores nulos em qualquer um dos horários foram excluídos, além disso, devido a grande maioria dos valores válidos no conjunto de dados ser composto de zeros ou valores próximos de zero, um cenário provável, resultante do treinamento com este arranjo, seria uma RNA que apresente apenas valores zeros como saída e mesmo assim tenha uma boa acurácia. Para evitar cenários como esses, o conjunto de dados foi balanceado excluindo os dias sem chuva, ainda assim, conforme a Figura 4.6(b) referente ao histograma dos acumulados sub-diários já com os dias sem chuva excluídos, é possível observar que grande parte do conjunto de dados horários ainda é composta de zeros, devido aos dias em que houve chuva em

apenas um, ou alguns horários do dia, o que é importante para que a RNA também aprenda a corrigir as situações onde a estimativa via satélite apresentou chuvas que não ocorreram realmente de acordo com as estações de superfície. Já a Figura 4.6(a) referente ao histograma dos acumulados diários, excluído os dias sem chuva, é possível perceber que com a exclusão a média dos valores subiu de 3,69 para 9,38, ainda assim uma maioria de valores próximos à zero permaneceu no conjunto de dados, preservando as características da natureza do dado.

Figura 4.6: Distribuição dos dados observados excluindo os dias sem chuva.




Distribuição do acumulado de precipitação dos dados de estações de superfície excluindo os dias sem chuva, onde (a) é o acumulado diário e (b) o acumulado sub-diário a cada 3 horas.

Fonte: Produção do Autor.

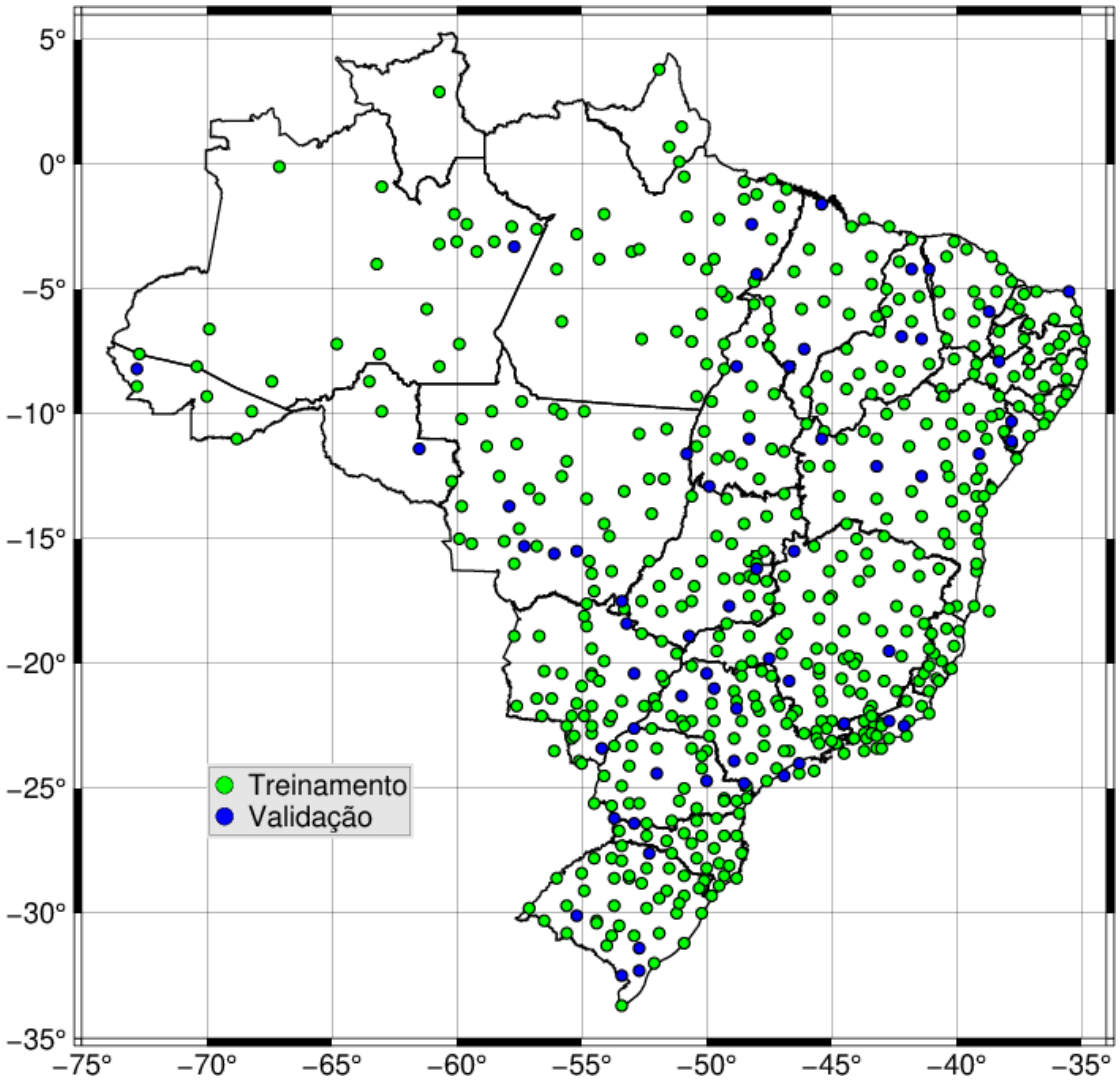
Outras análises mais robustas acerca de balanceamento de dados, que infelizmente não estão no escopo deste trabalho, mas que foram sugeridas em trabalhos futuros, poderiam ser realizadas, contudo, este desbalanceamento é um problema da natureza do dado, e muitas alterações poderiam descaracterizá-lo. Com a exclusão dos valores nulos e dos dias sem chuva restaram aproximadamente 814 mil séries temporais diárias. Espera-se que com isso a tendência a subestimativa diminua e também o custo

computacional. Essas medidas analisadas estatisticamente estão dentro dos limites considerados normais para precipitação no Brasil, e possuem quantidade suficiente para realização de treinamento supervisionado, orientando a RNA em quais valores espera-se alcançar ao executá-la.

As medidas sub-diárias neste estudo são os valores acumulados a cada 3 horas, sendo calculadas através da soma do valor medido no horário de referência acrescido do valor anterior e do valor seguinte em ±1,5 horas. Estes são os dados que melhor representam a chuva por serem medidas diretas da precipitação. O conjunto de dados observados diário, organizado inicialmente com valores horários, foi reorganizado em um conjunto de oito valores sub-diários que foram divididos e utilizados em três situações, o último ano da série, no caso 2020, foi reservado para validação da RNA em ocorrências fora do período utilizado para o treinamento, o restante equivalente a 20 anos de 2000 a 2019 foi dividido, parte para alvo do treinamento e teste da RNA (~90%) e parte para validação das informações geradas (~10%), respeitando a espacialidade entre as estações. Sendo este o conjunto alvo, a RNA desenvolvida neste estudo possuirá oito neurônios na camada de saída, um para cada estimativa sub-diária. Os 10% não utilizados no treinamento servem como referência para os locais onde não existem medidas em superfície. Uma vez que os dados de satélite permitem observar todo o território brasileiro, correções precisam ser realizadas nos locais onde não existem essas informações observadas. Logo, espera-se avaliar o comportamento destas estimativas a partir da RNA sobre estes locais utilizando estações em pontos geográficos que não foram utilizadas no treinamento e atestar a generalização da RNA.

Como dito anteriormente, a separação de parte do conjunto de dados observacionais para validação quanto à espacialidade foi realizada por região, separando aleatoriamente 10% da quantidade de estações de superfície de cada estado brasileiro individualmente, para serem independentes do treinamento. A Figura 4.7 exibe a divisão da distribuição das estações. Note que algumas regiões apresentam maior densidade do que outras, o que reflete diretamente nos resultados regionais como será discutido nos resultados.

Figura 4.7: Localização espacial das estações pluviométricas utilizadas.




Estações utilizadas no treinamento da RNA em verde e na validação em azul, no período de 2000 a 2019.

Fonte: Produção do Autor.

Quanto aos registros das estações de superfície os valores foram organizados em séries diárias de oito valores, acumulados a cada 3 horas e os dias sem chuva e com valores nulos foram excluídos, totalizando 691.753 séries diárias no conjunto utilizado para treinamento e teste da RNA (90,6%) e 71.474 no conjunto de validação (9,4%) para o período de 2000 a 2019. O conjunto de dados referente ao ano de 2020, reservado para validação temporal, totalizou outros 50.576 registros diários.

Além da precipitação, informações sobre as nuvens a qual estas estão relacionadas, foram utilizadas através de medidas da temperatura de brilho no espectro do infravermelho, provindas do satélite GOES a partir do conjunto de dados do MERGIR. Como medidas instantâneas podem ter baixa representatividade devido à alta variabilidade de algumas nuvens precipitantes, outras variáveis estatísticas foram utilizadas para inserir dentro da RNA esta variação espaço-temporal. Neste sentido, foram calculadas a média aritmética e a variância da temperatura de brilho que deveriam fornecer a RNA informações sobre a quantidade e a variação da temperatura do topo das nuvens, indicando condições favoráveis ou não para a ocorrência da precipitação.

As outras variáveis meteorológicas escolhidas para auxiliar a RNA durante a ponderação dos valores, umidade relativa, direção e magnitude do vento e a coluna de água total, também tiveram a média aritmética calculada a cada 3 horas. Todas as métricas, com exceção do acumulado diário e as informações de geolocalização, foram utilizadas como input para a RNA em séries temporais de oito valores, referentes a um dia, iniciando às 15 UTC do dia anterior ao de referência e encerrando as 12 UTC do dia de referência, em conformidade com a organização dos dados observados.

Ao todo foram selecionadas doze variáveis conforme a Tabela 4.1, que foram submetidas a um processo de análise exploratória, ao qual será apresentado no capitulo seguinte, contudo algumas dessas variáveis como a direção e magnitude do vento podem estar mais associadas com o local em que a chuva precipita e não tanto com a intensidade da chuva, outras variáveis podem apresentar uma correlação assíncrona como a temperatura, portanto mesmo apresentando baixa correlação em relação a chuva observada, cada uma das variáveis foi submetida isoladamente ao treinamento da RNA, verificando se houve melhora no aprendizado, sendo descartada quando o resultado se manteve negativo em ambas análises.

Tabela 4.1: Variáveis de input selecionadas para treinamento da RNA.

| **Variável** | **Dataset** | **Fonte** | **Resolução Espacial** | **Colocação Espacial** | **Resolução Temporal** | **Colocação Temporal** |
|---|---|---|---|---|---|---|
| Longitude | MERGE | INPE | 0.1° | Vizinho mais próximo | Fixo | Fixo |
| Latitude | MERGE | INPE | 0.1° | Vizinho mais próximo | Fixo | Fixo |
| Altitude | SRTM | NASA | 0.1° | Vizinho mais próximo | Fixo | Fixo |
| Dia Juliano | -- | -- | Pontual | Referência | Fixo | Fixo |
| Precipitação diária (mm) | MERGE | INPE | 0.1° | Vizinho mais próximo | Diário | Diário |
| Média 10 anos de Precipitação (mm) | MERGE | INPE | 0.1° | Vizinho mais próximo | Mensal | Acumulado 3 horas |
| Precipitação Micro-onda (mm) | IMERG | NASA | 0.1° | Vizinho mais próximo | 30 minutos | Acumulado 3 horas |
| Temperatura de brilho no IR (K) | MERGIR | NASA | 4km | Vizinho mais próximo | 30 minutos | Média e Variância em 3 horas |
| Umidade relativa média entre 1000-500 mb (%) | ERA5 | ECMWF | 0.28° | Vizinho mais próximo | 1 hora | Média em 3 horas |
| Direção do vento (graus) | ERA5 | ECMWF | 0.28° | Vizinho mais próximo | 1 hora | Média em 3 horas |
| Magnitude do vento (m/s) | ERA5 | ECMWF | 0.28° | Vizinho mais próximo | 1 hora | Média em 3 horas |
| Coluna de água total (kg $m^{-2}$) | ERA5 | ECMWF | 0.28° | Vizinho mais próximo | 1 hora | Média em 3 horas |

Fonte: Produção do Autor.

### 4.2.2 Modelagem da RNA

Em processamentos utilizando séries temporais é comum à utilização de RNA do tipo recorrente, onde diferentemente das RNA *feedfoward*, na qual o fluxo vai sempre de uma camada para a posterior, a RNN possui conexões ponderadas dentro de uma mesma camada, ou com camadas anteriores, que permitem o armazenamento e/ou a releitura dos estados, um exemplo dessa aplicação é o algoritmo de tradução de texto, onde cada palavra não é traduzida diretamente, mas são utilizados ciclos que armazenam os estados das variáveis durante a leitura, e a tradução ocorre apenas no término, possibilitando à RNN uma abstração acerca do sentido das frases e tornando a tradução mais realista, da mesma forma, além da correlação síncrona entre cada uma das variáveis selecionadas, é possível que uma RNN consiga abstrair correlações entre as flutuações assíncronas das variáveis de *input* com o ciclo diário de precipitação, como por exemplo a queda na temperatura ocasionada em decorrência da chuva precipitada em um horário anterior, contudo a RNA do tipo DNN, apesar de não possuir ciclos, pode possuir neurônios suficientes para abstrair a mesmas características de forma menos complexa, sendo assim, foram realizados testes com ambas as metodologias evoluindo a que mostrou-se mais promissora.

A fim de explorar a correlação entre todas as variáveis disponíveis, a RNA proposta tem um alto grau de conectividade. Quanto à especificação da camada intermediária não há, até o momento, consenso na literatura quanto a métodos e metodologias eficazes para determinar o número ideal de neurônios e de camadas ocultas, sendo alguns desses métodos descritos na Tabela 4.2.

Tabela 4.2: Métodos científicos determinísticos para o número de neurônios na camada oculta.

| **Referência** | **Equação** |
|---|---|
| Li at al., 1995 | $Nh = \frac{\sqrt{1 + 8Ni} - 1}{2}$ |
| Tamura e Tateishi, 1997 | $Nh = Ni - 1$ |
| Xu e Chen, 2008 | Se $\frac{Nt}{Ni} > 30$ então $Nh = \frac{1}{2}.\frac{Nt}{Ni \log Nt}$<br>Se $\frac{Nt}{Ni} \leq 30$ então $Nh = \frac{Nt}{Ni}$ |
| Shibata e Ikeda, 2009 | $Nh = \sqrt{Ni.No}$ |
| Hunter at al., 2012 | $Nh = \log_2(Ni + 1) - No$ |
| Sheela e Deepa, 2013 | $Nh = \frac{(4Ni^2 + 3)}{Ni^2 - 8}$ |

Onde $Ni$ é o número de *inputs*, $No$ é o número de *outputs*, $Nh$ é o número de neurônios na camada oculta e $Nt$ é o número de pares de treino.

Fonte: Produção do Autor.

Vujičić et. al. avaliou os métodos apresentados na Tabela 4.2 aplicando-os a dois conjuntos de dados distintos quanto a quantidade de elementos, o que não é considerado em nenhuma das metodologias mencionadas, e comparou o erro associado a cada um, que diferentemente do esperando foram controversos, concluindo que tais métodos não generaliza para todo tipo de conjunto de dados e podem ser utilizados apenas como ponto de ignição para a modelagem da topologia, porém a melhor configuração é obtida através do treinamento, teste e avaliação dos resultados (*tuning*). Yotov et. al. ressalta que

muitas dessas fórmulas apresentaram sucesso episódico e que a experiência trouxe a compreensão de que é necessário uma abordagem individual para cada tipo de problema, porém é consenso que um maior número de camadas ocultas possibilita uma maior capacidade da RNA extrair as características do ambiente, ou seja, uma RNA com poucas camadas e poucos neurônios teria menor capacidade de encontrar todos os padrões para modelar a saída adequadamente ocasionando um *underfitting*, assim como, um número excessivo pode levar a RNA a reproduzir os erros e ruídos do conjunto de dados utilizados como alvo, falhando na generalização da RNA e causando um *overfiting*. Portanto, a camada intermediaria do modelo de RNA proposto neste estudo foi configurada inicialmente de forma estocástica e calibrada de acordo com os resultados obtidos em testes de execução. Os dados de entrada possuem características estatísticas muito distintas e que podem ser prejudiciais para a evolução da RNA, dessa forma, a fim de minimizar tais problemas, cada uma das variáveis passou inicialmente por um processo de normalização, convertendo os valores de cada série para valores entre 0 e 1, e da mesma forma, cada neurônio da camada intermediária aplicará a função de ativação *Sigmoid* em suas saídas, antes de prosseguir com os dados para a camada seguinte, mantendo os estados no intervalo de 0 a 1. Apenas a camada de saída deverá utilizar a função de ativação *Rectified Linear Unit* (ReLU), convertendo o resultado final, em valores maiores ou iguais a zero equivalente aos milímetros de chuva do período.

É recomendado que os pesos dos neurônios de uma camada sejam inicializados com valores aleatórios e com baixa variância, para isso é preciso definir uma função de inicialização para cada camada da RNA. Dentre as funções mais populares e difundidas atualmente, estão às funções de inicialização *GlorotNormal* e *GlorotUniform*, proposta por Glorot e Bengio (2010), que utilizam o número de neurônios da camada de entrada e o número de neurônios da camada intermediária como variáveis das funções, porém segundo He et al. (2015) as funções de Glorot são falhas para camadas com a função de ativação ReLu, e para esses casos o autor propõe outras duas funções denominadas *HeGlorot* e *HeUniform*. Baseando-se nestes

comentários, para este estudo foi utilizado as funções *GlorotUniform* para as camadas de entrada e intermediária, onde foi definido anteriormente a *sigmoid* como função de ativação e na camada de saída, onde a função de ativação é a ReLu, foi utilizado a função *HeUniform*.

Para o treinamento, em vez de utilizar todo o conjunto de dados (*batch*), equivalente a 691.753 registros, para computar o gradiente em cada interação, será utilizado amostras menores do conjunto (*minibatchs*), com valores selecionadas aleatoriamente, isso permite otimizar o uso da memória. Os *minibatchs* terão 512 amostras cada, possibilitando o máximo de 1173 interações, durante cada interação um *minibatch* será dividido em duas parcelas, onde 80% serão utilizados para ajustar o valor dos pesos e 20% para avaliar o gradiente com os pesos ajustados. O treinamento será encerrado quando finalizada o número possível de interações ou após um total de 20 interações sem melhora no aprendizado. O algoritmo de otimização utilizado é o *adaptative moment estimation* (ADAM), que é uma evolução do gradiente descendente estocástico (SGD, *Stochastic Gradient Descent* em inglês), e que combina os conceitos de decaimento, momento e aprendizado adaptativo. A taxa de aprendizado será inicialmente definida como 0,001 e ajustada conforme os resultados das execuções. Como a maioria dos valores do alvo são zeros ou valores próximos de zero, é esperado que métricas como o erro médio (ME, *Mean Error* em inglês) ou viés (BIAS, em inglês), apresentado na Equação 4.1, e o erro médio absoluto (MAE, *Mean Absolute Error* em inglês), apresentado na Equação 4.2, também apresente valores muito baixos, dessa forma, se a chuva ocorrer, por exemplo, em apenas um dos oito horários do dia, e a RNA identificar que ela ocorreu em outro horário, o valor do erro será dividido por oito e não será representativo, causando uma falsa impressão de que a RNA é eficiente, neste sentido será utilizada, para avaliação de desempenho da RNA o erro médio quadrático (MSE, *Mean Square Error* em inglês), apresentado na Equação 4.3, onde os erros são elevados ao quadrado, antes do cálculo da média aritmética, ressaltando o valor do erro.

Apesar de não ter sido o índice utilizado para avaliação de desempenho durante o treinamento da RNA, o BIAS, será utilizado para analisar a tendência

do modelo em subestimar ou superestimar a precipitação, além dele o MAE e a raiz quadrada do erro médio quadrático (RMSE, *Root Mean Square Error* em inglês), Equação 4.4, serão analisados para avaliar o desempenho da RNA aplicada nos conjuntos de dados de validação e teste.

$$BIAS = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{y}_i \tag{4.1}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{4.2}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{4.3}$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{4.4}$$

Onde:

$y$ é o valor observado

$\hat{y}$ é o valor estimado

Para o desenvolvimento foi utilizado a linguagem Python, juntamente com as bibliotecas pandas, numpy, math entre outras para manipulação dos dados e operações matemáticas, as bibliotecas sklearn e tensorflow para construção do modelo neuronal e as bibliotecas seaborn, matplotlib e outras para a produção dos gráficos. Parte do código fonte desenvolvido pode ser encontrado em: https://colab.research.google.com/drive/1TnrfMerLBH2cPykp4aw6fOw7Ufo-PlGX?usp=sharing

# 5 RESULTADOS

Neste capítulo serão discutidos os resultados observados durante este estudo. As seções a seguir irão abordar questões técnicas e científicas como a comparação dos dados de entrada de precipitação do MERGE a serem feitos o *downscaling* com dados observados (Seção 5.1), uma análise exploratória dos dados de modo a definir aqueles que apresentam uma melhor correlação com a chuva (Seção 5.2), as análises que ajudaram a definir a RNA, objetivo deste estudo (Seção 5.3) e um experimento de caso de uso (Seção 5.4)

## 5.1 Intercomparação entre os estimadores (MERGE/IMERG) e as observações

Para verificar se os dados que serão usados para *downscaling* estão próximos dos dados observados foram feitas intercomparações. Na Figura 5.1 e na Tabela 5.1 é possível visualizar os números destas comparações para todo o período de análise utilizada no treinamento (2000 a 2019). Ao comparar a estimativa do MERGE (só os dados com informações do acumulado diário) com os acumulados diários das estações de superfície, foi possível observar correlação alta de 86% entre elas, Figura 5.1(a), no conjunto de dados separados para treinamento e um resultado parecido de 85%, Figura 5.1(c), no conjunto de dados separado para validação. Já os valores de MAE foram de 3 e 3,33 mm e o RMSE de 7,68 e 8,06 mm. O BIAS observado foi de -0,96 e -1,1 mm para o treinamento e validação, respectivamente, o que mostra uma subestimativa dos valores com relação à observação. Convencionalmente se analisa apenas o período de avaliação, contudo, quando o objetivo é criar um algoritmo baseado em redes neurais é importante também se conhecer o desvio existente durante o período de treinamento, neste sentido, ficou claro que os períodos são similares. Note que, além da alta correlação, as distribuições observadas pelos histogramas nas Figuras 5.1(b) e 5.1(d) são semelhantes e o gráfico de espalhamento, apresentado nas Figuras 5.1(a) e 5.1(c), mostram certa tendência linear, que no geral pode ser representada pela linha destacada em vermelho nas figuras (x=y). Contudo, existe uma

grande dispersão dos dados próximos à zero. As diferenças observadas podem estar associadas justamente aos efeitos das interpolações espaciais aplicadas no algoritmo e da má representatividade do modelo de estimativa de chuva usado (i.e. IMERG) como destacado por Huffman et al. (2020). A comparação entre os conjuntos de dados utilizados para treinamento e validação apresentaram singularidades apesar da diferença expressiva de medidas o que reforça o resultado do comparativo.

Figura 5.1: Distribuição da precipitação diária do MERGE e pluviômetros.

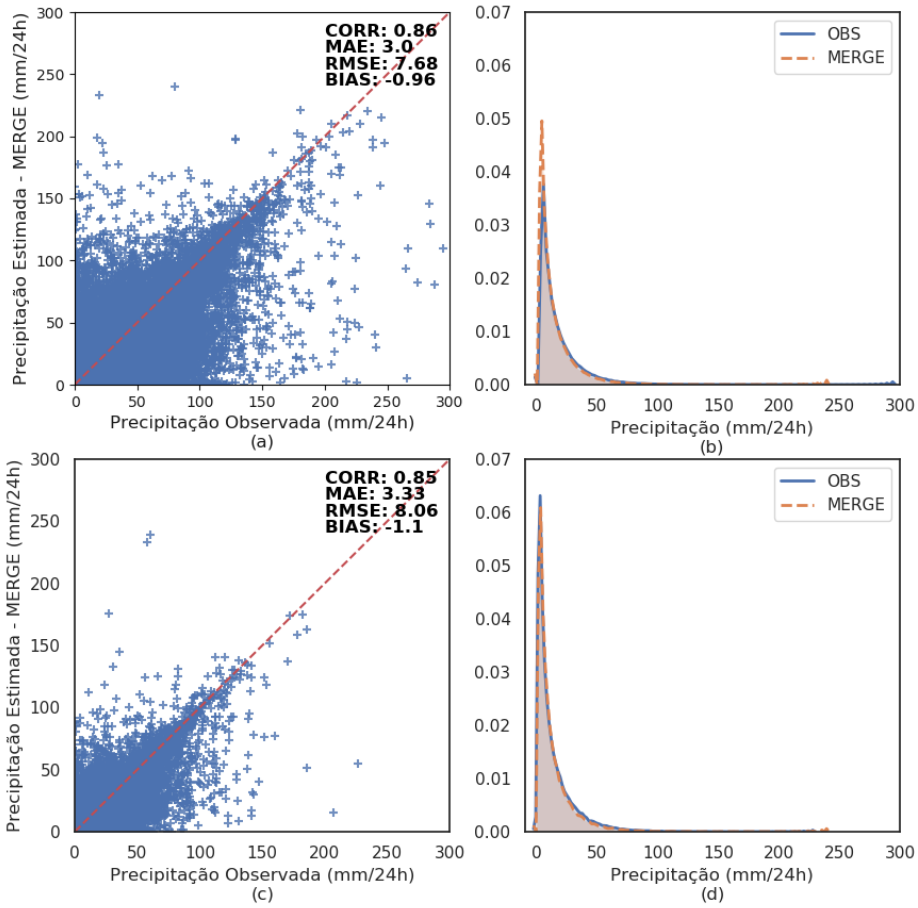



Gráfico de espalhamento das estimativas de precipitação MERGE diário em relação às medidas de pluviômetros acumuladas diariamente, onde (a) refere-se ao conjunto de dados de treinamento e (c) ao conjunto de validação, e o comparativo do histograma de cada um dos conjuntos de dados, onde (b) refere-se ao conjunto de dados de treinamento e (d) ao conjunto de validação.

Fonte: Produção do Autor.

A Tabela 5.1 mostra ainda que existem valores próximos para diversas métricas estatísticas como a média, desvio padrão, máximos e mínimos e os quartis. O que mostra que o dado é relevante para este estudo e pode ser utilizado como principal informação de acumulado diário a ser reduzido em escala temporal, aqui chamado apenas de *downscaling*. Como será notado em seções posteriores, a RNA aplicada foi capaz de realizar correções que melhoraram essas diferenças.

Tabela 5.1: Comparativo estatístico entre a estimativa diária do MERGE e IMERG com relação ao acumulado diário observado.

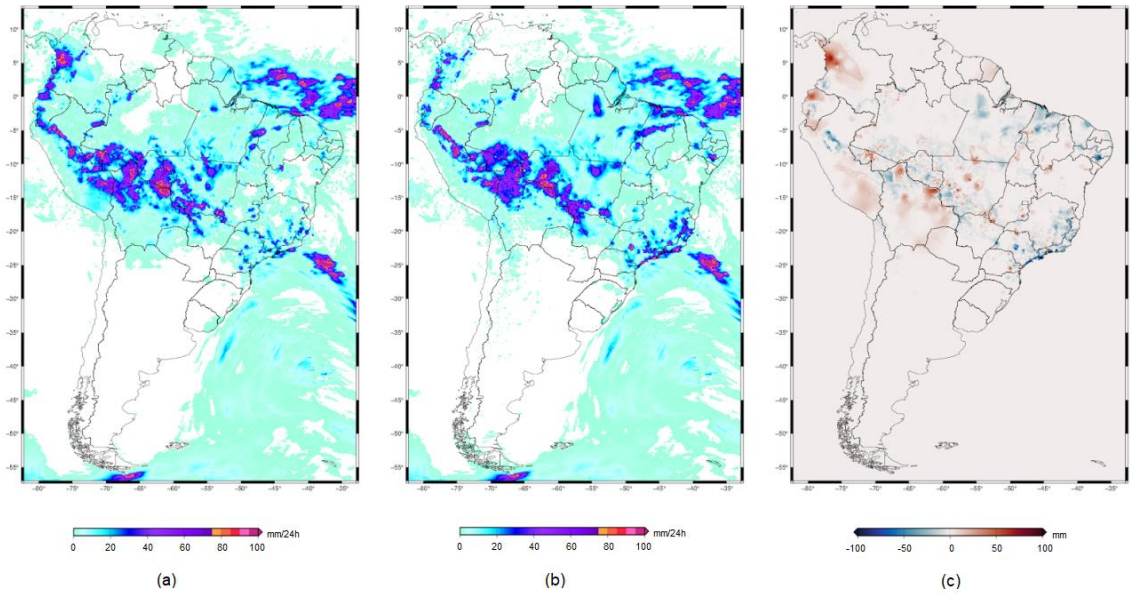
| **Período** | **Treinamento** | | | **Validação** | | |
|---|---|---|---|---|---|---|
| **Dado** | **OBS** | **MERGE** | **IMERG** | **OBS** | **MERGE** | **IMERG** |
| **População** | 667699 | 667699 | 667699 | 68897 | 68897 | 68897 |
| **Média** | 9,34 | 8,38 | 9,84 | 9,68 | 8,58 | 10,65 |
| **Desvio Padrão** | 14,90 | 13,29 | 17,37 | 14,91 | 13,23 | 17,41 |
| **Mínimo** | 0,20 | 0,00 | 0,00 | 0,20 | 0,00 | 0,00 |
| **1° Quartil** | 0,60 | 0,50 | 0,06 | 0,60 | 0,62 | 0,27 |
| **2° Quartil** | 3,20 | 3,00 | 2,89 | 3,40 | 3,25 | 3,94 |
| **3° Quartil** | 11,80 | 10,50 | 12,12 | 12,40 | 11,00 | 13,64 |
| **Máximo** | 294,60 | 239,88 | 405,20 | 226,80 | 238,75 | 313,24 |

Fonte: Produção do Autor.

O MERGE apresenta duas versões, uma usa apenas dados diários e a outra, dados horários, sendo este último apenas períodos mais recentes. No intuito de averiguar as diferenças entre elas, uma vez que este estudo visa apenas usar a versão diária, uma comparação entre ambos também foi realizada. Como pode ser observado na Figura 5.2, os campos de chuva para o dia 22 de fevereiro de 2020 e a diferença entre eles, é possível perceber tanto superestimativas como subestimativas na diferença do acumulado diário do

MERGE, gerado a partir da soma dos produtos sub-diários, e o produto MERGE diário. Estas diferenças são devido à utilização exclusiva das estações de superfície convencionais no cálculo diário, que devido a maior densidade de dados cobre áreas maiores, representando assim a estimativa diária da chuva com maior propriedade. Contudo em anos anteriores o número de estações de superfície, tanto as convencionais quanto as automáticas, que são esparsas atualmente eram ainda menores, além disso, a distribuição das redes pluviométricas em cada um dos estados Brasileiros é diferente, sendo algumas muito mais esparsas que em outros. Fatores esses que causam sub-representatividade nas estimativas mais antigas e em algumas regiões.

Figura 5.2: Diferenças do MERGE diário e MERGE sub-diário acumulado.




Produtos referentes ao dia 22 de fevereiro de 2020: MERGE diário à esquerda (a), acumulado 24 horas do MERGE sub-diário ao centro (b) e a diferença entre os dois produtos à direita (c).

Fonte: Produção do Autor.

Considerando que os dados diários observados são superiores em quantidade, principalmente em anos passados, quando comparados aos dados sub-diários, o *downscaling* desta variável será realizado, na expectativa de que as

diferenças entre essas estimativas e os dados observacionais sub-diários (estações de superfície) não sejam significativas. Contudo, as estimativas do MERGE em resolução temporal sub-diárias também serão utilizadas para produzir uma média decenal (2010-2019) do ciclo diurno para cada mês, maiores detalhes sobre o uso desta informação serão apresentados em seções posteriores.

Como o MERGE usa as estimativas do IMERG, as análises que seguem são justamente para quantificar o quão perto das observações são desses dados. A Figura 5.3 e a Tabela 5.1, apresentam os resultados da relação entre a estimativa diárias (acumulados dos dados horários) do IMERG e sua contrapartida observacional. A Figura 5.3(a) assim como a Figura 5.3(c), mostram alta dispersão entre a estimativa e a observação, enquanto a frequência relativa mostrou distribuições semelhantes. Este tipo de comportamento mostra uma má representação espacial e/ou temporal das chuvas, ou seja, a chuva ocorreu, mas não exatamente no local e horário esperado. A linha destacada em vermelho nas figuras é uma referência de valores iguais entre as séries, apenas para análise visual. O MAE observado para este conjunto de dados foi de 8,2 mm e o RMSE foi de 15,25 mm para os dados de treinamento, já com relação à validação, esses foram 8,46 e 15,32 mm, o que mostra uma distribuição semelhante entre ambas as séries. Observa-se que o erro deste algoritmo é muito maior que aquele observado para o MERGE, como é esperado, já que este último se aplica correções baseadas por dados na superfície, pluviômetros. O BIAS mostra um comportamento contrário ao do MERGE, onde existe uma superestimava dos valores de 0,51 e 0,97 mm para o treinamento e validação, respectivamente. No entanto, a correlação foi de 0,56 em ambos os conjuntos de dados, baixa se comparada ao MERGE, o que mostra também a importância dos dados observados em superfície na correção deste algoritmo, e corrobora com as análises acima sobre a má representatividade espacial. Percebe-se também pela Tabela 5.1 que os valores do IMERG são quase sempre superiores aqueles observados pelos pluviômetros e pelo MERGE. A exceção se dá nos percentis mais baixos, onde ele mostra que sua distribuição está muito próxima

a zero, contudo, as maiores percentis são superiores. Tal comportamento mostra que sua distribuição não deve estar em fase com as observações.

Figura 5.3: Comparativo entre o IMERG diário e o acumulado de precipitação observada no mesmo período.

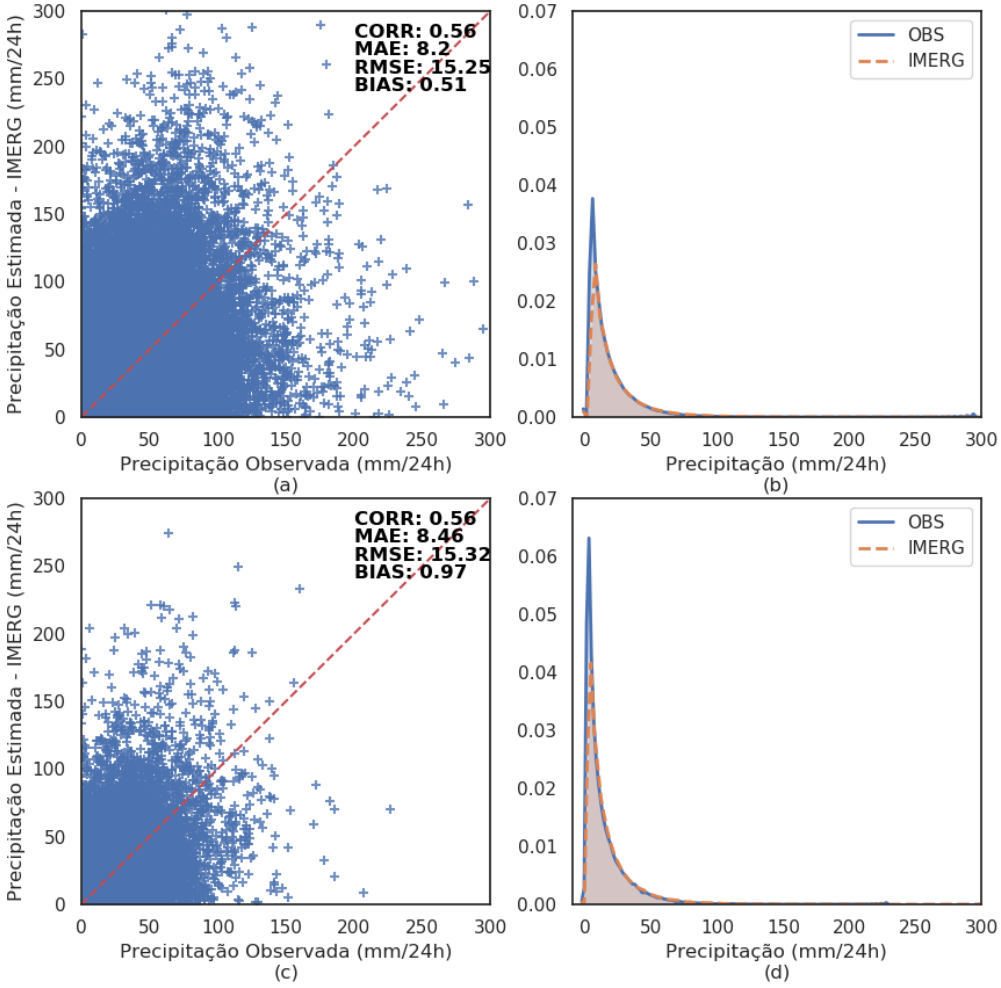



Gráfico de espalhamento das estimativas de precipitação IMERG diária em relação às medidas de pluviômetros também diários, onde (a) refere-se ao conjunto de dados de treinamento e (c) ao conjunto de validação, e o comparativo do histograma de cada um dos conjuntos de dados, onde (b) refere-se ao conjunto de dados de treinamento e (d) ao conjunto de validação.

Fonte: Produção do Autor.

## 5.2 Análise exploratória dos dados de entrada

Uma vez avaliada a performance dos algoritmos estimadores que servem de entrada primária para a RNA, nesta etapa será analisado de forma exploratória os outros dados que fazem parte deste estudo. Logo, a fim de definir a relação entre as variáveis selecionadas para o treinamento da RNA foi verificado a correlação síncrona entre cada uma delas. Este tipo de análise permite verificar se as variáveis (entrada) apresentam relação direta síncrona com a precipitação observada (alvo). Isto dará suporte na definição de quais variáveis seriam mais interessantes para o treinamento do sistema de aprendizado. Cabe ressaltar que a assincronicidade pode existir entre as variáveis, contudo esta não foi levada em consideração nesta análise e sugerida como trabalho futuro. Além disso, algumas relações que não mostram aparente representação física foram disponibilizadas no Apêndice A.

Uma das variáveis utilizadas para averiguar a distribuição espacial e temporal da precipitação via satélite é a chuva horária estimada pelo IMERG. A Figura 5.4 apresenta os resultados da relação entre as estimativas deste algoritmo e os mais de cinco milhões de registros sub-diários disponíveis para este estudo. A Figura 5.4(a) assim como a Figura 5.4(c), mostra alta dispersão entre a estimativa e a observação, enquanto a frequência relativa mostrou que os dados apresentam distribuições semelhantes. Este tipo de comportamento geralmente está associado a uma má representação espacial dos dados, ou seja, a precipitação ocorreu, mas em um lugar diferente. A linha destacada em vermelho nas figuras se refere ao modelo linear que melhor expressa à relação entre as duas grandezas. O MAE observado para este conjunto de dados foi de 1,39 mm e o RMSE foi de 4,32 mm no conjunto de dados de treinamento e valores próximos foram encontrados no conjunto utilizado para validação, o que caracteriza uma boa relação entre eles. O BIAS mostra uma superestimava dos valores de 0,06 e 0,12 mm para o treinamento e validação respectivamente. No entanto, a correlação foi de 0,47 em ambos os conjuntos de dados, baixa se comparada ao MERGE, como discutido anteriormente para o acumulado diário.

Figura 5.4: Comparativo entre o IMERG sub-diário e o acumulado de precipitação observada no mesmo período.

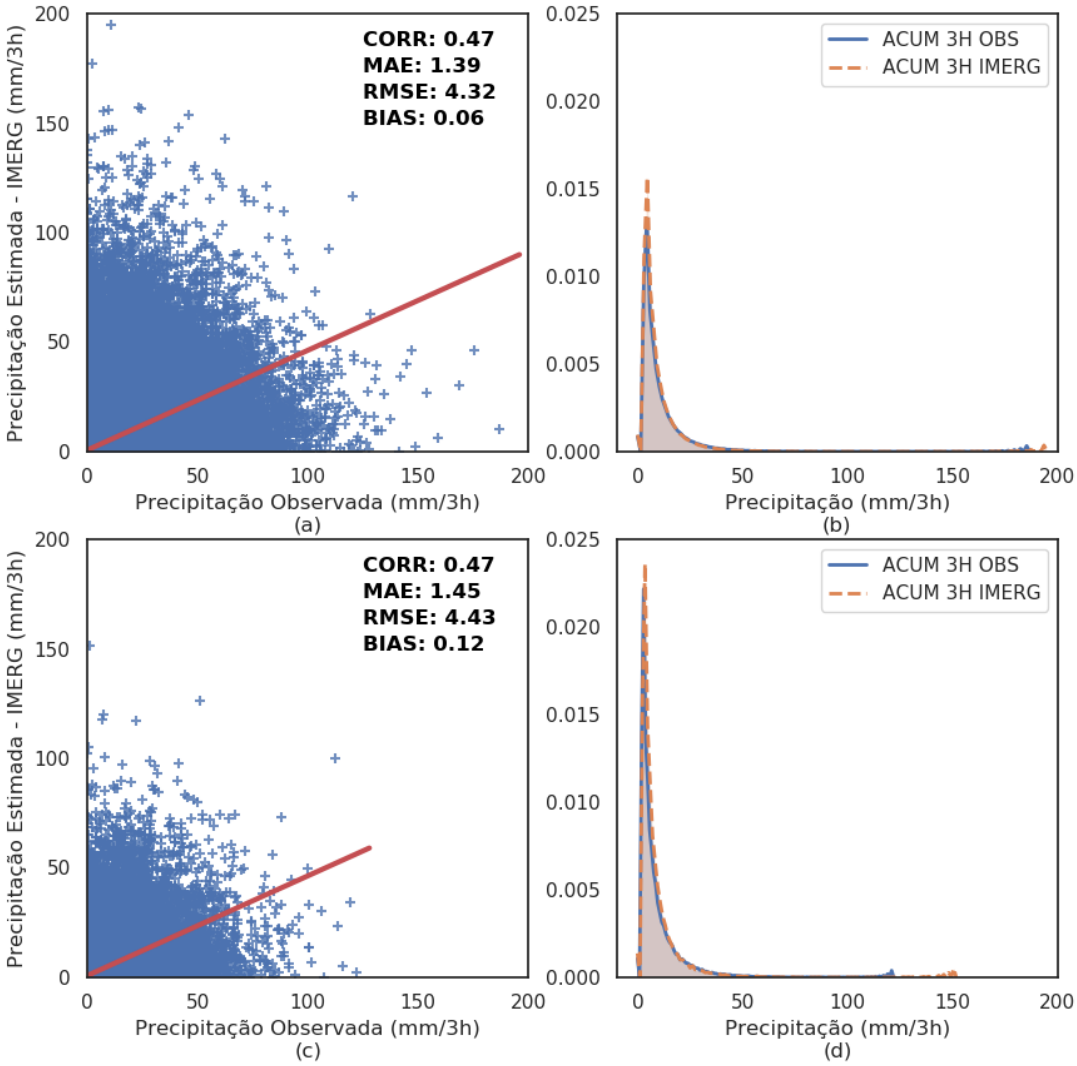



Gráfico de espalhamento das estimativas de precipitação IMERG sub-diária em relação às medidas de pluviômetros acumuladas a cada 3 horas, onde (a) refere-se ao conjunto de dados de treinamento e (c) ao conjunto de validação, e o comparativo do histograma de cada um dos conjuntos de dados, onde (b) refere-se ao conjunto de dados de treinamento e (d) ao conjunto de validação.

Fonte: Produção do Autor.

Muitas vezes a utilização de alguns dados não visa aplicar uma relação direta com o alvo (chuva na superfície), mas sim inserir alguns comportamentos para que a rede possa vir a assimilar. A terceira variável a ser analisada é justamente uma das quais não se tem relação direta, mas que implica na representação de um comportamento médio das chuvas sobre diferentes

regiões, neste caso o ciclo diurno médio das chuvas para cada mês, medido a partir de um período de 10 anos provinda do MERGE sub-diário. A ideia por trás do uso desta informação é a tentativa de representar a climatologia no processo de *downscaling*. Mesmo conhecendo a má representatividade direta dessas informações com relação à chuva é comum verificar seu comportamento em análises exploratórias, de modo a identificar possíveis desempenhos da rede neural já construída. A Figura 5.5 mostra a relação entre as medidas observadas e o ciclo diurno médio (valores fixos para cada mês) da precipitação. Na figura é possível observar o gráfico de espalhamento entre as variáveis, Figura 5.5(a), assim como a correlação entre elas (número no canto superior direito) e o histograma de frequência relativa, Figura 5.5(b). Apesar de o produto instantâneo apresentar uma correlação muito alta (Vide figura anterior), a correlação com a média da mesma variável é bem menor. Contudo, a distribuição dos dados é similar, mantendo as devidas proporções dos valores, muito mais baixos na média. Isso se dá devido à alta variabilidade dos dados com relação à média decenal, o que já é esperado. Trabalhos como o de Garcia et al. (2015) mostraram que o uso da informação da climatologia em técnicas automáticas de início e fim da estação chuvosa são importantes para melhor modelar a chuva. Logo, é justamente a intenção desses dados, inferir o que seria a climatologia, o que pode dar uma certa memória ao algoritmo quanto ao comportamento da chuva para determinado ponto, mesmo que em intensidades menores. Sendo assim, apesar dos resultados mostrarem que de maneira síncrona essas informações não são altamente correlacionadas, ainda são úteis ao processo físico climatológico da chuva. De todo modo, pode ser notado pela linha de tendência da Figura 5.5(a) que quanto maior a chuva maior a sua média climatológica dentro do ciclo diurno.

Figura 5.5: Comparativo entre a média decenal mensal do MERGE sub-diário e o acumulado de precipitação observada a cada 3 horas.

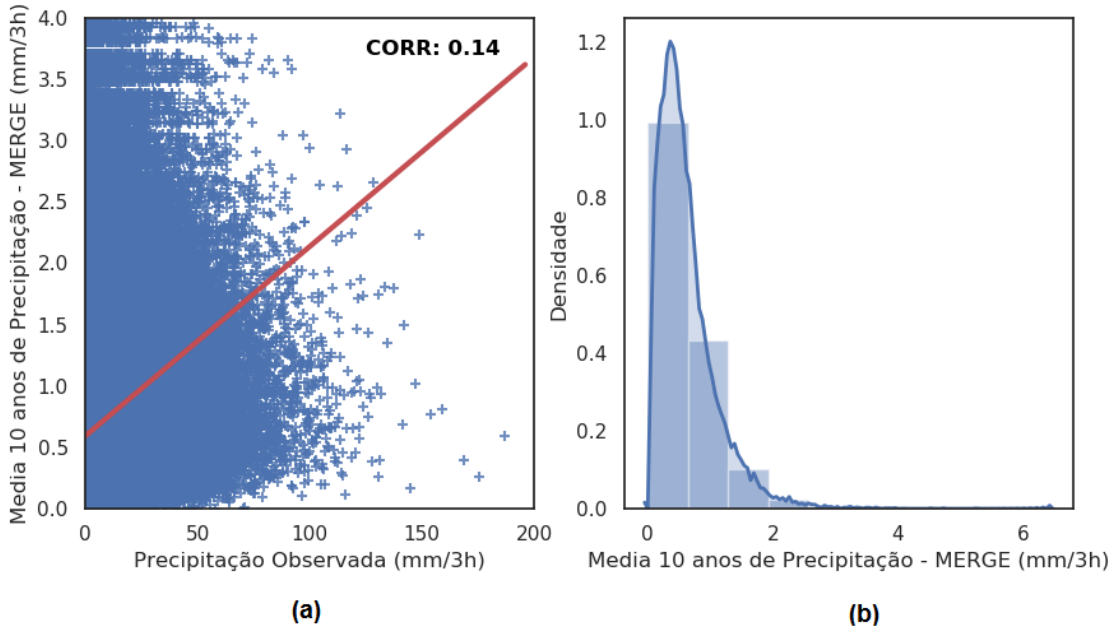



Gráfico de espalhamento da média decenal mensal de precipitação sub-diária calculada com o MERGE sub-diário em relação às medidas de pluviômetros (a) e o histograma da média decenal (b).

Fonte: Produção do Autor.

Dentre os dados utilizados neste estudo aqueles que melhor representam as nuvens são as temperaturas de brilho no espectro do infravermelho, providos dos satélites GOES sobre a América do Sul. Com esses dados foram calculados a média aritmética (i.e. um *proxy* da altura média das nuvens) e a variância da grandeza (i.e. representação de como esses topos de nuvens variaram) a cada três horas. A Figura 5.6 apresenta o espalhamento e a distribuição das médias de temperatura de brilho, no qual comparado com as medidas de chuva observada é possível perceber uma relação inversamente proporcional, reafirmada pelo valor de correlação negativo, ou seja, quanto menor a temperatura de brilho, maior a precipitação. Este tipo de comportamento já existe na literatura e já foi usado para desenvolver algoritmos de estimativa de precipitação, como o Autoestimator (VICENTE et al., 1998).

Figura 5.6: Comparativo entre a média a cada 3 horas da temperatura de brilho e o acumulado de precipitação observada no mesmo período.

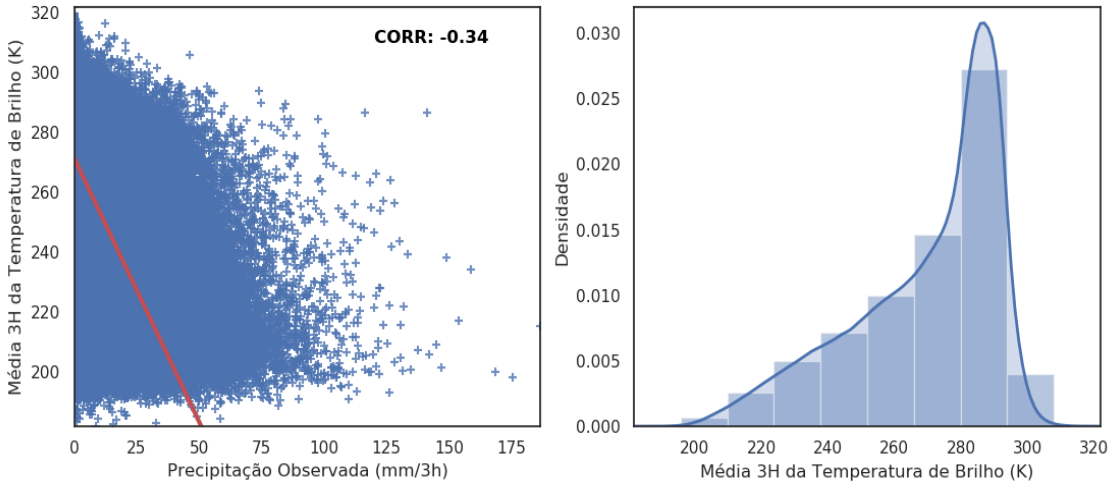



Gráfico de espalhamento da média da temperatura de brilho em relação à precipitação (a) e histograma da média das medidas de temperatura de brilho (b).

Fonte: Produção do Autor.

É possível perceber no gráfico acima que os valores de temperatura do topo da nuvem não se limitam a temperaturas baixas, variando em torno de aproximadamente 190 a 320 K ou -83 a 47 °C. Valores de temperatura de brilho acima de 273K (0°C) geralmente não são associados a chuvas por algoritmos de estimativa de precipitação por regressão como o autoestimador (VICENTE), esta é uma limitação das técnicas que não consideram a chuva provinda de nuvens quentes. Neste sentido, a RNA pode apresentar um ganho, aprendendo uma maneira de considerar este tipo de chuva em suas análises, contudo este é um problema antigo em sensoriamento remoto da atmosfera e não está incluído no escopo deste trabalho.

A variável anterior relaciona a chuva à temperatura de brilho média no intervalo de horas, no caso da ocorrência da precipitação durante este intervalo, a água presente na atmosfera, em geral na forma de gelo, vai para a superfície e é esperado que ocorra uma variabilidade dos valores de temperatura brilho no topo das nuvens e/ou nos locais da ocorrência. Para poder representar esta variação, foi introduzida nos dados de entrada a sua variância dentro do

mesmo intervalo. Esta variável possibilitaria que a RNA identifique possíveis variações que estejam associadas à presença ou não de chuva, uma complementação a média. Com relação à variância da temperatura de brilho é possível perceber, conforme a Figura 5.7(a), valores muito dispersos e uma baixa correlação em relação à precipitação, porém a distribuição dos dados de variância, Figura 5.7(b) se assemelha muito a distribuição da precipitação apresentada em exemplos anteriores. Tal comportamento indica que existe, de modo geral, uma certa associação entre os dados, contudo, esta relação pode não estar, tanto espacialmente, como temporalmente, bem colocalizada. Se a RNA for capaz de identificar um padrão espaço-temporal, esses dados podem ser importantes para definir onde deve estar ou não chovendo. Neste caso, só a partir da validação da rede será possível quantificar esta informação. Mais detalhes da verificação do desempenho da RNA serão dados nas seções posteriores.

Figura 5.7: Comparativo entre a variância a cada 3 horas da temperatura de brilho e o acumulado de precipitação observada no mesmo período.

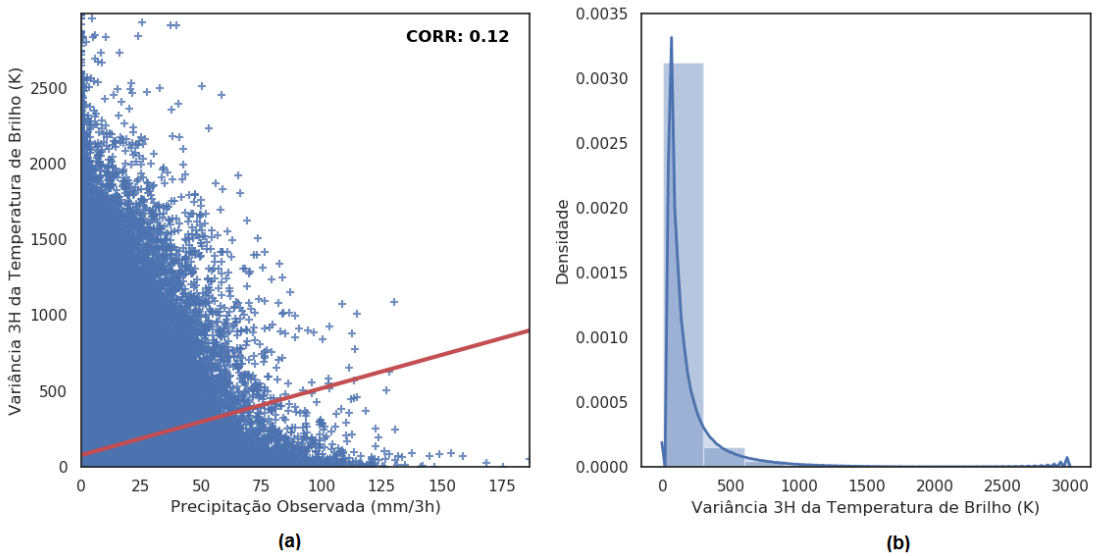



Gráfico de espalhamento da variância da temperatura de brilho em relação à precipitação (a) e a frequência relativa dos valores de variância da temperatura de brilho (b).

Fonte: Produção do Autor.

Outra variável importante para verificar a disponibilizada de água na atmosfera é a coluna total de água. Em uma comparação direta entre a chuva e esta variável notou-se que a correlação foi baixa, por volta de 0,16. O gráfico de dispersão apresentado na Figura 5.8(a) mostra que a medida é menos dispersa quando comparada a chuvas de maior intensidade, o que mostra que essa variável pode ser importante na determinação da chuva de sistemas mais severos. O histograma apresentado na Figura 5.8(b) mostra que a maioria das medidas encontra-se entre os valores 40 e 60, e considerando que a maioria das medidas de precipitação observada é igual ou próxima de zero, é possível afirmar que mesmo sem a presença de chuva na superfície a quantidade de água na atmosfera é significativamente alta, o que pode interferir nas estimativas satelitais.

Figura 5.8: Comparativo entre a média a cada 3 horas da coluna de água total e o acumulado de precipitação observada no mesmo período.

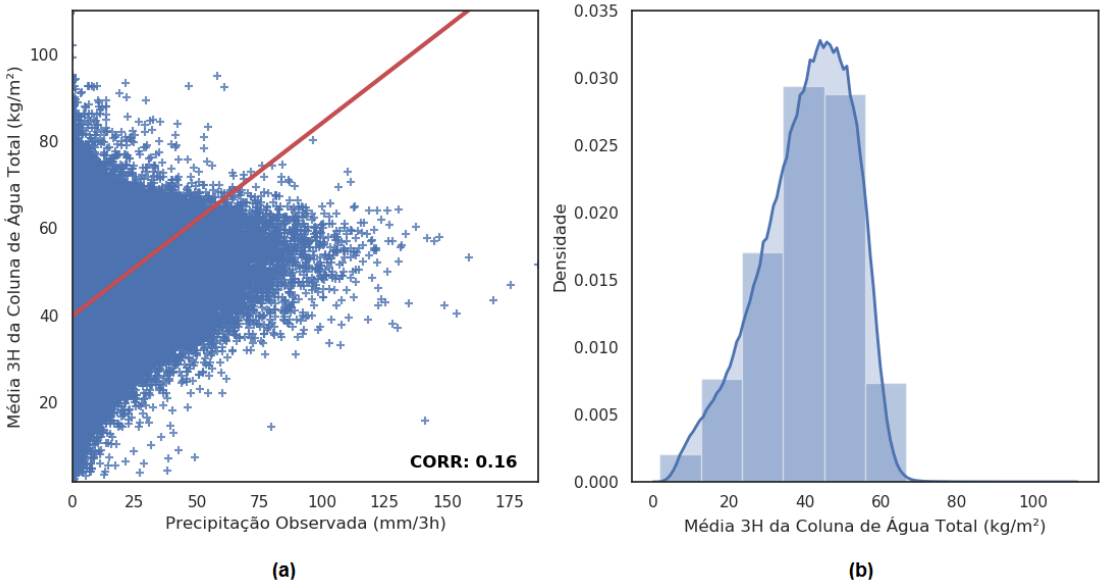



Gráfico de espalhamento da variável coluna de água total em relação à precipitação (a) e a distribuição dos valores de coluna de água total (b).

Fonte: Produção do Autor.

A intrusão de umidade em certas camadas da atmosfera providas de baixos níveis, ou o secamento a partir da advecção de ar seco proveniente de camadas da atmosfera superior, pode estar associada à intensificação ou não das chuvas. Neste estudo a representação da umidade relativa do ar foi a partir de uma média aritmética entre 16 níveis, gerando uma estimativa da umidade relativa média em uma coluna vertical entre 1000 e 500 hPa, em seguida foi realizada a média para cada 3 horas. A Figura 5.9 abaixo mostra a correlação entre a medida calculada e a medida de interesse. As medidas apresentaram baixa correlação e valores dispersos. Contudo, considerando que todos os dias sem chuva foram removidos do conjunto de dados, e todos os registros analisados apresentam chuva em algum horário do dia, é esperado que nos horários que não tenha ocorrência de chuva, haja ainda uma quantidade de umidade no ar decorrente da chuva que precipitou em um horário anterior ou que ainda irá precipitar em um horário posterior. Além disso, ambientes mais úmidos pós-precipitação ou associados a nuvens de chuva estratiformes, que produzem baixa precipitação e são mais frequentes, podem influenciar nestas condições.

Figura 5.9: Comparativo entre a média a cada 3 horas de umidade relativa e o acumulado de precipitação observada no mesmo período.

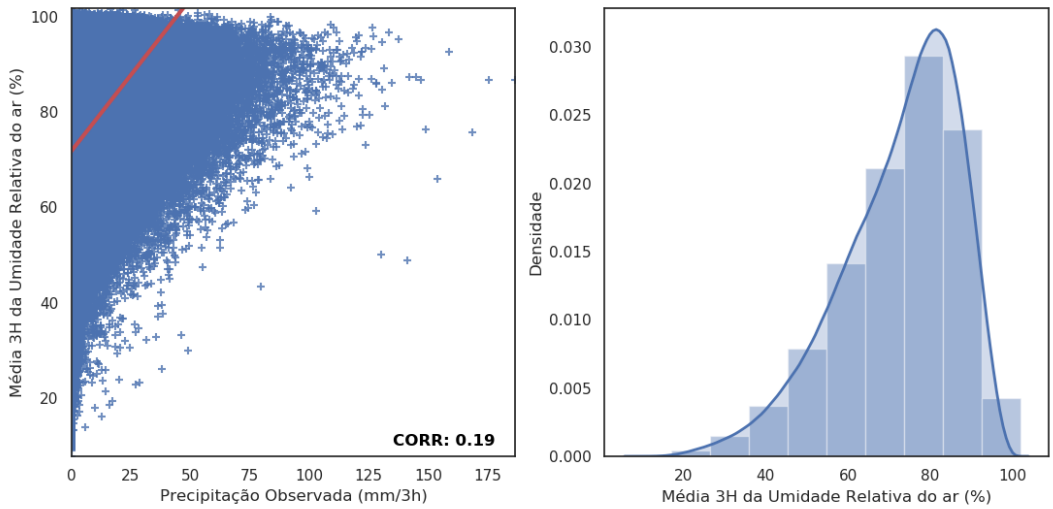



Gráfico de espalhamento da variável umidade relativa do ar em relação à precipitação (a) e a distribuição dos valores de umidade relativa do ar (b).

Fonte: Produção do Autor.

De acordo com o histograma apresentado na Figura 5.9(b), a maioria das medidas de umidade encontra-se próximo de 80%, indicativo da presença de uma condição atmosférica mais propícia à formação de nuvens. Essa informação mesmo sem uma correlação alta com a intensidade da chuva pode fornecer a RNA subsídios quanto ao estado físico das nuvens e do ambiente circundante e auxiliar na classificação dos horários com e sem chuva.

Outra variável importante, principalmente para diagnosticar as chuvas orográficas, é a intensidade e direção do vento. As variáveis referentes ao vento foram definidas a partir de estimativas representadas pela componente $u$ e pela componente $v$. Esses dois vetores são utilizados para calcular a direção e a magnitude do vento, e para essa conversão foram utilizadas as equações 5.1 e 5.2 descritas pela ECMWF no manual de utilização do produto ERA5 (GUILLORY, 2020).

$$mag = \sqrt{u^2 + v^2} \tag{5.1}$$

$$dir = 180 + arctan\left(\frac{v}{u}\right) * \left(\frac{180}{\pi}\right) \tag{5.2}$$

Depois de calculado direção e magnitude do vento em 850 hPa, nível já usado em outros algoritmos, como o hidroestimator, versão atualizado do autoestimator, para aplicar efeitos orográficos na estimativa de chuva, foi calculada a média para cada 3 horas das duas variáveis. Para cálculo da magnitude média foi realizada a média aritmética entre o valor do horário de referência e os horários anterior e seguinte, conforme a equação 5.3.

$$\overline{mag_{HH}} = \frac{mag_{(HH-1)} + mag_{HH} + mag_{(HH+1)}}{3} \tag{5.3}$$

Para o cálculo da média em 3 horas da direção do vento foi aplicada a equação de média circular, descritas nas equações 5.4, 5.5 e 5.6 para $\overline{cos} \neq 0$:

$$\overline{sin} = \frac{1}{n}\sum_{i=1}^{n} \sin(dir_i) \tag{5.4}$$

$$\overline{cos} = \frac{1}{n}\sum_{i=1}^{n} \cos(dir_i) \tag{5.5}$$

$$\overline{dir} = 180 + arctan\left(\frac{\overline{sin}}{\overline{cos}}\right) * (\frac{180}{\pi}) \tag{5.6}$$

Como as medidas de direção do vento não trouxeram informações relevantes, devido a grande variabilidade e pouco significado físico quando analisadas sem levar em consideração sua localização e os sistemas atuantes, suas figuras foram disponibilizadas na Figura A1, no Apêndice A. Contudo, a magnitude do vento apresentou um certo grau de informação que pode ser fisicamente relacionada com a precipitação acumulada em 3 horas, como pode ser observada na Figura 5.10. Cabe ressaltar que a média em 3 horas da magnitude do vento ficou entre 0 e 35 m/s, e que ventos mais fracos estavam associados as chuvas mais intensas. Apesar desta relação e da baixa correlação observada nos comparativos, à linha de tendência da Figura 5.10(a) indica que há uma relação em que quanto maior a magnitude do vento, maior seria a intensidade da precipitação. Ainda, acredita-se que essa variável pode ser determinante para indicar o local em que ocorrerá a precipitação, ou ter uma relação assíncrona com a mesma. Estas baixas correlações (-0.04 e 0.02) são esperadas, uma vez que diversos sistemas meteorológicos e a própria condição ambiente local, em todo o território brasileiro, podem mudar

drasticamente, logo, estes são suscetíveis à variabilidade altíssima. Contudo, esta é uma variável importante para representar a chuva sob os efeitos de brisas marítimas e pluviais, ou mesmo, associada à presença de sistemas que organizam a precipitação em grande escala, como as zonas de convergência, e em regiões de topografia mais alta. Sendo esta última, a orografia, que também não apresentou valores significantes de correlação síncrona com a precipitação, Figura A2.

Figura 5.10: Comparativo entre a média a cada 3 horas da magnitude do vento em 850 hPa e o acumulado de precipitação observada no mesmo período.

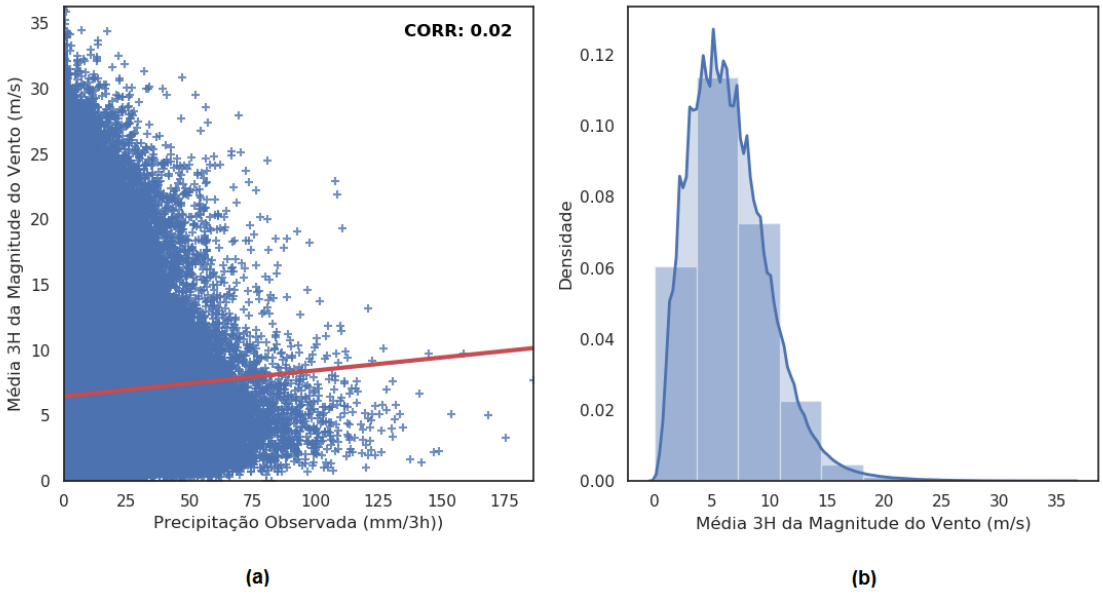



Gráfico de espalhamento da variável magnitude do vento em relação à precipitação (a) e a distribuição dos valores (b).

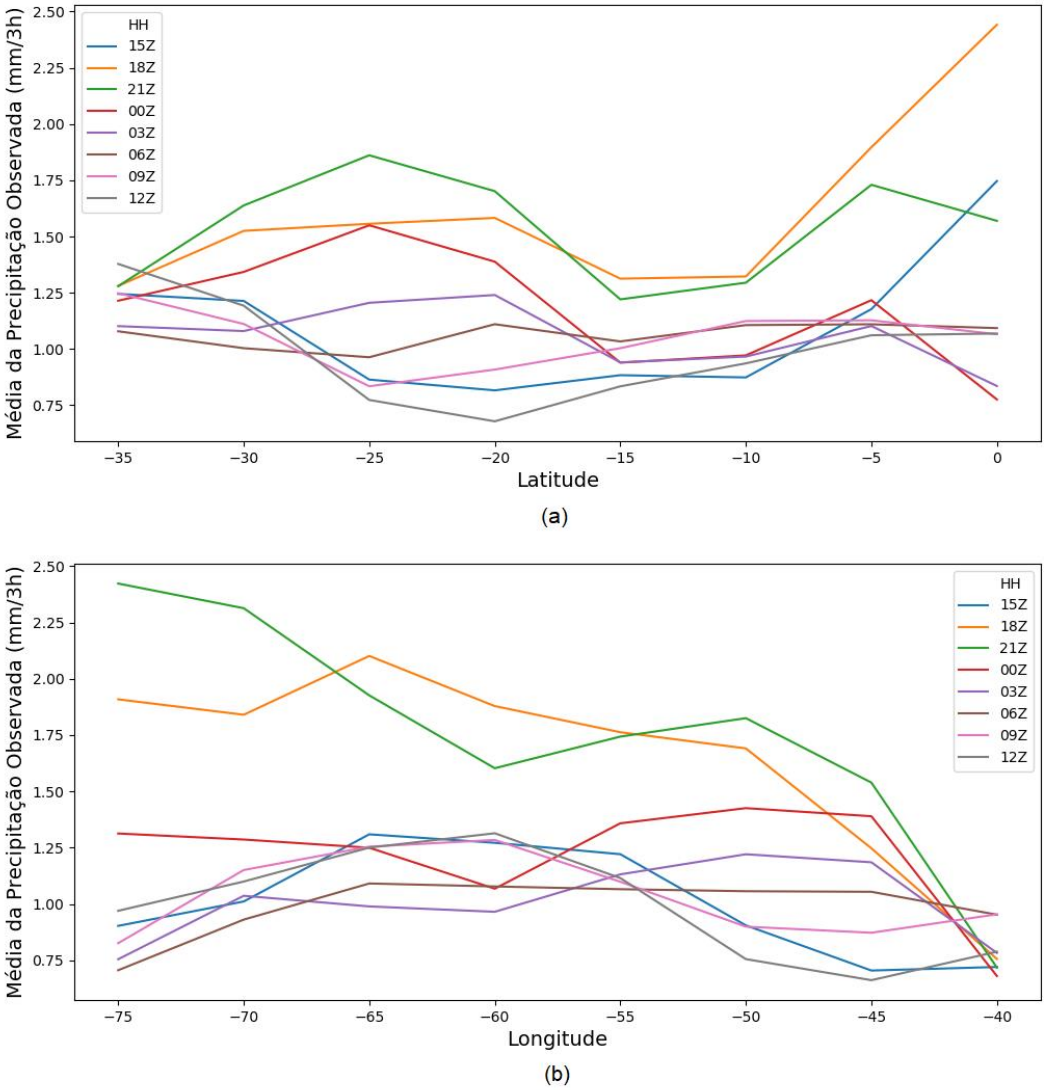
Fonte: Produção do Autor.

A baixíssima correlação apresentada pela orografia talvez possa ser justificada devido à interação da mesma com a precipitação ocorrer em regiões específicas e sob condições que ainda precisam ser melhor estudadas, como foi analisado na região Amazônica por Machado et al. (2018) e no Rio de Janeiro por Ceron et al. (2019). Cabe ressaltar que os pluviômetros utilizados nesta pesquisa estão distribuídos por diferentes altitudes, o que é ideal para a generalização da RNA.

As variáveis de latitude e longitude, também apresentaram baixa correlação com a intensidade da chuva. Como as relações mostradas nos gráficos de espalhamento e frequência relativa não trazem informações relevantes, estes não foram mostrados aqui e sim no Apêndice A, Figura A3. Contudo, cabe ressaltar que é uma variável usada justamente para que a RNA entenda a alta variabilidade das chuvas em diferentes localizações geográficas. Como existem diversos regimes de precipitação e condições climáticas que variam sazonalmente, anualmente e alguns ciclos diurnos podem ser bem definidos em certos lugares do Brasil, uma alta variabilidade é esperada.

As variações médias por latitude e longitude podem ser observadas na Figura 5.11. Nota-se na Figura 5.11(a) que os ciclos diurnos da precipitação por latitude podem ser observados. Nela é possível verificar que as precipitações mais ao norte mostram picos a tarde e início da noite que já são conhecidos na literatura (COHEN et al., 1995). Além disso, é possível notar o ciclo bem definido em latitudes que estão associadas a sudeste e centro-oeste do Brasil onde há um ciclo diurno bem definido. Nota-se também que as maiores taxas estariam associadas a regiões na faixa latitudinal do Sudeste e da região Amazônica. Já com relação ao comportamento longitudinal, Figura 5.11(b), nota-se que há um aumento da taxa de chuva, principalmente entre 18 e 21h, de leste para oeste, o que pode ser um reflexo das baixas taxas de chuva na região nordeste do Brasil, a leste, e das altas taxas de precipitação observadas nas regiões mais a oeste. Em resumo, observou-se que a representação da precipitação média está de acordo com a literatura, (KIKUCHI E WANG, 2008), o que pode ajudar a RNA a melhor estimar a chuva em diferentes localidades.

Figura 5.11: Variação da precipitação média por latitude e longitude.



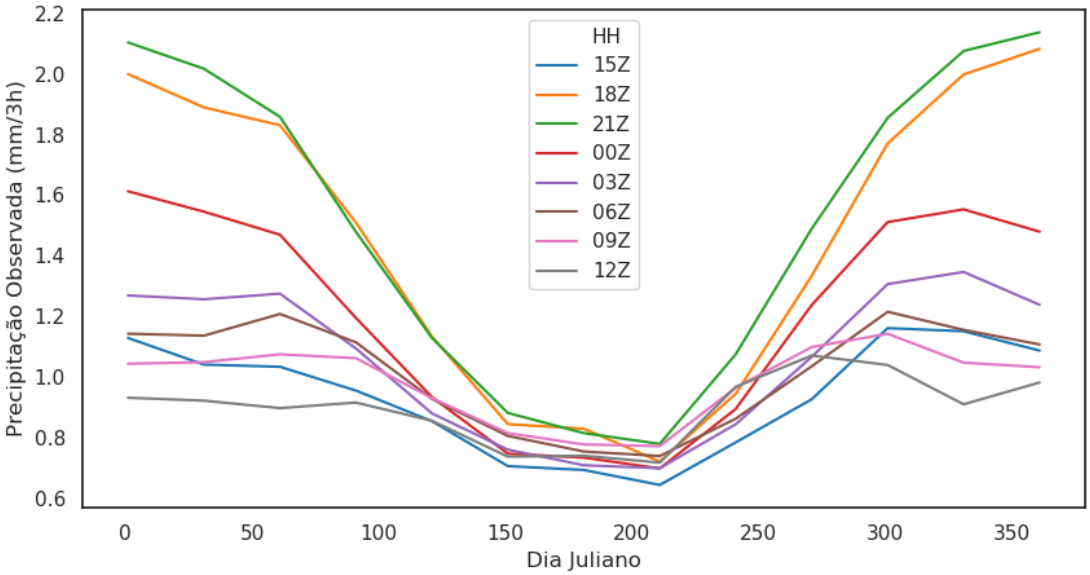

Variação média por latitude (a) e variação média por longitude (b)

Fonte: Produção do Autor.

Assim como as longitudes e latitudes, que foram utilizadas para representar a variabilidade da chuva em diferentes regiões e locais do Brasil, outra variável que pode representar os ciclos temporais é justamente o tempo, representado aqui pelo dia juliano. Apesar das variáveis apresentar baixíssima correlação, conforme Figura A4, é possível observar claramente o ciclo anual em todos os oito horários pela Figura 5.12. Nota-se pela figura que chuvas mais intensas ocorrem no verão (início e fim do ano), e o tempo mais seco pode ser

observado no inverno (meio do ano). Também é possível observar que no geral os horários de maior intensidade são entre 18 e 00 horas. Essas informações podem ser valiosas para definir alguns ciclos diurnos em regiões onde eles são bem definidos, como a Amazônica.

Figura 5.12: Precipitação média horária (3h) para todos os dias do ano (1-366) para o período de análise (2000-2020).




Fonte: Produção do Autor.

## 5.3 *Downscaling* por RNA

### 5.3.1 Definindo as RNAs

A definição da RNA a ser usada neste trabalho será explicada nesta seção. Foi testado inicialmente dois tipos distintos de RNAs, a DNN (*Deep Neural Network*) e a RNN (*Recurrent Neural Network*), ambas para toda extensão do país. Conforme susodito, a definição do número de camadas ocultas e neurônios nessas camadas são definidas através de uma abordagem individual considerando o problema, nesse caso um número estocástico de neurônios e camadas ocultas na camada intermediaria foi definido para construção do modelo inicial da RNA, e ajustados através de *tuning*, assim como a taxa de aprendizado e o tamanho dos *minibatches* de treino. Nesta etapa as variáveis

que apresentaram baixíssimas (<|0.1|) ou nenhuma correlação foram descartadas de modo a diminuir o custo computacional, estas variáveis são listadas na Tabela 5.2. Ressalta-se que, uma vez definido a melhor rede, estas variáveis serão adicionadas para verificar sua contribuição na versão final do algoritmo.

Tabela 5.2: Variáveis utilizadas conforme correlação.

| **Produto** | **Resolução temporal** | **Quant. (valores)** | **Corr.** | **Utilizada** |
|---|---|---|---|---|
| Longitude | Pontual | 1 | 0,04 | Não |
| Latitude | Pontual | 1 | 0,01 | Não |
| Altitude/Topografia | Pontual | 1 | 0,00 | Não |
| Dia Juliano | Pontual | 1 | 0,00 | Não |
| Precipitação diária MERGE (mm) | Diária | 1 | --- | Sim |
| Média 10 anos de Precipitação (mm) | Sub-diária | 8 | 0,14 | Sim |
| Precipitação Micro-onda (mm) | Sub-diária | 8 | 0,47 | Sim |
| Média da Temperatura de brilho no IR (K) | Sub-diária | 8 | 0,34 | Sim |
| Variância da Temperatura de brilho no IR (K) | Sub-diária | 8 | 0,11 | Sim |
| Umidade relativa média entre 1000-500 mb (%) | Sub-diária | 8 | 0,19 | Sim |
| Direção do vento (graus) | Sub-diária | 8 | 0,04 | Não |
| Magnitude do vento (m/s) | Sub-diária | 8 | 0,02 | Não |
| Coluna de Água Total (kg $m^{-2}$) | Sub-diária | 8 | 0,16 | Sim |

Fonte: Produção do Autor.

Partindo do modelo inicial, após o treinamento as informações referentes aos erros MAE e MSE, tempo de execução e consumo de memória, foram analisados, em seguida os hiperparâmetros foram ajustados e novamente

submetidos ao treinamento, observando piora ou melhora do modelo, até a obtenção do modelo ideal considerando o equipamento utilizado, além disso, cada configuração foi submetida ao treino mais de uma vez para evitar resultados referentes a um mínimo local. Após sucessivas execuções e ajustes da camada intermediária um modelo de RNA para cada um dos dois tipos DNN e RNN foi definido, conforme Tabela 5.3, sendo os neurônios da camada de entrada equivalentes ao número de variáveis analisadas com correlação maior que 0.1 e o número de neurônios na camada de saída equivalente aos 8 valores sub-diários pretendidos.

Tabela 5.3: Hiperparâmetros dos modelos iniciais.

| | **DNN** | **RNN** |
|---|---|---|
| **Neurônios na camada de entrada** | 49 | 49 |
| **Neurônios na camada intermediária** | 128, 64, 32 e 16 | 128,256,128,256, 128,32 e 16 |
| **Neurônios na camada de saída** | 8 | 8 |
| **Taxa de aprendizado** | 0,001 | 0,001 |
| **Tamanho dos subconjuntos (batch)** | 512 | 512 |

Fonte: Produção do Autor.

### 5.3.2 Validação das RNAs

Na tabela 5.4, a coluna treinamento apresenta os erros obtidos pelos modelos iniciais aplicados em um subconjunto aleatório equivalente a 10% do conjunto de dados separado para treinamento, ou seja, são dados que não foram utilizados no processo de aprendizagem, mas provenientes de estações de superfície que foram utilizadas, já a coluna validação apresenta o erro obtido na execução de cada um dos modelos no conjunto de dados validação, o qual as medidas provem de estações localizadas em posições geográficas diferentes das utilizadas no processo de aprendizagem da RNA. É possível

observar que os resultados de ambas foram próximos, contudo, os valores de MAE e MSE foram melhores para a DNN do que aqueles observados pela RNN para quase todos os parâmetros, exceto para o MAE aferido no treinamento, onde a RNN foi levemente superior (-0,02 mm).

Tabela 5.4: Resultados dos modelos iniciais.

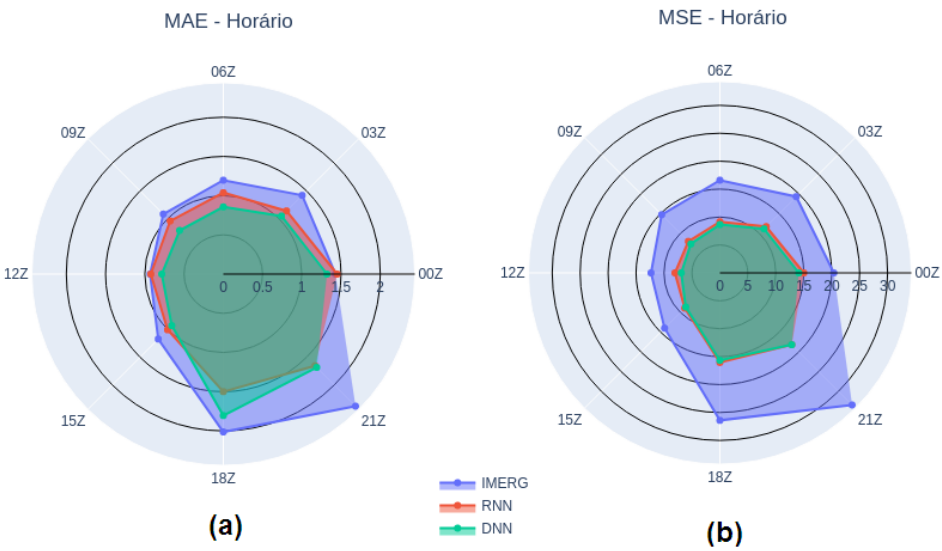
| | **DNN** | | **RNN** | |
|---|---|---|---|---|
| | **Treinamento** | **Validação** | **Treinamento** | **Validação** |
| **MSE** | 10,51 | 11,33 | 10,79 | 11,88 |
| **MAE** | 1,14 | 1,15 | 1,12 | 1,20 |

Fonte: Produção do Autor.

Para melhor compreender o ganho desses estimadores, usaremos o IMERG como comparativo, já que este é vastamente utilizado pela comunidade científica, e a versão mais atualizada do programa GPM (*Global Precipitation Mesurements*) para estimativa de chuva, além de também aquele utilizado pelo MERGE para suprir informações de precipitação em áreas sem medidas na superfície. O índice utilizado aqui para definir se os resultados foram melhores ou não, será principalmente o MSE, outros índices (e.g. BIAS, MAE e RMSE) que não foram redundantes também serão apresentados nesta seção, quando não, estarão no apêndice B. Ainda, nota-se que os índices aqui apresentados foram calculados apenas para o todo o período para as estações de validação. Na comparação entre os diferentes algoritmos de estimativa de precipitação com relação à observação, notou-se que ambas as RNAs apresentaram índices MAE e MSE melhores que o IMERG para todos os horários, sendo a DNN ligeiramente melhor que a RNN, conforme pode ser observado pela Figura 5.13. A figura em questão mostra que apesar da diferença, os estimadores apresentaram um comportamento similar, sendo os horários das 18 e 21 GMT aqueles que apresentaram o maior erro em todas as três

estimativas, o que é justificado pela ocorrência maior de chuvas intensas nesses horários na maioria das regiões do Brasil.

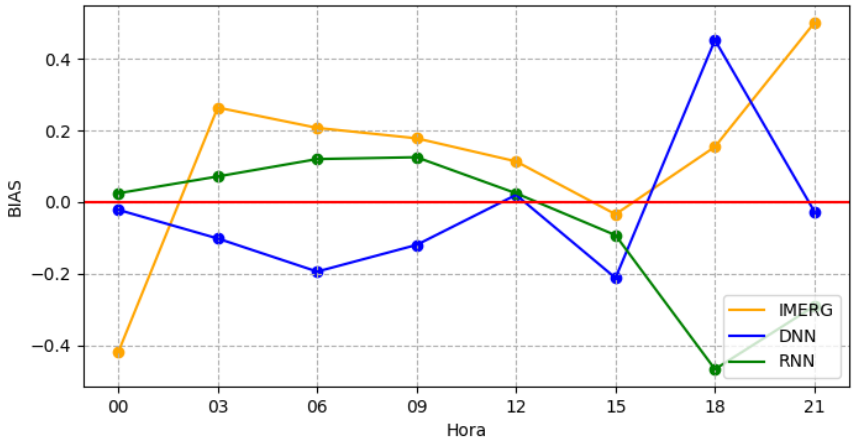
Figura 5.13: MAE e MSE por horário para IMERG, RNN e DNN.




Fonte: Produção do Autor.

Contudo, a Figura 5.14, referente ao BIAS, mostrou um comportamento bem diferente entre os estimadores, inclusive apresentando tendências opostas em alguns horários. Nota-se pela figura que a RNN, apesar de apresentar bons resultados em horários diurnos, mostrou maior subestimativa nos horários de maior intensidade de chuva, 18 e 21 GMT. Enquanto o IMERG superestimou nesses horários. Já a DNN, apresentou comportamento similar a RNN, com bons resultados ao longo do dia, com a diferença de que nos horários de maior intensidade de chuva, a DNN apresentou superestimativa apenas às 18 GMT. No geral o BIAS de cada um dos estimadores foi 0,12 para o IMERG, -0,02 para DNN e -0,06 para a RNN. Logo, isto mostra também que o ciclo diurno da convecção tem grande papel na modelagem da chuva em algumas regiões e que ainda é necessárias melhorias em suas estimativas.

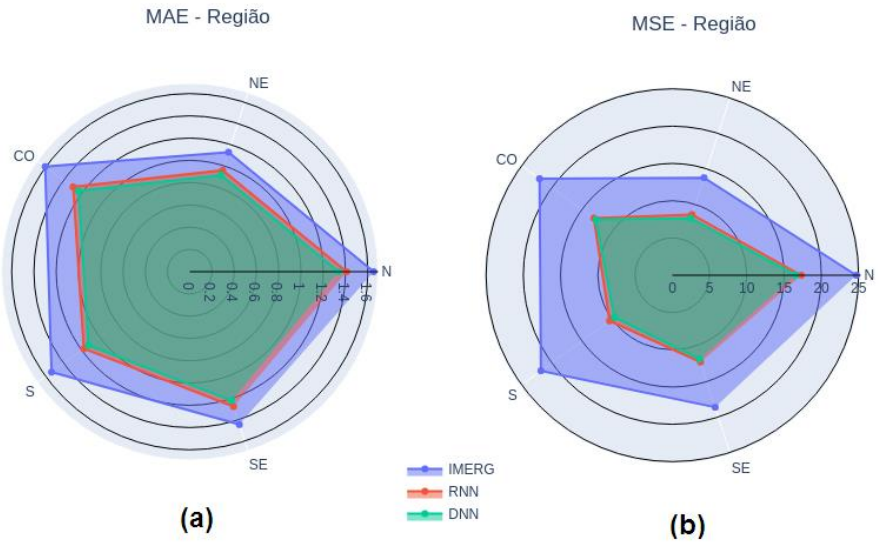
Figura 5.14: BIAS por horário para IMERG, RNN e DNN.




Fonte: Produção do Autor.

De modo a verificar o desempenho de ambas RNAs em função dos aspectos regionais e da estação do ano, validações foram feitas levando em consideração esses aspectos. Conforme pode ser observado na Figura 5.15 as regiões que apresentaram melhor desempenho foram o Sudeste e Nordeste, principalmente esta última.

Figura 5.15: MAE e MSE por região para IMERG, RNN e DNN.

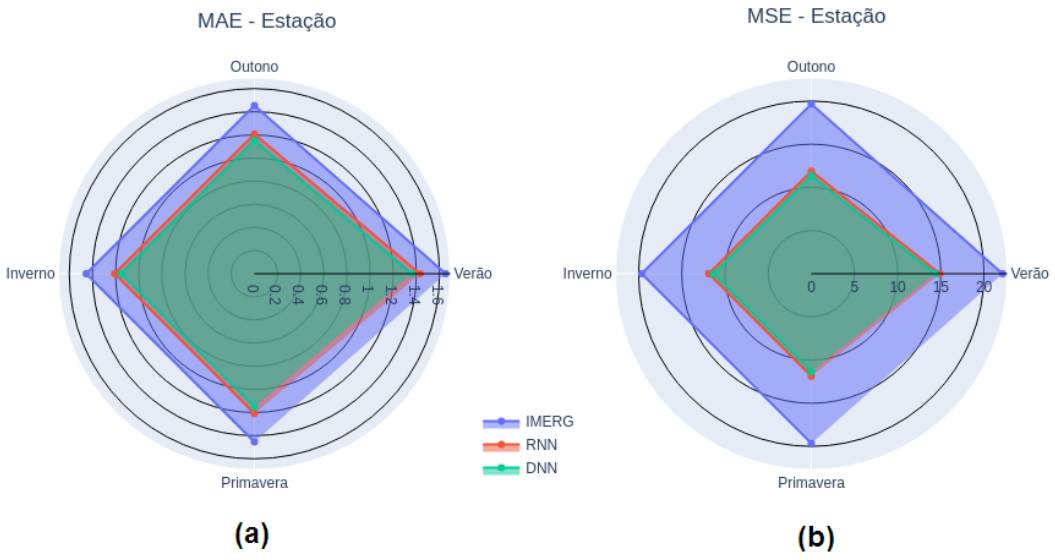



Fonte: Produção do Autor.

Considera-se que como a região Nordeste é aquela que tem as taxas de precipitação mais baixas, devido ao regime pluviométrico local, e existe um desbalanceamento dos dados, onde chuvas mais fracas são predominantes, espera-se que o algoritmo a represente melhor.

A região que apresentou os piores resultados foi a região Norte, uma possível causa para esta performance provavelmente está associada a alta variabilidade das chuvas, proporcionada, principalmente, pelo ciclo diurno da convecção sobre a região, que por sua vez provocam altos valores de taxa de precipitação associados a tempestades, a região climatologicamente também apresenta o maior acumulado pluviométrico anual que segundo Nimer, 1989 é causada pela falta de homogeneidade e sazonalidade em relação à pluviosidade. Já em relação às estações do ano, Figura 5.16, nota-se pouca variação, contudo o verão foi a que mostrou o maior erro, o que corrobora com as análises anteriores, onde uma das prováveis fontes de erros são as tempestades locais que são frequentes nesta época do ano e provocam alta variabilidade na precipitação. Com relação às técnicas, ambas RNAs se mostraram mais eficientes do que a técnica do IMERG baseando-se apenas no MAE e no MSE. Sendo a DNN um pouco melhor que a RNN.

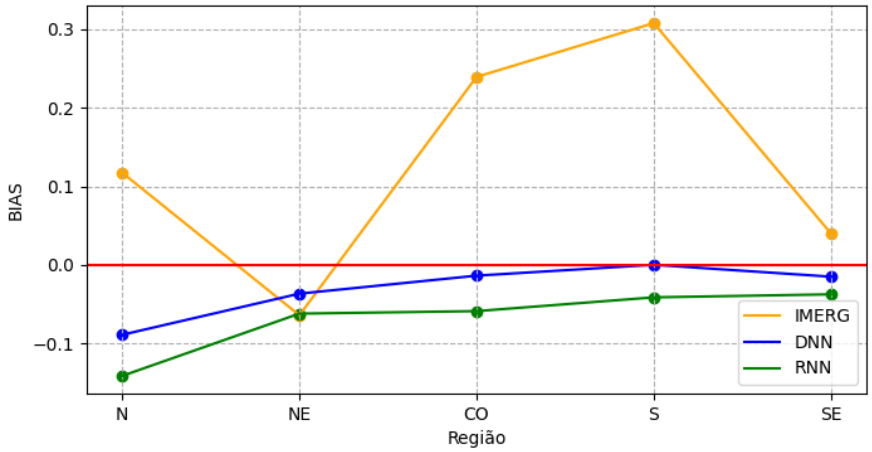
Figura 5.16: MAE e MSE por estação do ano para IMERG, RNN e DNN.




Fonte: Produção do Autor.

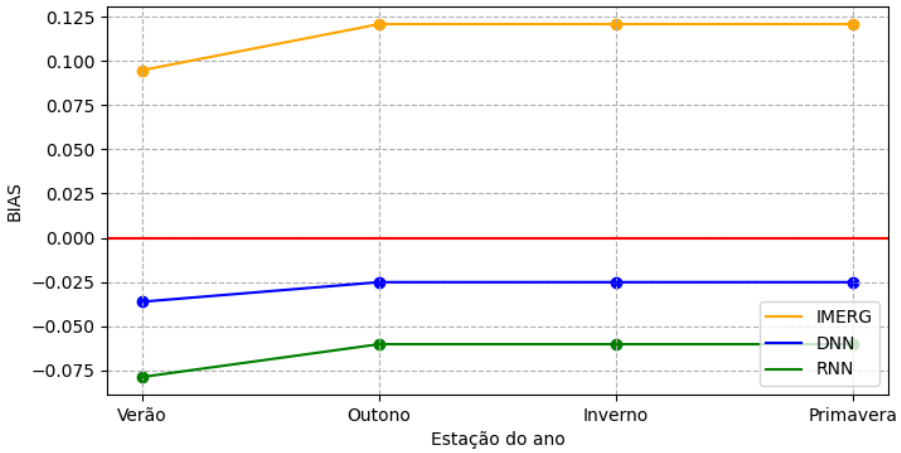
O BIAS por região e por estação do ano, conforme Figuras 5.17 e 5.18, mostraram resultados mais estáveis indicando uma tendência do IMERG em superestimar, subestimando apenas na região Nordeste, enquanto as RNAs apresentaram tendência para subestimar a chuva, principalmente a RNN. Contudo, as diferenças são relativamente pequenas.

Figura 5.17: BIAS por região para IMERG, RNN e DNN.




Fonte: Produção do Autor.

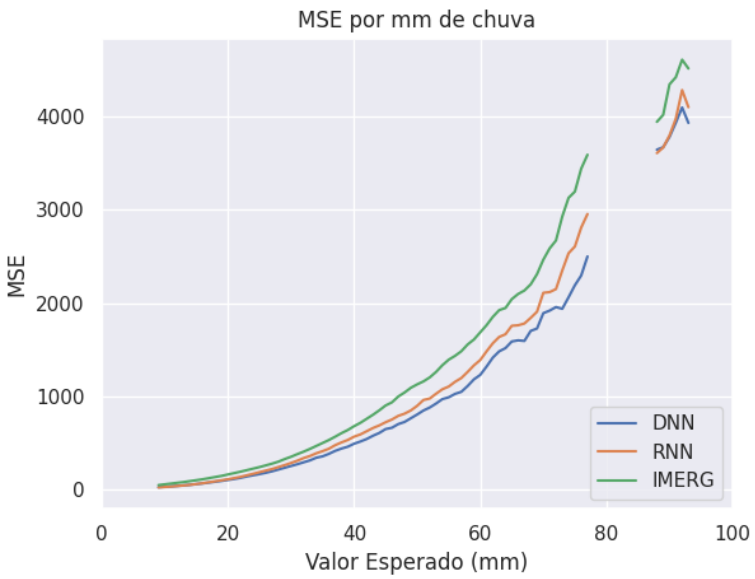
Figura 5.18: BIAS por estação do ano para IMERG, RNN e DNN.




Fonte: Produção do Autor.

Também é possível observar, conforme Figura 5.19 (erro por taxa de chuva), que em relação aos milímetros de chuva o MSE aumenta progressivamente, o que pode ser explicado pela distribuição dos dados utilizados para treinamento, que em sua maioria encontra-se próximo a zero e somente uma minoria acima de 60 mm, onde inclusive pode ser observado no gráfico uma ausência de dados entre 75 e 90 mm na amostra utilizada, porém esse comportamento é devido à própria natureza da chuva, ou seja, com uma menor distribuição de dados de chuva intensa, o modelo tende a ter uma performance pior para essas taxas de precipitação.

Figura 5.19: Evolução do MSE conforme mm de chuva.



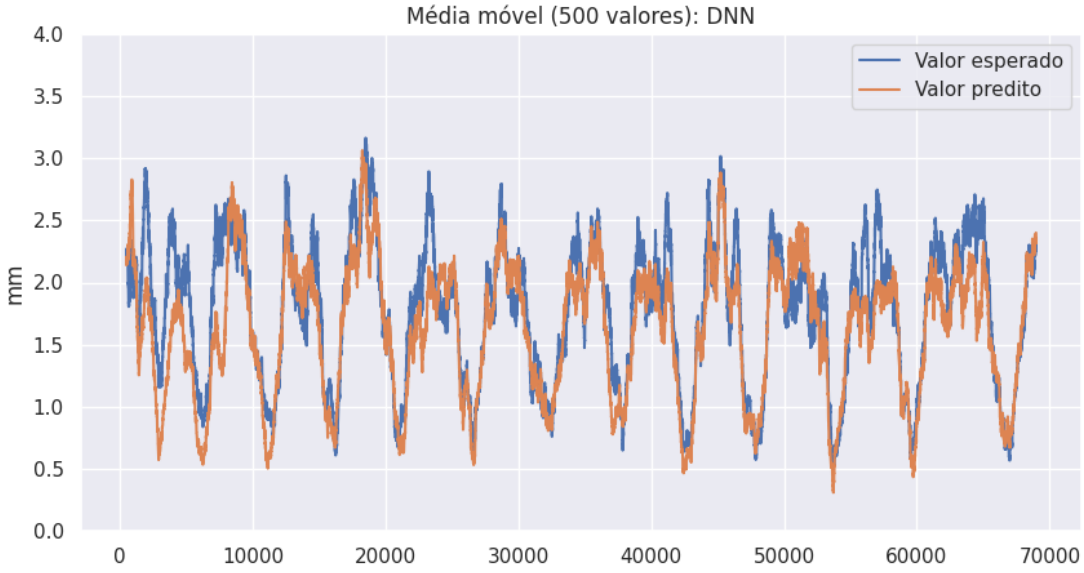

Fonte: Produção do Autor.

Apesar disso, ainda assim, houve melhora em relação à estimativa do IMERG. Uma provável razão para estes resultados deve estar no fato que a RNA apresenta taxas de chuvas menores em suas estimativas, e quando há eventos com maior taxa de chuva, como as tempestades, é provável que haja uma subestimativa da precipitação. Contudo, globalmente, as estatísticas mostram

um melhor resultado devido ao fato de que as taxas de chuvas menores são muito mais frequentes do que as altas.

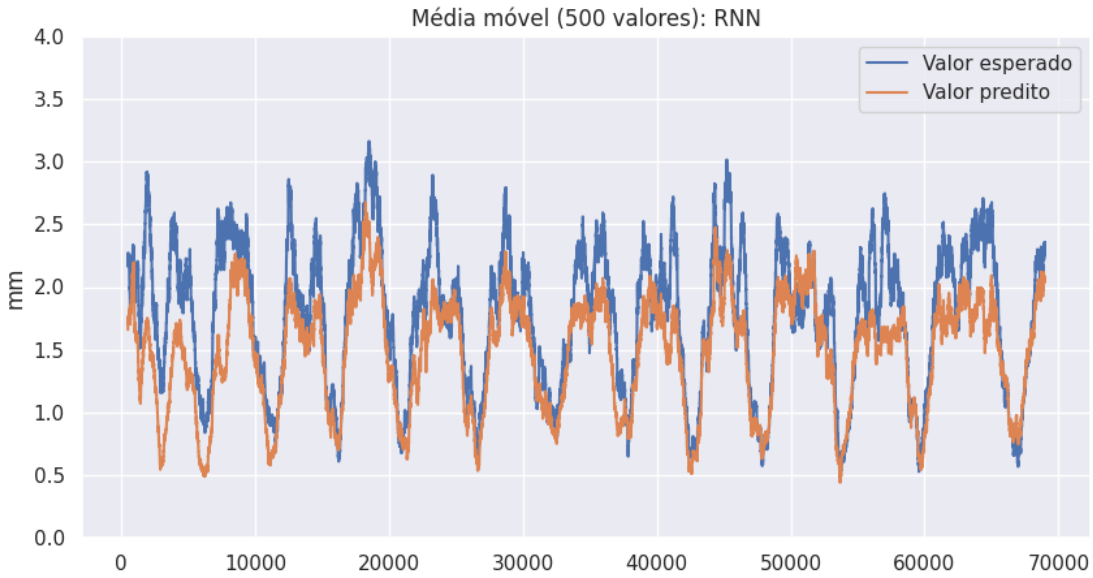
Pode-se notar pelo recorte realizado na série temporal da média móvel da chuva estimada pela DNN, Figuras 5.20 e pela RNN, Figura 5.21, que ambas RNAs mantiveram um comportamento semelhante à observação. A DNN se mostrou mais representativa do que sua concorrente, diferentemente do IMERG representado pela Figura 5.22, onde os picos máximos estimados ficaram acima dos observados, caracterizando uma tendência, já observada anteriormente de superestimar a chuva.

Figura 5.20: Média móvel da DNN.



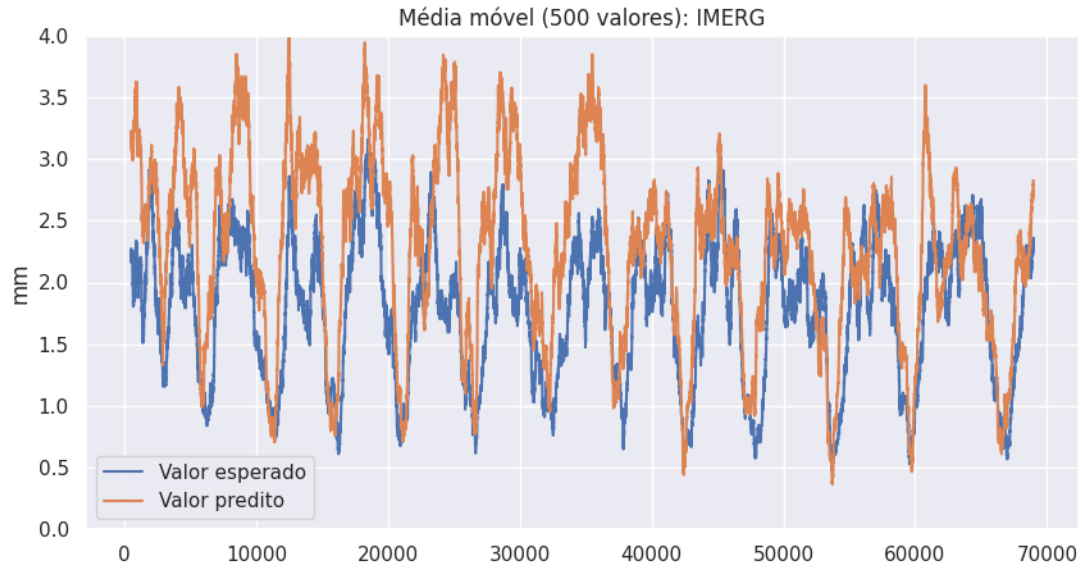

Fonte: Produção do Autor.

Figura 5.21: Média móvel da RNN.




Fonte: Produção do Autor.

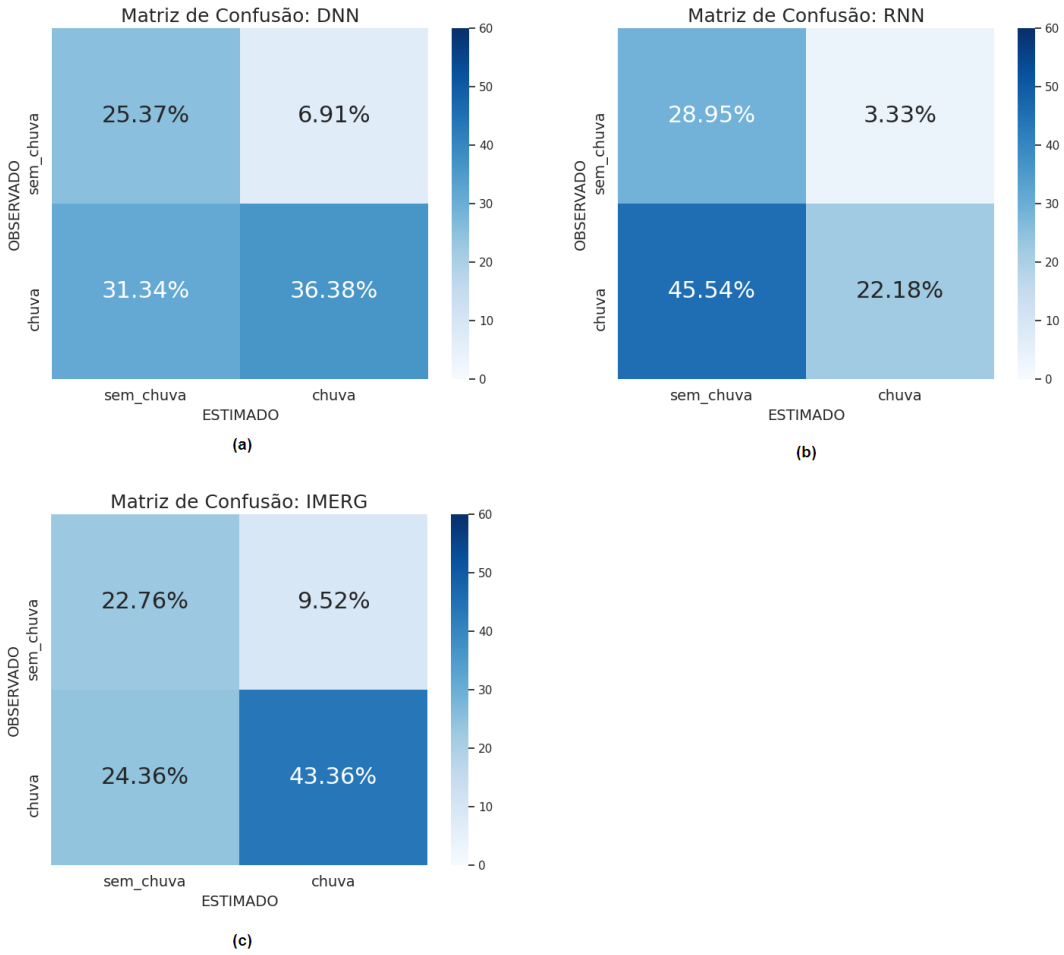
Figura 5.22: Média móvel do IMERG.




Fonte: Produção do Autor.

Para verificar a acurácia de cada uma das RNAs, foi realizada a validação cruzada dos dados, exibido na Figura 5.23.

Figura 5.23: Validação cruzada com valores mínimos de chuva em 0 mm.

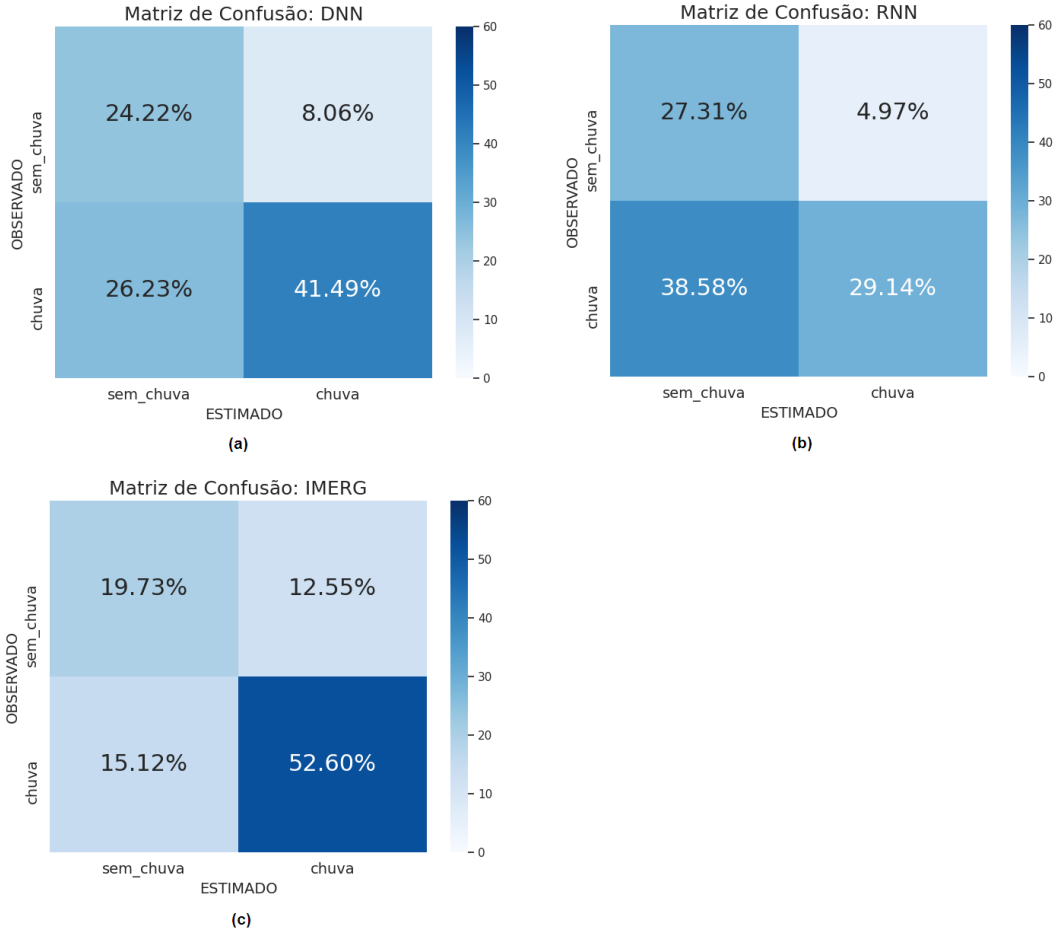



Fonte: Produção do Autor.

O resultado não se refere a capacidade da RNA em acertar a intensidade da chuva, mas sim em acertar o horário de ocorrência de chuva. Nota-se pela figura que a DNN apresenta valores intermediários, entre a RNN o IMERG, tanto para os valores assertivos de não-chuva e chuva, assim como, para as falhas, o que estatisticamente pode ter dado a ela o melhor desempenho geral mostrado acima. Ou seja, a RNN é melhor para determinar quando não há chuva, contudo o IMERG é melhor para quando há chuva, e ambos erram mais nas condições opostas. Entretanto, é sabido que alguns algoritmos de

precipitação são criados para estimar chuva apenas acima de um certo limiar, de modo a testar esta condição, a Figura 5.24 mostra a validação cruzada onde não-chuva são valores menores que 0,1 mm. Nesse sentido, observa-se que à maioria das medidas em ambas RNAs foram mais assertivas, sendo 65,71% de acerto para a DNN e 56,41% para a RNN. Porém, ambas ficaram abaixo do IMERG que obteve 72,33% de acerto, ou seja, este teve mais precisão na relação dicotômica (i.e. aquilo chamado de "*screening*" da chuva ou triagem). Contudo, conforme avaliações anteriores, o IMERG apresentou uma tendência a superestimar os valores, enquanto as RNAs a subestimar. Tomando como referência o IMERG, a menor diferença (4,37%) foi observada para DNN.

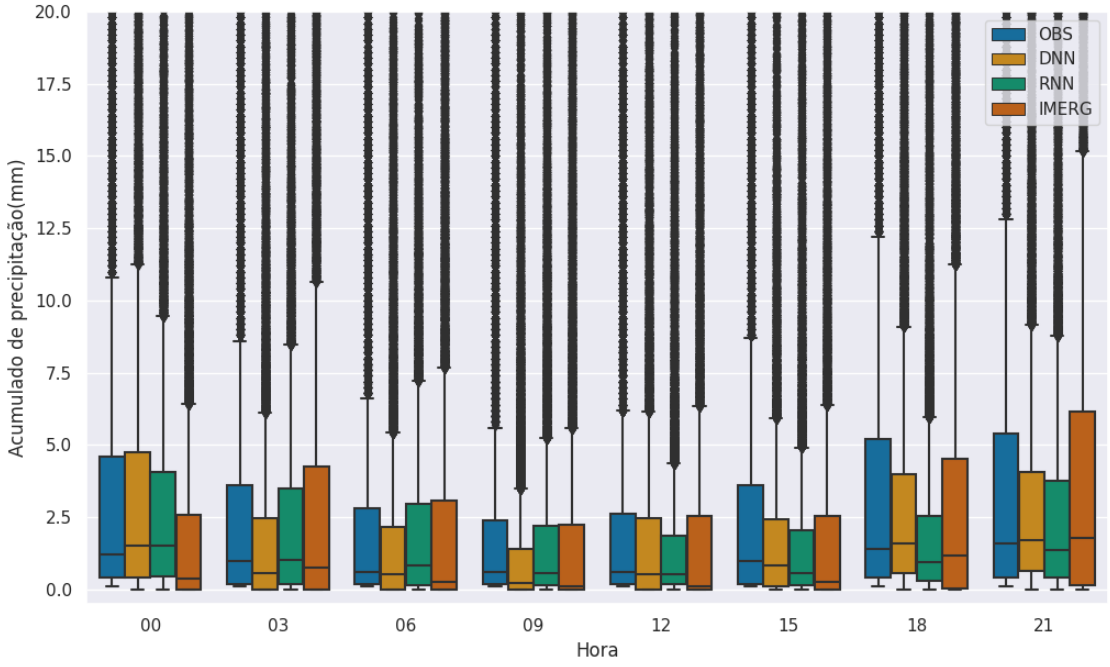
Figura 5.24: Validação cruzada com valores mínimos de chuva em 0,1 mm.




Fonte: Produção do Autor.

Apesar da maioria dos índices até aqui apontarem para um bom resultado, algumas análises como aquelas por meio de boxplot, exibido na Figura 5.25, trouxeram algumas incertezas. A figura mostra os valores de chuva para cada horário em todos os métodos e a observação, calculado apenas onde há chuva observada. Aqui, o melhor resultado seria aquele cujas caixas forem parecidas a da observação (azul). Devido ao grande número de valores próximos a zero, a estatística das medidas até o 3° quartil, também resultou em valores muito baixos.

Figura 5.25: Comparativo estatístico dos dados observados e estimados.




Fonte: Produção do Autor.

No gráfico é possível observar as variações entre as estatísticas dos dados observados e os dados estimados, todas as métricas apresentam diferenças, os limites inferiores e o primeiro quartil de todas as estimativas, como esperado, ficaram muito próximos de zero não ultrapassando 1 mm, o 2° quartil ou mediana em todas as estimativas ficou abaixo de 2,5 mm, o 3° quartil ficou em torno de 6 mm sendo a maior variação ocorrida as 18 horas, onde o dado

observado apresentou 6,2 mm e a RNN 3,1 mm. A métrica que mais apresentou variação foram os limites superiores, o que também era esperado devido à quantidade menor de medidas altas de chuva, fazendo a RNA errar mais nesses casos. Analisando a figura, do ponto de vista subjetivo, nota-se que a DNN foi melhor para os horários das 00, 12 e 15 GMT. Enquanto a RNN aparenta ser melhor para as 03, 06 e 09 GMT. Os horários mais chuvosos, ou seja, as 18 e as 21 GMT, são aqueles que tanto o IMERG como a DNN chegam próximos, mas um aparenta superestimar e o outro subestimar, respectivamente. Esta análise sugere que talvez o uso de multitécnicas seja um caminho a ser investigado no futuro.

### 5.3.3 Validação das RNAs para um período fora do treinamento

Apesar do objetivo deste estudo ser o de analisar dados para o período de 20 anos (2000 a 2019) e produzir métricas para o cálculo do ciclo diurno, é interessante conhecer o comportamento da RNA quando aplicada em um período fora do utilizado para treinamento, para isso, ambas as RNAs foram aplicadas em dados do ano de 2020 e comparadas aos dados observados do período. No geral conforme Tabela 5.5, apesar dos erros apresentados serem um pouco maior que o observado anteriormente para as séries dentro do período do treinamento, o comportamento das RNAs se manteve.

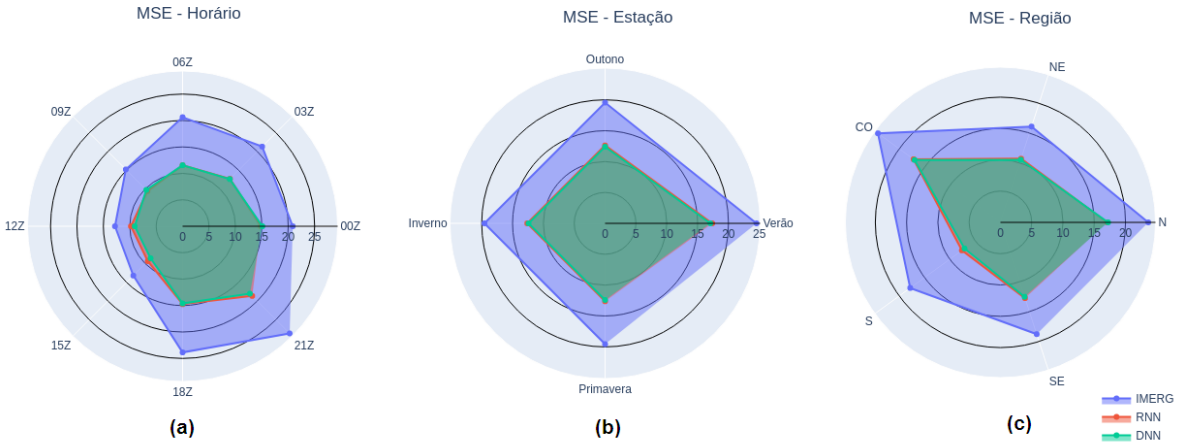
Tabela 5.5: Resultados dos estimadores aplicados no período de janeiro a dezembro de 2020 para a validação via MSE.

| | **DNN** | **RNN** | **IMERG** |
|---|---|---|---|
| **MSE** | 12,40 | 12,62 | 19,56 |
| **MAE** | 1,23 | 1,28 | 1,41 |
| **BIAS** | 0,15 | 0,11 | -0.14 |

Fonte: Produção do Autor.

O comportamento observado na primeira validação se manteve também para os horários, estação do ano e região, conforme observado na Figura 5.26 para o MSE, persistindo o resultado da DNN ligeiramente melhor que a RNN.

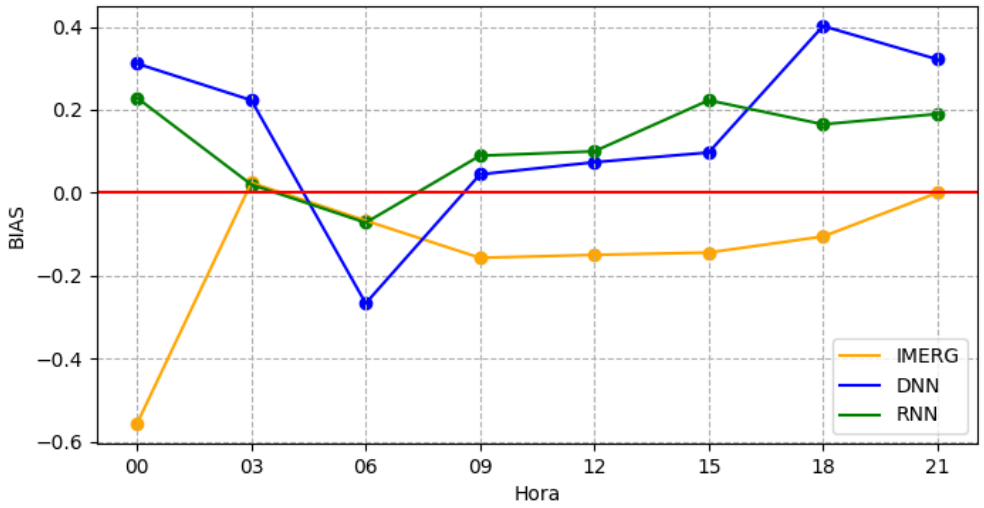
Figura 5.26: MSE por horário, estação e região para o ano de 2020.




Fonte: Produção do Autor.

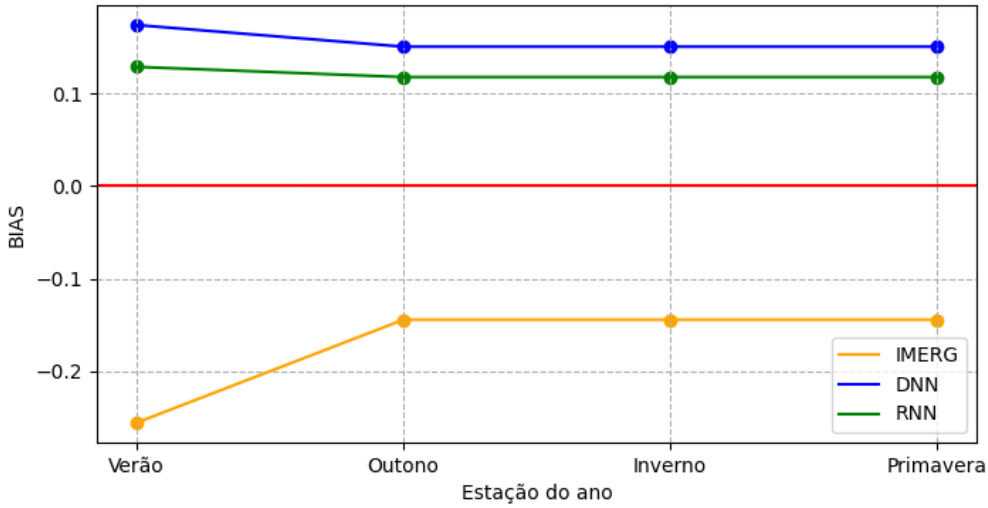
Apesar dos resultados dos estimadores para 2020 serem muito semelhantes aos resultados obtidos durante o treinamento, o BIAS para esse ano apresentou valores completamente opostos ao encontrado anteriormente para ambas as metodologias, conforme Figuras 5.27, 5.28 e 5.29. Onde as redes superestimaram em quase todos os horários, exceto para as 06 GMT.

Figura 5.27: BIAS por horário do IMERG, RNN e DNN para 2020.



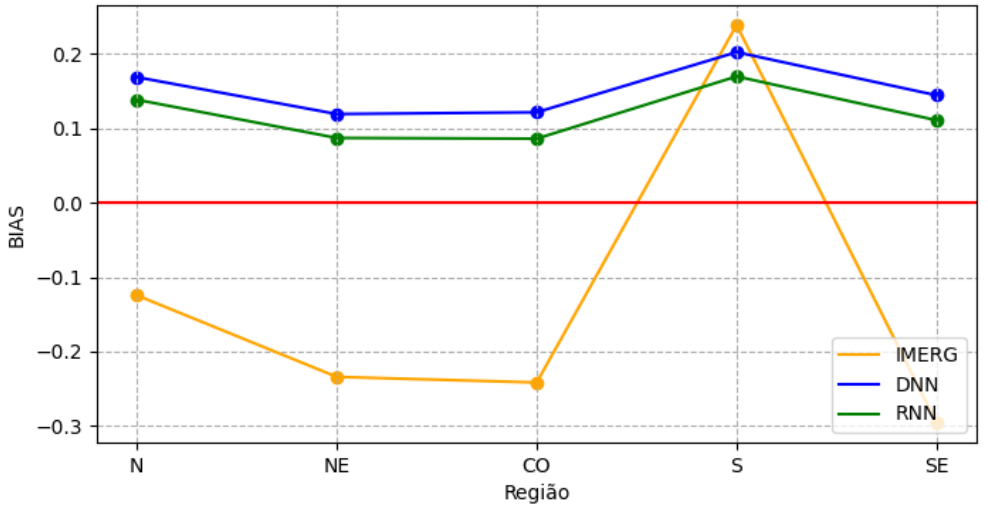

Fonte: Produção do Autor.

Figura 5.28: BIAS por estação do IMERG, RNN e DNN para 2020.




Fonte: Produção do Autor.

Figura 5.29: BIAS por região do IMERG, RNN e DNN para 2020.




Fonte: Produção do Autor.

Uma possibilidade para o resultado divergente pode ser devido a 2020 ter sido um ano atípico do ponto de vista climático. Segundo o INMET, vários estados extrapolaram marcas históricas de temperatura e chuva. Em São Paulo, por exemplo, o mês de fevereiro teve o maior acumulado de chuva desde que se

tem registro (1943), já o mês de março de 2020 foi considerado o quinto mais seco da história de São Paulo. No fim de junho um ciclone bomba se formou no Sul provocando temporais na região, em agosto nevou em Santa Catarina, Rio Grande do Sul e Curitiba, além de vários outros eventos. Apesar disso os modelos apresentaram resultados satisfatórios para o ano em conformidade com o treinamento.

### 5.3.4 Teste de sensibilidade da DNN

Como verificado nas seções anteriores, a DNN apresentou melhores resultados que a RNN e o IMERG para quase todas as condições. Neste sentido, esta seria a escolha mais prudente de modelo de estimativa da chuva para *downscaling*. O primeiro experimento para verificar um possível ganho na execução da RNA, foi acrescentar individualmente as variáveis ignoradas anteriormente devido à baixa correlação (<0.1), considerando que o MSE sem essas variáveis foi de 11,33 mm. Os resultados incluindo cada uma das variáveis, excluídas inicialmente, ao treinamento são apresentados na Tabela 5.6.

Tabela 5.6: Performance do modelo DNN com as variáveis pouco correlacionadas.

| **Produto** | **MSE (Treinamento)** | **MSE (Validação)** |
|---|---|---|
| Dia Juliano | 10,114 | 69,411 |
| Longitude | 10,470 | 152,034 |
| Latitude | 10,099 | 217,142 |
| Altitude/Orografia | 10,506 | 116,940 |
| Direção do vento (graus) | 10,257 | 11,330 |
| Magnitude do vento (m/s) | 10,415 | 11,323 |
| Direção e Magnitude do vento | 9,735 | 11,092 |

Fonte: Produção do Autor.

Durante o treinamento da DNN todas as seis variáveis apresentaram bons resultados, porém quando aplicadas ao conjunto de dados de validação, que possui dados de estações que não foram utilizadas no treinamento, as variáveis dia juliano, longitude, latitude e altitude afetaram negativamente a estimativa, aumentando o erro drasticamente. Já os dados de direção e magnitude do vento proporcionaram resultados melhores. Uma hipótese para a baixa performance da DNN com os dados de geolocalização seria:

a) a baixa densidade de pluviômetros em algumas regiões;

b) a organização da precipitação em grande escala teria mais influência do que as questões locais (segunda ordem);

c) as estações com erros sistemáticos e descontinuidade poderiam acentuar ainda mais o erro em um modelo menos generalizado.

A combinação entre direção e magnitude do vento proporcionou uma pequena melhora na DNN, mostrando que apesar da baixa correlação síncrona, o comportamento da variável é importante para definir algumas características da precipitação.

Um segundo experimento consistiu em treinar e executar a DNN por região, sendo estas, Norte (N), Nordeste (NE), Centro-Oeste (CO), Sul (S) e Sudeste (SE). Devido à quantidade variada de estações de superfície em cada região, o número de registros disponíveis para treinamento é diferente em cada uma delas, sendo assim o tamanho dos subconjuntos utilizados durante o treinamento (*batches)* para todo o Brasil, precisou ser revisto, isso porque utilizando todos os aproximadamente 670 mil registros com *batches* de 512 registros, era possível executar aproximadamente 1300 épocas de treinamento, reduzir o número de registros, e manter o mesmo tamanho de *batch* significaria menos épocas de treino, o que poderia não ser suficientes para encontrar o melhor resultado, sendo assim, o tamanho dos batches foi ajustado para cada região conforme Tabela 5.7.

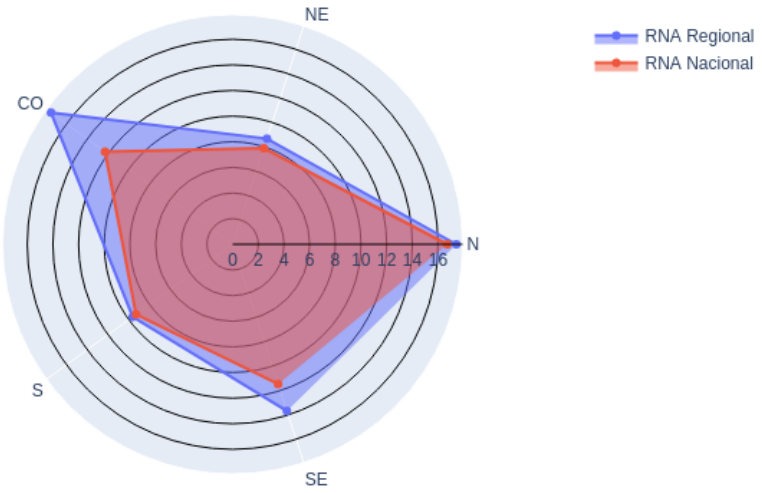
Tabela 5.7: Ajuste do tamanho de *batch* por região.

| **Região** | **Registros** | ***Batchs*** |
|---|---|---|
| Norte | 102467 | 64 |
| Nordeste | 153271 | 128 |
| Centro-Oeste | 107165 | 64 |
| Sul | 135742 | 128 |
| Sudeste | 169054 | 128 |

Fonte: Produção do Autor.

Comparando o MSE de uma determinada região obtido a partir da DNN treinada exclusivamente com os registros dessa região, com o mesmo erro obtido a partir da DNN treinada com todos os registros disponíveis (667.699), conforme a Figura 5.30 e a Tabela 5.8, é possível verificar que a DNN treinada com dados de todas as cinco regiões (DNN-Brasil) apresentou resultados melhores que quando treinada individualmente. O que mostra que a rede foi generalizada o suficiente para ter uma performance melhor do que aquelas regionais.

Figura 5.30: Comparativo do MSE resultante do treinamento regionalizado.




Fonte: Produção do Autor.

Tabela 5.8: Resultado do treinamento por região.

| Região | MSE (Treino) | MSE (Teste) | MSE (Validação) | MSE (DNN-Brasil para cada região) |
|---|---|---|---|---|
| Brasil | 9,379 | 9,735 | 11,092 | |
| Norte | 13,001 | 16,220 | 17,430 | 16,721 |
| Nordeste | 6,747 | 8,643 | 8,671 | 7,880 |
| Centro-Oeste | 11,739 | 14,450 | 17,495 | 12,288 |
| Sul | 7,514 | 9,486 | 9,623 | 9,301 |
| Sudeste | 9,940 | 12,516 | 13,675 | 11,470 |

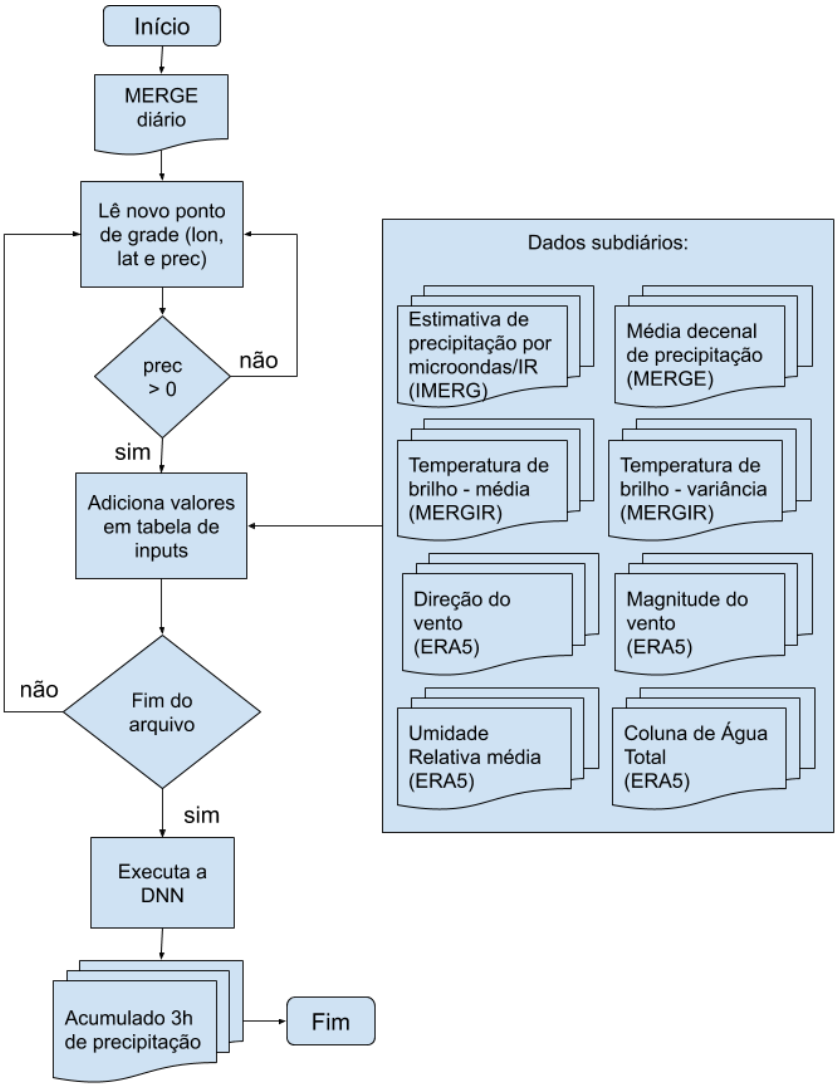
Fonte: Produção do Autor.

Note que a última coluna da Tabela 5.8 refere-se apenas aos MSE calculados usando a DNN com os parâmetros para todo o Brasil, e não regionalmente como listado, sua adição foi apenas para questões de visualização e comparação dos resultados. Outros índices da DNN-Brasil podem ser visualizados no anexo B, Figuras B1, B2 e B3 uma vez que não divergiram ou trouxeram informações adicionais a esta análise.

## 5.4 Aplicação da RNA em dados de grade e simulação de caso de uso

Para verificar a eficiência da utilização da RNA no *downscaling* da precipitação diária em uma situação real, foi realizada uma simulação conforme o fluxograma apresentado na Figura 5.31, onde o processo inicia com a entrada da estimava de precipitação diária MERGE, em seguida é realizado um *loop* para leitura de cada valor de ponto de grade e quando este for maior que zero, os demais dados auxiliares para a mesma coordenada geográfica são lidos e a RNA é executada, ao final da leitura do arquivo diário os dados gerados são agrupados em oito arquivos sub-diários.

Figura 5.31: Fluxograma para execução da RNA.




Fonte: Produção do Autor.

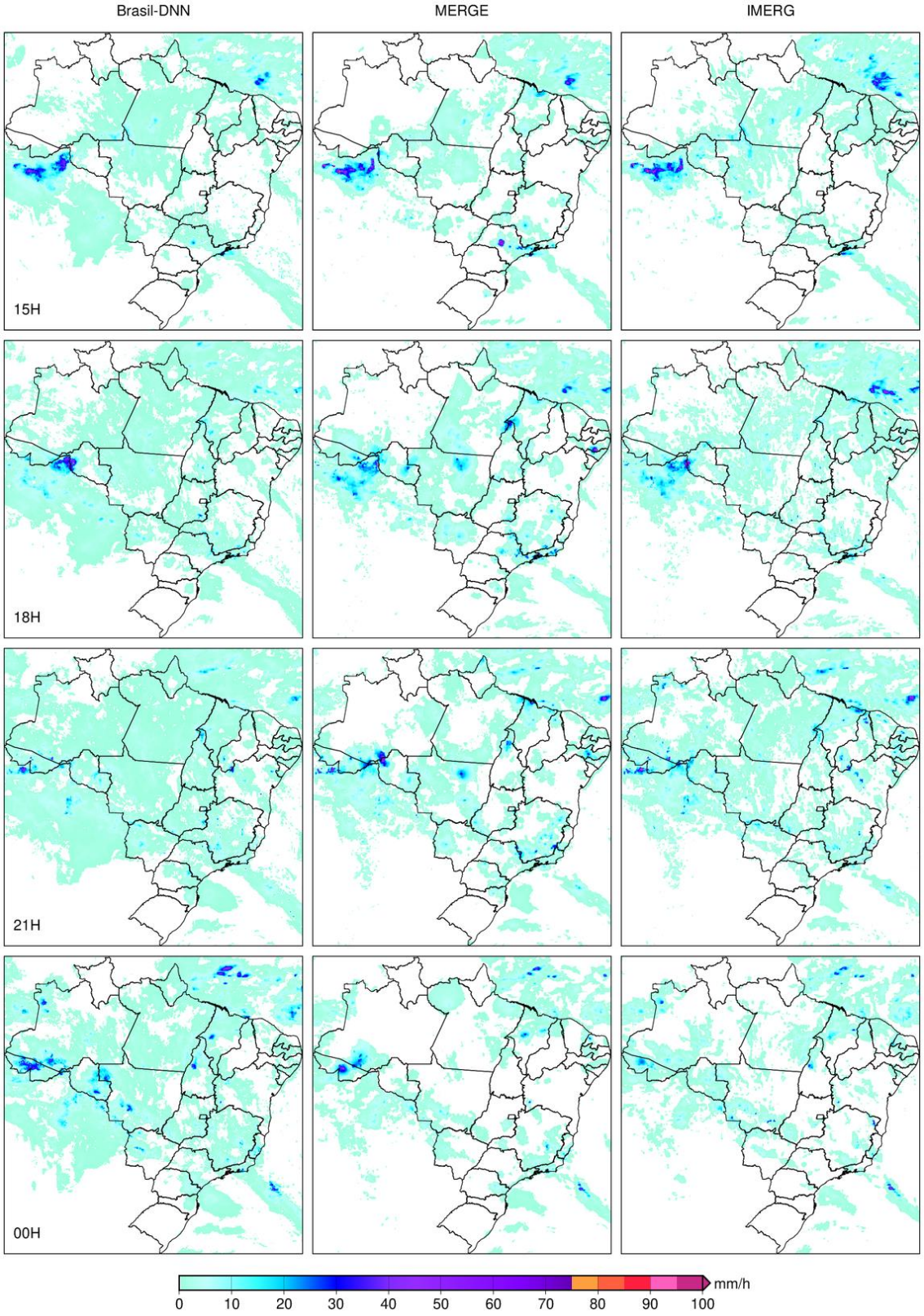
Para simulação foi utilizado uma data fora do período utilizado para treinamento, segundo o INMET o mês de fevereiro de 2020 foi o mês mais chuvoso para a cidade de São Paulo levando em consideração os 77 anos de dados históricos do instituto, assim como a cidade do Rio de Janeiro em comparação aos últimos 24 anos, e várias outras cidades superaram a média histórica neste mês. Estas chuvas levaram aproximadamente 200 munícipios a decretarem situação de emergência. O dia 22 de fevereiro foi o dia do mês que registrou o maior acumulado em 24 horas, com 146,8 mm em Iguapé-SP, além

de registrar valores acima de 45 mm em 11 dos 26 estados brasileiros, e por essa razão foi o dia escolhido para simulação.

As Figuras 5.32 e 5.33 mostram as saídas da execução da DNN-Brasil para o dia 22 de fevereiro de 2020 comparado às saídas correspondentes do produto MERGE e IMERG. Note que essas figuras ilustram apenas as feições dos três dados, de modo a demostrar espacialmente como as chuvas entre eles podem ser observadas. Nota-se uma certa semelhança entre as três estimativas, onde o DNN-Brasil mostra-se por vezes mais próximas do MERGE e outras do IMERG, enquanto essas duas estimativas exibem alguns valores altos de precipitação em alguns pontos isolados que apresentam prováveis erros em dados observacionais ou de estimativa por satélite. Como exemplo, podemos citar um ponto isolado no centro do estado de São Paulo de valores altos de chuva no MERGE às 15H na Figura 5.32, que provavelmente corresponde a uma estação com problemas em suas medidas, cuja a DNN conseguiu diminuir seus efeitos. Já na Figura 5.33 é possível ver o excesso de chuva gerado pelo IMERG sobre o oeste do estado do Maranhão às 03 GMT que é melhor representado pela DNN. De todo a DNN subestimou alguns eventos de forte chuva em horários que geralmente não são convectivos, como as chuvas que ocorreram em 06 GMT do dia 22 de fevereiro de 2020 (Figura 5.33).

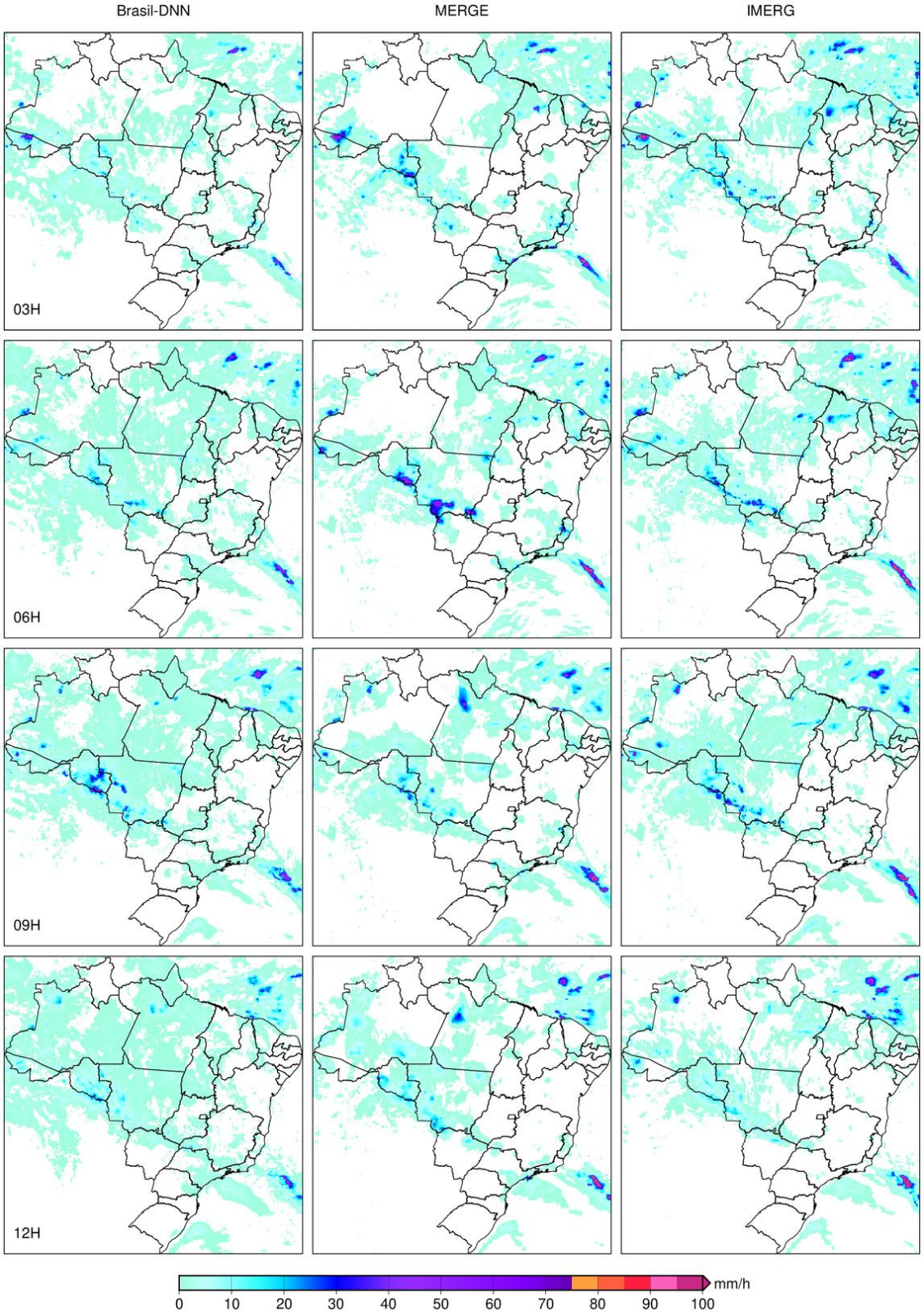
Pouca informação objetiva é possível de ser coletada sem uma verdade em superfície, ou seja, um mapa das observações apenas. Contudo, nos locais onde foram registradas medidas observacionais foi possível calcular a correlação e o erro de cada uma das estimativas em relação aos dados observados e compará-las. O produto MERGE, nos locais onde há observação e nos vizinhos ao redor, tem seus valores corrigidos pela observação, por isso, quando correlacionados apresenta resultados muito altos, entre 85 e 95%. Contudo, este percentual não é válido para toda grade, os locais descobertos por pluviômetros no produto MERGE tem as medidas equivalentes ao IMERG, diferentemente da RNA onde a correção é realizada em toda a grade independente de dados observados.

Figura 5.32: Acumulado 3h de precipitação entre os dias 21 e 22 de fevereiro de 2020 entre 15 e 00 GMT.




Fonte: Produção do Autor.

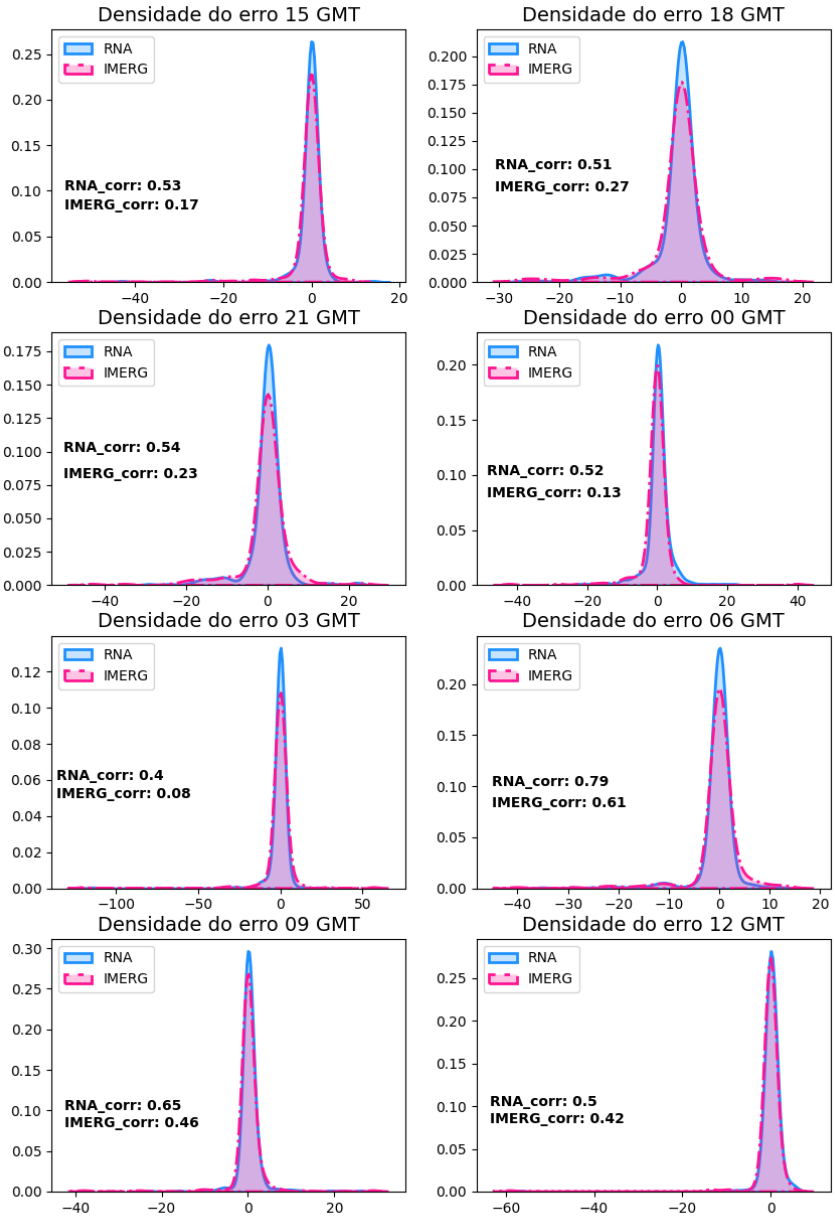
Figura 5.33: Acumulado 3h de precipitação para o dia 22 de fevereiro de 2020 entre 03 e 12 GMT.




Fonte: Produção do Autor.

O resultado da correlação entre a RNA e o IMERG com os dados observados, por horário, é apresentado na Figura 5.34, em todos os horários a estimativa provinda da DNN-Brasil apresentou maior correlação que o IMERG, sendo em média 25,87% maior, além disso, na distribuição do erro é possível observar a gaussiana centrada em zero em ambas às estimativas, porém em todos os horários a RNA também se mostrou superior.

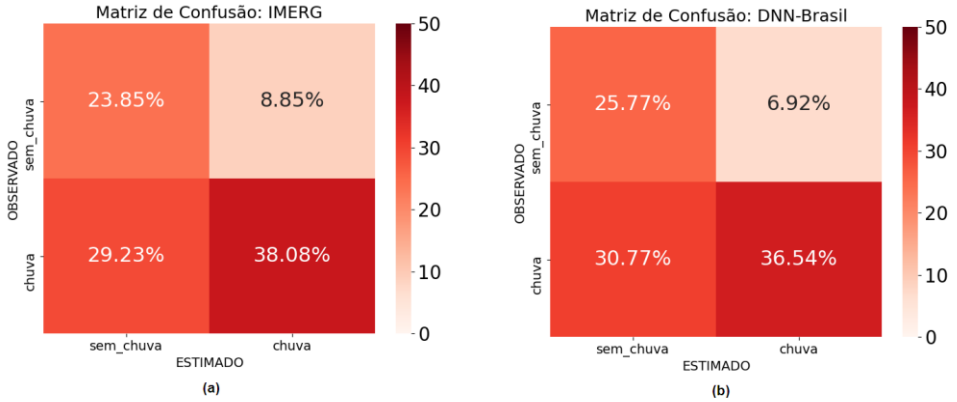
Figura 5.34: Comparativo de desempenho dos estimadores RNA (DNN-Brasil) e IMERG para o dia 22 de fevereiro de 2020.




Fonte: Produção do Autor.

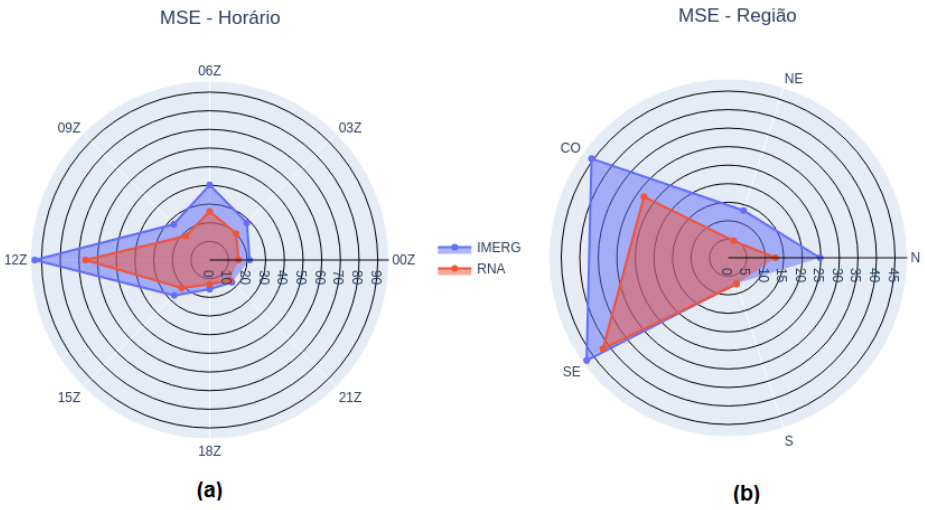
A Figura 5.35, mostra que ambos os estimadores variaram muito pouco quanto a assertividade para o dia em questão, porém na Figura 5.36 é apresentado os comparativos entre o MSE, dos estimadores IMERG e da RNA, por horário e por região. Em todos os horários a RNA apresentou MSE menor que o IMERG. Quanto às regiões a RNA apresentou valores menores com exceção da região Sul onde o IMERG apresentou MSE de 7,31 e a RNA apresentou MSE de 7,46.

Figura 5.35: Comparativo da acurácia dos estimadores RNA (DNN-Brasil) e IMERG para o dia 22 de fevereiro de 2020.




Fonte: Produção do Autor.

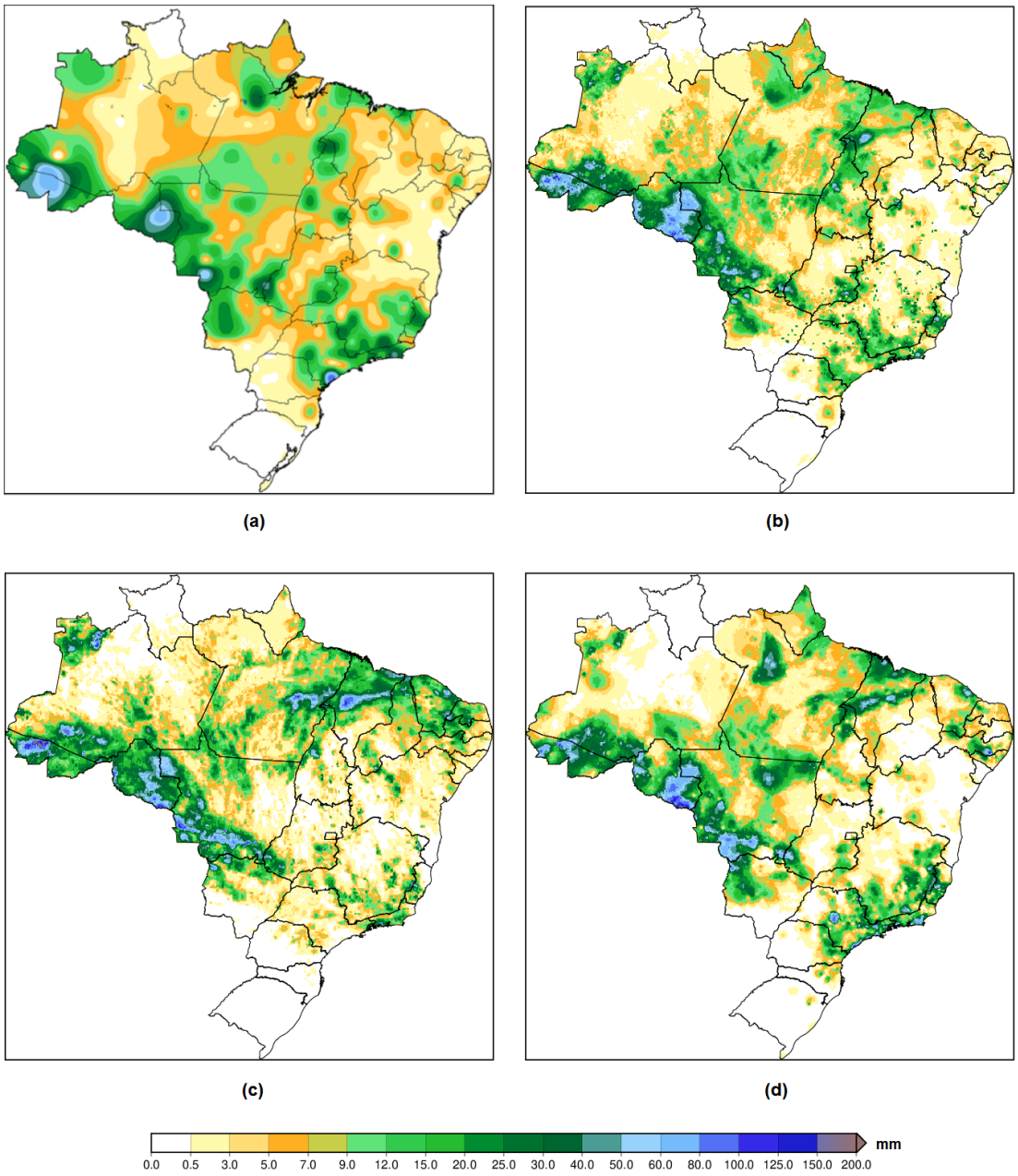
Figura 5.36: Comparativo do MSE dos estimadores RNA (DNN-Brasil) e IMERG para o dia 22 de fevereiro de 2020.




Fonte: Produção do Autor.

A Figura 5.37, mostra o resultado do acumulado diário de cada uma das estimativas para o dia em questão, comparada ao produto observado diário extrapolado para uma grade regular pelo INMET.

Figura 5.37: Comparativo dos diferentes produtos diários de precipitação para o dia 22 de fevereiro de 2020.




Onde (a) é o acumulado diário observado disponibilizado pelo INMET, (b) é o resultado da execução da RNA (DNN-Brasil), (c) é o acumulado diário do IMERG e (d) é o acumulado diário do MERGE sub-diário.

Fonte: Produção do Autor.

No produto diário é possível ver anomalias discrepantes nas feições das estimativas em relação ao produto observado, principalmente entre o acumulado do IMERG, onde é possível observar a ausência de sistemas em vários locais incluindo São Paulo, onde no dia em questão foi registrada precipitação que ultrapassou 140 mm. Assim como no norte do Pará, onde um grande núcleo de precipitação pode ser observado. Quanto ao acumulado dos produtos sub-diários da RNA também é possível observar anomalias positivas e negativas em relação ao produto alvo, porém existe muitas similaridades entre as feições dos dois produtos e considerando que o produto observado também é uma extrapolação é possível afirmar que o produto diário da RNA pode ser satisfatório para simulação da realidade. Contudo, ainda é possível ver algum resquício de pontos isolados de precipitação, aparentemente não realísticos, sobre o sudeste do Brasil que podem estar associados a necessidade de definir um limiar mínimo de chuva mais assertivo.

# 6 CONCLUSÃO

Conforme o resultado obtido nesta pesquisa é possível afirmar que as redes neurais artificiais são uma alternativa plausível para o *downscaling* de dados de precipitação diária para nosso país. Neste estudo foram avaliados dois tipos de redes neurais, para resolver o problema em questão, as redes neurais profundas e as redes neurais recorrentes, e ambas se mostraram promissoras obtendo erros inferiores aos de metodologias muito utilizadas atualmente como o IMERG da NASA.

Para a ponderação dos valores foram selecionadas variáveis meteorológicas que apresentavam indicativos de correlação de acordo com a literatura, estas foram avaliadas quanto à correlação síncrona comparando-as com a precipitação sub-diária, objeto de interesse deste estudo, e submetidas individualmente ao treinamento da RNA. Além disso, foi verificado a partir da análise da performance se há melhorias no desempenho causado por correlações assíncronas entre as variáveis intrínsecas ao sistema.

Os resultados foram avaliados quanto ao horário, estação do ano, e região do país. Ambos os modelos propostos apresentaram resultados superiores ao modelo IMERG, para quase todos os experimentos. Concluiu-se que a rede profunda para todo o território brasileiro, aqui chamada de DNN-Brasil, foi aquela que mostrou o melhor resultado.

A metodologia proposta além de ser uma técnica computacionalmente mais econômica que a utilizada para produzir estimativas através de modelos dinâmicos, que por vezes necessitam de supercomputadores, além de menos complexa, é também uma alternativa para minimizar a deficiência da rede pluviométrica automática esparsa do Brasil, além dos gastos com equipamentos e manutenção. Uma vez que, além de produzir dados para todo o território brasileiro, usando informações de satélites e modelos numéricos, também foi capaz de inferir medidas com precisão próxima a das observações.

Por fim a RNA proposta neste estudo, DNN-Brasil, foi aplicada em um caso de uso real (como em um sistema de monitoramento operacional) e reproduziu resultados satisfatórios, como aqueles encontrados no estudo.

Cabe ressaltar que, apesar do parecer positivo, esta é uma primeira abordagem e estudos mais profundos sobre alguns aspectos ainda precisam ser mais bem elaborados. Sugere-se que em estudos futuros: balancear os dados para diferentes regiões e períodos, de modo a diminuir alguns efeitos locais de regiões com maior densidade de pluviômetros; implementar técnicas mais objetivas de definição dos hiperparâmetros das redes neurais; implementar novas variáveis que possam melhorar as estimativas da chuva; testar o uso de dados de modelos numéricos de previsão de tempo em mais alta resolução; definir ainda mais controles de qualidade dos dados de entrada da rede; ampliar o número de dados de estações de superfície através de outras redes; balancear os dados talvez incluindo amostras de chuvas intensas provenientes de radares; entre outros aspectos. Além disso, notou-se que anos como de 2020, onde os aspectos climáticos foram adversos, será necessário um melhor aprofundamento do sistema de aprendizado da técnica para melhor representar extremos.

## REFERÊNCIAS BIBLIOGRÁFICAS


AHRENS, C. D. **Meteorology today:** an introduction to weather, climate, and the environment. 19. ed. Belmont, CA: Brooks/Cole, 2009.

ALMAZROUI, M. Calibration of TRMM rainfall climatology over Saudi Arabia during 1998–2009. **Atmospheric Research**, v.99, n.3/4, p.400-414, 2011.

ALMEIDA, C. T. et al. Avaliação das estimativas de precipitação do produto 3B43-TRMM do estado do Amazonas. **Floresta e Ambiente**, v.22, n.3, p.279-286, 2015. Disponível em: http://dx.doi.org/10.1590/2179-8087.112114.

AGÊNCIA NACIONAL DE ÁGUAS - ANA. **Diretrizes e análises recomendadas para consistência de dados pluviométricos**. Brasília: ANA, 2012, p. 10-13. Disponível em: https://arquivos.ana.gov.br/infohidrologicas/cadastro/DiretrizesEAnalisesRecomendadasParaConsistenciaDeDadosPluviometricos-VersaoJan12.pdf

AGÊNCIA NACIONAL DE ÁGUAS - ANA. **Conjuntura dos recursos hídricos no Brasil 2020**. Brasília: ANA, 2020, p. 88-89. Disponível em: http://conjuntura.ana.gov.br/static/media/conjuntura-completo.23309814.pdf.

BELLERBY, T. et al. Rainfall estimation from a combination of TRMM precipitation radar and GOES multispectral satellite imagery through the use of an artificial neural network. **Journal of Applied Meteorology**, v.39, n.12, p.2115-2128, 2000.

BRITO, S. S. B.; OYAMA, M. D. Daily cycle of precipitation over the northern coast of Brazil. **Journal of Applied Meteorology and Climatology**, v.53, n.11, p. 2481–2502, 2014.

BRUHN, J. A.; FRY, W. E.; FICK, G. W. Simulation of daily weather data using theoretical probability distributions. **Journal of Applied Meteorology**, v.19, n.9, p. 1029-1036, 1980.

CALHEIROS, A. J. P. **Propriedades radiativas e microfísicas das nuvens continentais: uma contribuição para a estimativa de precipitação de nuvens quentes por satélite**. Tese (Doutorado em Meteorologia) - Instituto Nacional de Pesquisas Espaciais. São José dos Campos, 2013. Disponível em: http://urlib.net/8JMKD3MGP7W/3EUU6BS.

CAMARO, E.; DRUNCK, S.; CÂMARA, G. Análise Espacial de Superfícies. In: EMBRAPA (Org.). **Análise espacial de dados geográficos**. Planaltina: [s.n.], 2004. ISBN ISBN: 85-7383-260-6. Disponível em: http://www.dpi.inpe.br/gilberto/livro/analise/cap3-superficies.pdf.

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.1/4, p. 136-151, 2002.

CERON, W. et al. Community detection in very high-resolution meteorological metworks. **IEEE Geoscience and Remote Sensing Letters**, p.2007-2010, 2019.

CHIANG, Y.-M. et al. Dynamic ANN for precipitation estimation and forecasting from radar observations. **Journal of Hydrology**, v.334, n.1/2, p.250-261, 2007.

COHEN, J. C. P.; DIAS, M. A. S.; NOBRE, C. A. Environmental conditions associated with amazonian squall lines: a case study. **Monthly Weather Review**, v.123, n.11, p. 3163-3174, 1995.

DAI, A.; GIORGI, F.; TRENBERTH, K. E. Observed and model-simulated diurnal cycles of precipitation over the contiguous United States. **Journal of Geophysical Research: Atmospheres**, v.104, n.D6, p.6377-6402, 1999.

DANELICHEN, V. H. D. M. et al. TRMM satellite performance in estimated rainfall over the midwest region of Brazil. **Revista Brasileira de Climatologia**, v.12, p.22-31, 2013. ISSN ISSN:1980-055x.

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS. DIVISÃO DE SATÉLITES E SENSORES METEOROLÓGICOS. **Órbitas**. Disponivel em: http://satelite.cptec.inpe.br/informacao/orbitas.jsp.

GABRIEL, K. R.; NEUMANN, J. A Markov chain model for daily rainfall occurrence at Tel Aviv. **Quarterly Journal of Royal Meteorological Society**, v.88, n.375, p.90-95, 1962.

GADELHA, A. N. et al. Grid box-level evaluation of IMERG over Brazil at various space and time scales. **Atmospheric Research**, p.231-244, 2019.

GARCIA, S. R.; CALHEIROS, A. J. P.; KAYANO, M. T. Revised method to detect the onset and demise dates of the rainy season in the South American Monsoon areas. **Theoretical and Applied Climatology**, v.126, n.3/4, p.481-491, 2015.

GLOROT, X.; BENGIO, Y. Proceedings of the thirteenth international conference on artificial intelligence and statistics. **JMLR Workshop and Conference Proceedings**, p.249-256, 2010.

GRIGOLETTO, J. C. et. al. Gestão das ações do setor saúde em situações de seca e estiagem. **Ciência e Saúde Coletiva**, Rio de Janeiro, v. 21, n. 3, p. 709-718, 2016. ISSN ISSN 1678-4561.

GUILLORY, A. ERA5: How to calculate wind speed and wind direction from u and v components of the wind? **Copernicus Knowledge Base**, 2020. Disponivel em: https://confluence.ecmwf.int/pages/viewpage.action?pageId=133262398. Acesso em: 14 jan. 2021.

HAYKIN, S. **Redes neurais, princípios e práticas**. 2.ed. [S.l.]: bookman, 2001.

HE, K. et al. Delving deep into rectifiers: surpassing human-Level performance on imagenet classification. In: INTERNATIONAL CONFERENCE ON COMPUTER VISION (ICCV), 2015. **Proceedings…** 2015.

HE, X. et al. Spatial downscaling of precipitation using adaptable random forests. **Water Resources Research**, v.52, n.10, p. 8217-8237, 2016.

HERSBACH, H. et al. ERA5 hourly data on pressure levels from 1979 to present. **Copernicus Climate Change Service (C3S) Climate Data Store (CDS)**, 2018. ISSN 10.24381/cds.bd0915c6. Disponivel em: https://cds.climate.copernicus.eu/cdsapp#!/dataset/10.24381/cds.bd0915c6?tab=overview. Acesso em: 14 jan. 2021.

HEWITSON, B. C.; CRANE, R. G. Large-scale atmospheric controls on local precipitation in tropical Mexico. **Geophysical Research Letters**, v.19, n.18, p.1835-1838, 1992.

HEWITSON, B. C.; CRANE, R. G. Climate downscaling: techniques and application. **Climate Research**, v.7, p.85-95, 1996.

HUFFMAN, G. J. et al. The TRMM Multisatellite Precipitation Analysis (TMPA): quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. **Journal of Hydrometeorology**, v.8, n.1, p.38-55, 2007.

HUFFMAN, G. J. et al. GPM IMERG early precipitation L3 half hourly 0.1 degree x 0.1 degree V06. **Goddard Earth Sciences Data and Information Services Center (GES DISC)**, 2019. ISSN 10.5067/GPM/IMERG/3B-HH-E/06. Disponivel em: https://disc.gsfc.nasa.gov/datasets/GPM_3IMERGHHE_06/summary. Acesso em: 14 jan. 2021.

HUFFMAN, G. J. et al. Integrated Multi-Satellite Retrievals for the Global Precipitation Measurement (GPM) Mission (IMERG). **Advances in Global Change Research**, p.343-353, 2020.

HUNTER, D. et al. Selection of proper neural network sizes and architectures: a comparative study. **Proceedings of the IEEE Transactions on Industrial Informatics**, v. 8, n. 2, p. 228-240, 2012.

IBARRA-BERASTEGI, G. et al. Downscaling of surface moisture flux and precipitation in the Ebro Valley (Spain) using analogues and analogues followed by random forests and multiple linear regression. **Hydrology and Earth System Sciences Discussions**, v.15, n.6, p.1895-1907, 2011.

INSTITUTO NACIONAL DE METEOROLOGIA - INMET. Histórico de dados Meteorológicos. **Instituto Nacional de Meteorologia**, 2020. Disponivel em: <https://portal.inmet.gov.br/dadoshistoricos>. Acesso em: 14 jan. 2021.

JANOWIAK, J.; JOYCE, B.; XIE, P. NCEP/CPC L3 half hourly 4km global (60S - 60N) merged IR V1. **Goddard Earth Sciences Data and Information Services Center (GES DISC)**, 2017. ISSN 10.5067/P4HZB9N27EKU. Disponivel em: <https://disc.gsfc.nasa.gov/datasets/GPM_MERGIR_1/summary>. Acesso em: 14 jan. 2021.

JET PROPULSION LABORATORY - JPL. Releases enhanced shuttle land elevation data. **Jet Propulsion Laboratory (NASA)**, 2021. Disponivel em: <https://www2.jpl.nasa.gov/srtm/>. Acesso em: 14 jan. 2021.

KHAN, M. S.; COULIBALY, P.; DIBIKE, Y. Uncertainty analysis of statistical downscaling methods. **Journal of Hydrology**, v.319, n.1/4, p.357-382, 2006.

KIDDER, S.; HAAR, T. V. Meteorological satellite instrumentation. In: KIDDER, S.; HAAR, T. V. (Ed.). **Satellite meteorology, an introduction**, San Diego: Academic Press, 1995. p.87-144. Disponível em: https://doi.org/10.1016/b978-0-08-057200-0.50008-0.

KIKUCHI, K.; WANG, B. Diurnal precipitation regimes in the global tropics. **Journal of Climate**, v.21, n.11, p.2680-2696, 2008.

KULIGOWSKI, R. J.; BARROS, A. P. Experiments in short-term precipitation forecasting using artificial neural networks. **Monthly Weather Review**, v.126, n.2, p.470-482, 1998.

KUMAR, J. et al. Sub-daily statistical downscaling of meteorological variables using neural networks. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL SCIENCE, 2012. **Proceesings…** ICCS 2012, p.887-896.

KUMMEROW, C. et al. The status of the Tropical Rainfall Measuring Mission (TRMM) after two years in orbit. **Journal of Applied Meteorology**, v.39, n.12, p. 1965-1982, 2000.

LARSEN, G. A.; PENSE, R. B. Stochastic simulation of daily climate data for agronomic models. **Agronomy Journal**, v.74, n.3, p. 510-514, 1982.

LI, J. Y.; CHOW, T. W. S.; YU, Y. L. Estimation theory and optimization algorithm for the number of hidden units in the higher-order feedforward neural network. In: INTERNATIONAL CONFERENCE ON NEURAL NETWORKS, 1995. **Proceedings…** 1995. p.1229-1233.

LI, M.; SHAO, Q. An improved statistical approach to merge satellite rainfall estimates and raingauge data. **Journal of Hydrology**, v.385, n.1/4, p. 51-64, 2010.

MACHADO, L. A. T. et al. Overview: precipitation characteristics and sensitivities to environmental conditions during GoAmazon2014/5 and ACRIDICON-CHUVA. **Atmospheric Chemistry and Physics**, v.18, p.6461-6482, 2018.

MARENGO, J. A. et al. A seca e a crise hídrica de 2014-2015 em São Paulo. **Revista USP**, São Paulo, v. 106, p. 31-44, 02 set. 2015. ISSN http://doi.org/10.11606/issn.2316-9036.v0i106p31-44.

MCCARTHY, J. et al. A Proposal for the dartmouth summer research project on artificial intelligence, August 31, 1955. **AI Magazine**, v.27, 2006. ISSN https://doi.org/10.1609/aimag.v27i4.1904.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **The Bulletin of Mathematical Biophysics**, v.5, n.4, p.115-133, 1943.

MELO, D. D. C. D. et al. Performance evaluation of rainfall estimates by TRMM Multi-satellite Precipitation Analysis 3B42V6 and V7 over Brazil. **Journal of Geophysical Research: Atmospheres**, v.120, n.18, p.9426-9436, 2015.

MICHELSON, D. et al. BALTRAD Advanced weather radar networking. **Journal of Open Research Software**, v.12, 2018. Disponivel em: <http://git.baltrad.eu/trac/wiki/cookbook/Anomaly-detection-applying-non-polarimetric-pattern-recognition>. Acesso em: 08 jan. 2021.

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION - NASA. **Algorithm Theoretical Basis Document (ATBD)**. Global Precipitation Measurement (GPM) National Aeronautics and Space Administration (NASA). [S.l.], v6, p. 9, 2019. Disponível em: https://gpm.nasa.gov/sites/default/files/document_files/IMERG_ATBD_V06.pdf.

NICKS, A. D.; HARP, J. F. Stochastic generation of temperature and solar radiation data. **Journal of Hydrology**, v.48, n.1/2, p.1-17, 1980.

OKI, T.; MUSIAKE, K. Seasonal change of the diurnal cycle of precipitation over Japan and Malaysia. **Journal of Applied Meteorology**, v.33, n.12, p.1445-1463, 1994.

OLIVEIRA, R. et al. Characteristics and diurnal cycle of GPM rainfall estimates over the Central Amazon Region. **Remote Sensing**, v.8, n.7, 2016.

OLIVEIRA, V. P. S.; ZANETTI, S. S.; PRUSKI, F. F. CLIMABR parte I: modelo para a geração de séries sintéticas de precipitação. **Revista Brasileira de Engenharia Agrícola e Ambiental**, Campina Grande, PB, v.9, n.3, p.348-355, 2005.

PEREIRA, G. et al. Avaliação dos dados de precipitação estimados pelo satélite TRMM para o Brasil. **Revista Brasileira de Recursos Hídricos**, v.18, n.3, p. 139-148, 2013. ISSN DOI: 10.21168/rbrh.v18n3.p139-148.

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.

RICHARDSON, C. W.; WRIGHT, D. A. **Wgen:** a model for generating daily weather variables. [S.l.]: U.S. Department of Agriculture, Agricultural Research Service, 1984.

RODGERS, E. et al. A statistical technique for determining rainfall over land employing Nimbus 6 ESMR measurements. **Journal of Applied Meteorology**, v.18, n.8, p. 978-991, 1979.

ROZANTE, J. R. et al. Combining TRMM and surface observations of precipitation: technique and validation over South America. **Weather and Forecasting**, v.25, n.3, p. 885-894, 2010.

ROZANTE, J. R. et al. Produto de precipitação MERGE. **Centro de Previsão do Tempo e Estudos Climáticos (CPTEC)**, 2020. Disponivel em: <http://ftp.cptec.inpe.br/modelos/tempo/MERGE/GPM/>. Acesso em: 14 jan. 2021.

SANTOS E SILVA, C. M. Ciclo diário e semidiário de precipitação na costa Norte do Brasil. **Revista Brasileira de Meteorologia**, v.28, n.1, p. 34-42, 2013.

SANTOS E SILVA, C. M.; FREITAS, S. R.; GELOW, R. Ciclo diário da precipitação estimada através de um radar banda S e pelo algoritmo 3B42_V6 do projeto TRMM durante a estação chuvosa de 1999 no sudoeste da Amazônia. **Revista Brasileira de Meteorologia**, v.26, n.1, p. 95-108, 2011.

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.

SCOFIELD, R. A.; KULIGOWSKI, R. J. Status and outlook of operational satellite precipitation algorithms for extreme-precipitation events. **Weather and Forecasting**, v.18, n.6, p. 1037-1051, 2003.

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.

SHEELA, K. G.; DEEPA, S. N. Review on methods to fix number of hidden neurons in neural networks. **Mathematical Problems in Engineering**, 2013. DOI: 10.1155/2013/425740

SHI, Y.; SONG, L. Spatial downscaling of monthly TRMM precipitation based on EVI and other geospatial variables over the Tibetan plateau from 2001 to 2012. **Mountain Research and Development**, v. 35, p. 180-194, 2015. ISSN https://doi.org/10.1659/mrd-journal-d-14-00119.1.

SHIBATA, K.; IKEDA, Y. Effect of number of hidden neurons on learning in large-scale layered neural networks. In: ICCAS-SICE INTERNATIONAL JOINT CONFERENCE, 2009. **Proceedings…** 2009. p.2008-5013.

SOARES, A. S. D.; DA PAZ, A. R.; PICCILLI, D. G. A. Avaliação das estimativas de chuva do satélite TRMM no Estado da Paraíba. **Revista Brasileira de Recursos Hídricos**, v. 21, p. 288-299, 2016. ISSN DOI: 10.21168/rbrh.v21n2.p288-299.

TAMURA, S. I.; TATEISHI, M. Capabilities of a four-layered feedforward neural network: four layers versus three. **IEEE Transactions on Neural Networks**, v. 8, n. 2, p. 251-255, 1997. ISSN DOI: 10.1109/72.557662.

TAPIADOR, F. J. et al. A neural networks--based fusion technique to estimate half-hourly rainfall estimates at 0.1 resolution from satellite passive microwave and infrared data. **Journal of Applied Meteorology**, v.43, n.4, p. 576-594, 2004.

TOHMA, S.; IGATA, S. Rainfall estimation from GMS imagery data using neural network. **WIT Transactions on Ecology and the Environment**, p. 123-130, 1994.

TUCCI, C. E. M. **Hidrologia:** ciência e aplicação. 3. ed. Porto Alegre: Editora da UFRGS: ABRH, v. 4, 2004.

VASILOFF, V. et al. Improving QPE and very short term QPF: an initiative for a community-wide integrated approach. **Bulletin of the American Meteorological Society**, v. 88, p. 1899-1911, 2007. ISSN https://doi.org/10.1175/BAMS-88-12-1899.

VICENTE, G. A.; SCOFIELD, R. A.; MENZEL, W. P. The operational GOES infrared rainfall estimation technique. **Bulletin of the American Meteorological Society**, v.79, n.9, p. 1883-1898, 1998.

VILA, D. A. et al. Statistical evaluation of combined daily gauge observations and rainfall satellite estimates over continental South America. **Journal of Hydrometeorology**, v.10, n.2, p. 533-543, 2009.

VIRGENS FILHO, J. S. et al. PGECLIMA_R: gerador estocástico para simulação de cenários climáticos Brasileiros. I - desenvolvimento do gerenciador do banco de dados climáticos. In: CONGRESSO BRASILEIRO DE AGROMETEOROLOGIA, 17., 2011, Guarapari. **Anais...** 2011. p. 1-5.

VUJIčIć, T. et al. Comparative analysis of methods for determining number of hidden neurons in artificial neural network. In: CENTRAL EUROPEAN CONFERENCE ON INFORMATION AND INTELLIGENT SYSTEMS, 2016, Varaždin, Croatia. **Proceedings…** 2016. p. 219-250.

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.

WILHEIT, T. T. et al. A Satellite technique for quantitatively mapping rainfall rates over the oceans. **Journal of Applied Meteorology**, v.16, n.5, p. 551-560, 1977.

WILLIAMS, J. R.; NICKS, A. D.; ARNOLD, J. G. Simulator for water resources in rural basins. **Journal of Hydraulic Engineering**, v.111, n.6, p. 970-986, 1985.

WU, L.; XU, Y.; WANG, S. Comparison of TMPA-3B42RT legacy product and the equivalent IMERG products over Mainland China. **Remote Sensing**, v.10, n.11, 2018.

XIAO, R.; CHANDRASEKAR, V. Development of a neural network based algorithm for rainfall estimation from radar observations. **IEEE Transactions on Geoscience and Remote Sensing**, v.35, n.1, p. 160-171, 1997.

XU, S.; CHEN, L. A novel approach for determining the optimal number of hidden layer neurons for FNN's and its application in data mining. In:

INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY AND APPLICATIONS, 2008. **Proceedings…** 2008. p. 683-686.

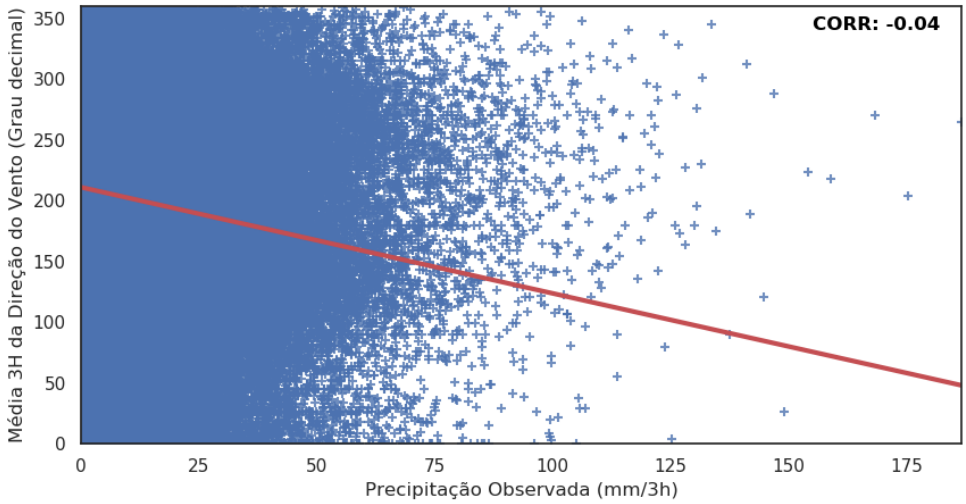
YANG, S.; SMITH, E. A. Convective–stratiform precipitation variability at seasonal scale from 8 Yr of TRMM observations: implications for multiple modes of diurnal variability. **Journal of Climate**, v.21, n.16, p. 4087-4114, 2008.

YOTOV, K.; HADZHIKOLEV, E.; HADZHIKOLEVA, S. Determining the number of neurons in artificial neural networks for approximation, trained with algorithms using the Jacobi matrix. **TEM Journal**, v. 9, n. 4, p. 1320-1329, 2020. ISSN ISSN 2217-8309, DOI: 10.18421/TEM94-02.

ZHOU, L. et al. Daily rainfall model to merge TRMM and ground based observations for rainfall estimations. In: INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM, 2016. **Proceedings...** IEEE, 2016. p. 601-604.

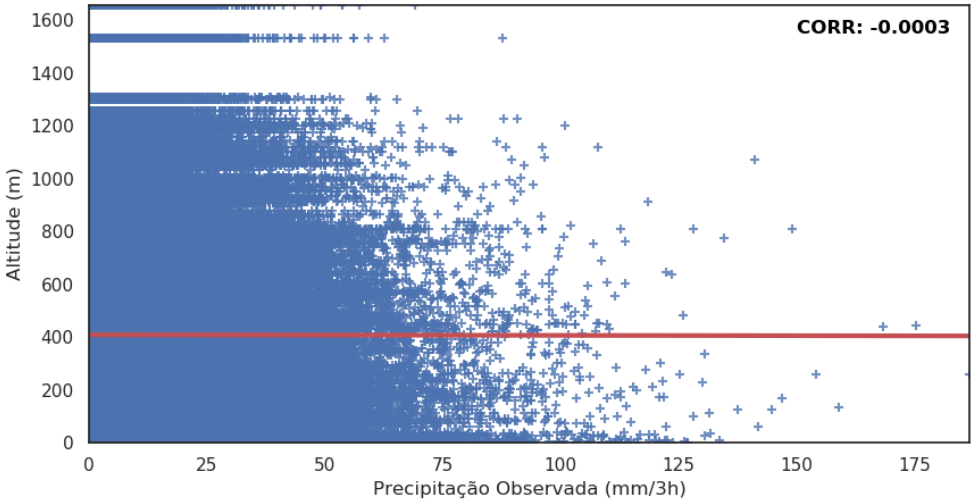
## APÊNDICE A – VARIÁVEIS QUE APRESENTARAM INFORMAÇÕES POUCO RELEVANTES PARA O ESTUDO

Figura A1: Comparativo entre a média 3 horas da direção do vento em 850 hPa e o acumulado de precipitação observada no mesmo período.



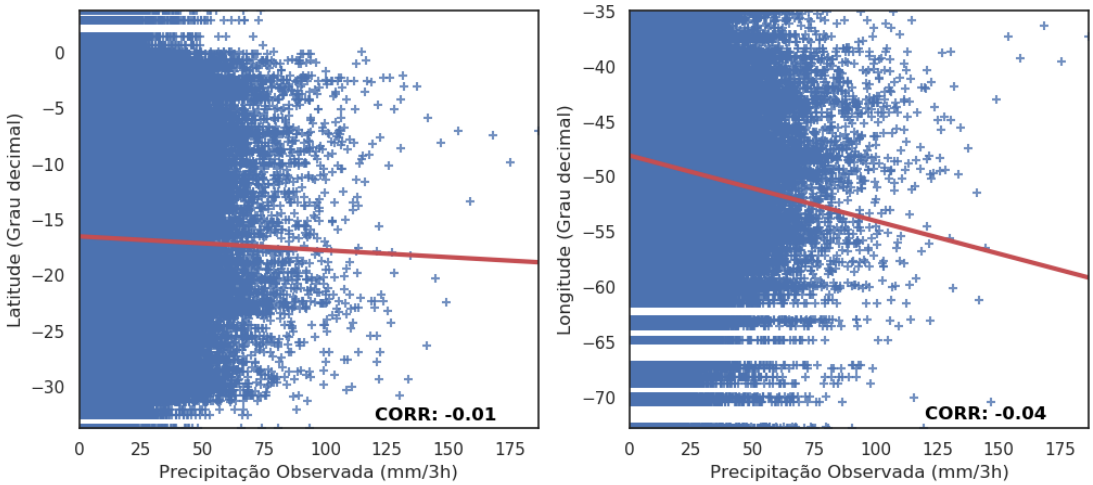

Fonte: Produção do Autor.

Figura A2: Comparativo entre a orografia e o acumulado de precipitação observada.



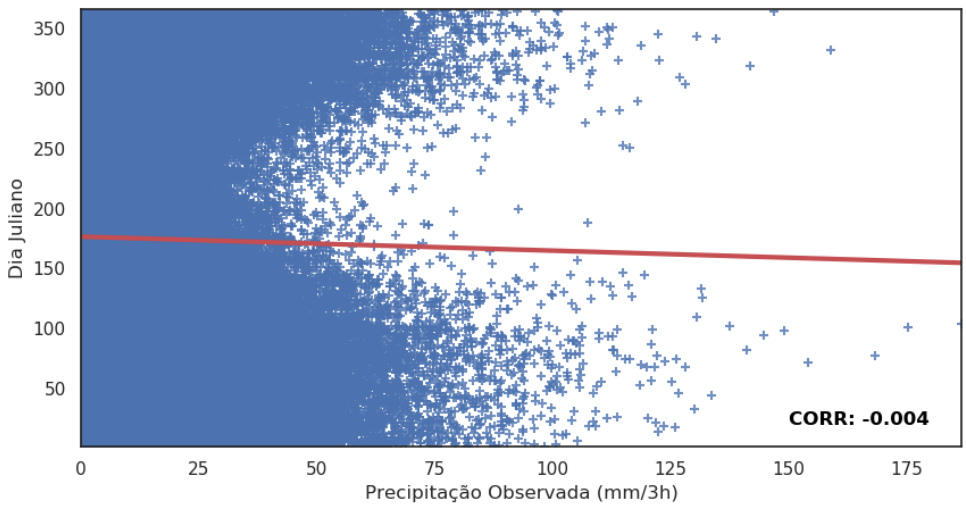

Fonte: Produção do Autor.

Figura A3: Comparativo entre a localização geográfica e o acumulado de precipitação observada.




Fonte: Produção do Autor.

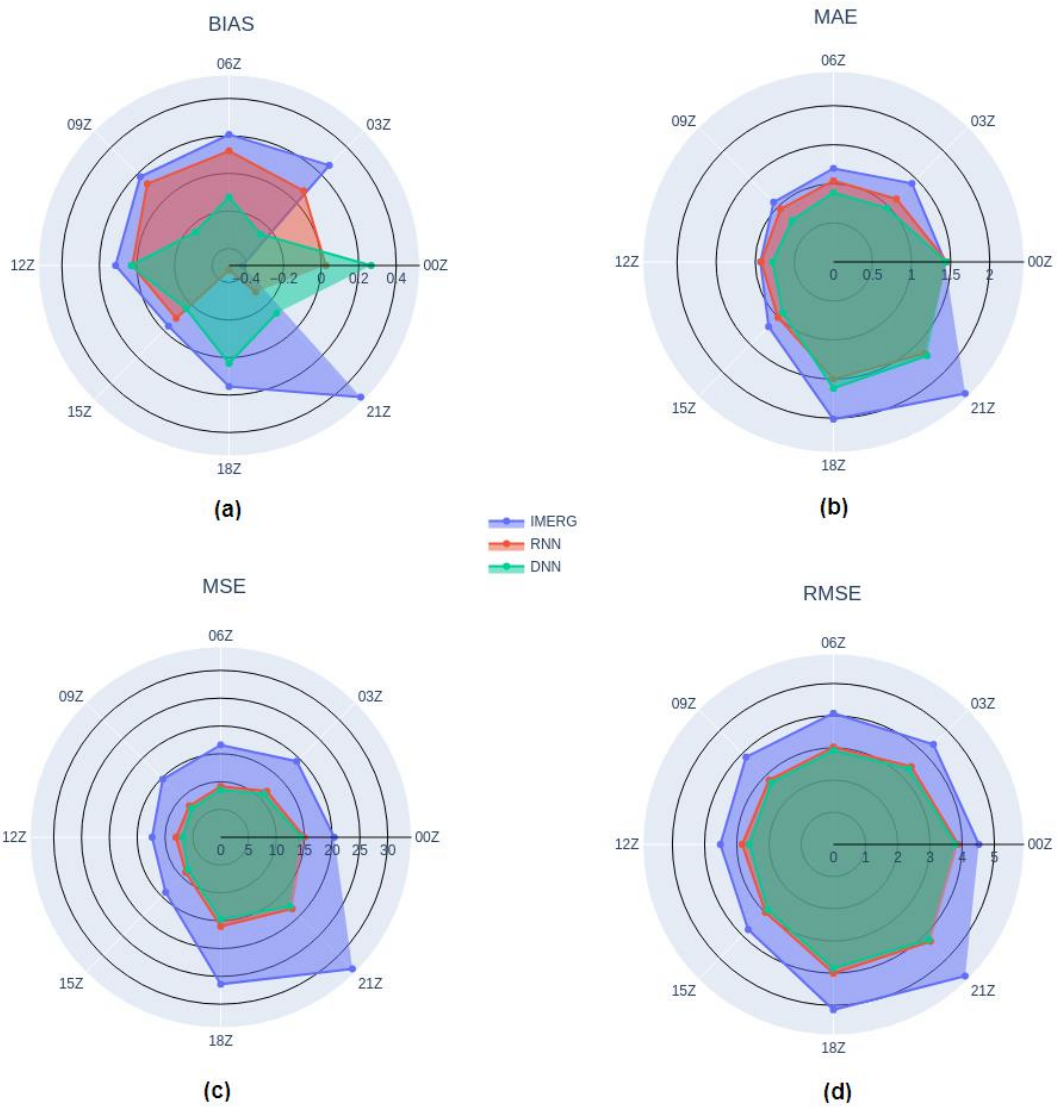
Figura A4: Comparativo entre o dia juliano e o acumulado de precipitação observada.




Fonte: Produção do Autor.

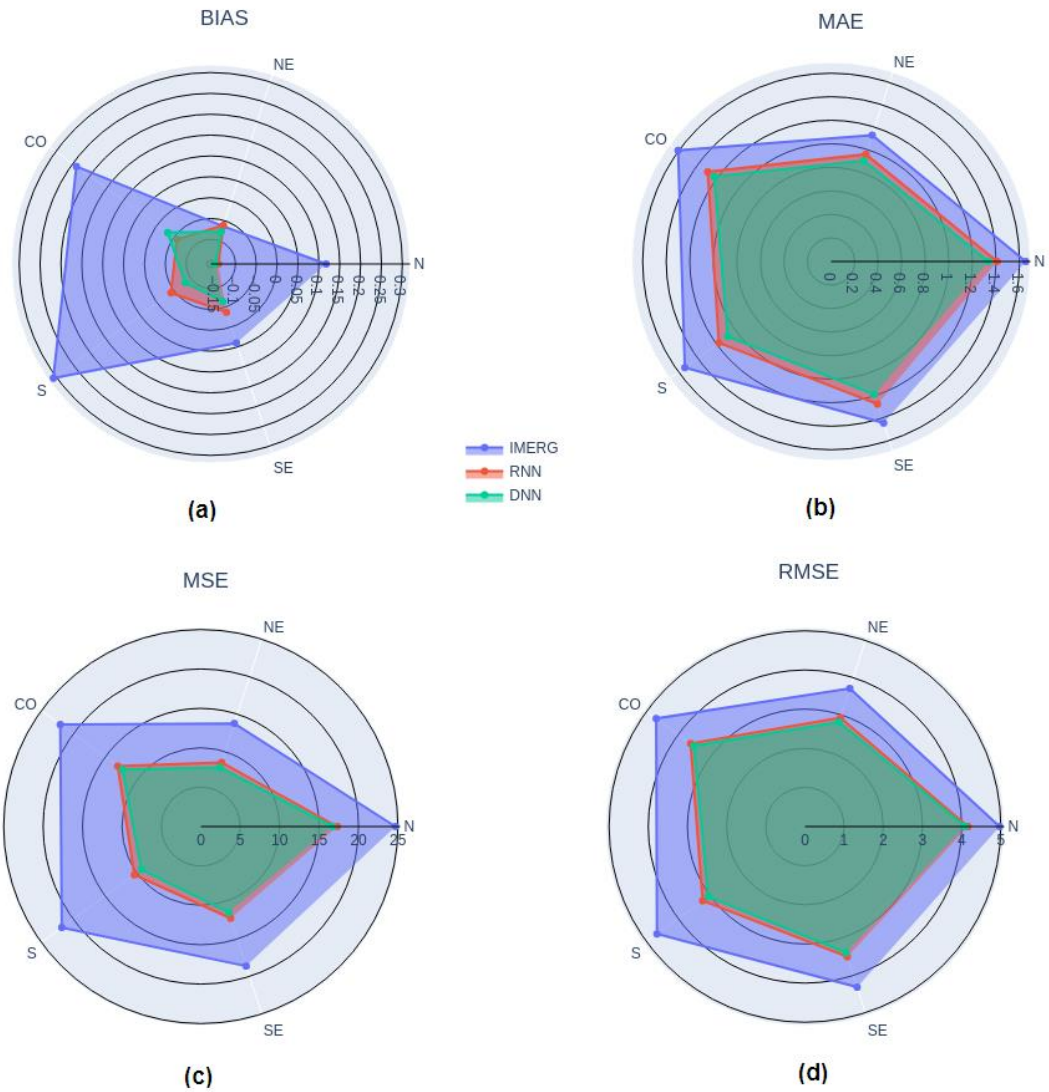
## APÊNDICE B – MÉTRICAS DNN-BRASIL

Figura B1: Métricas para a DNN-Brasil por horário.



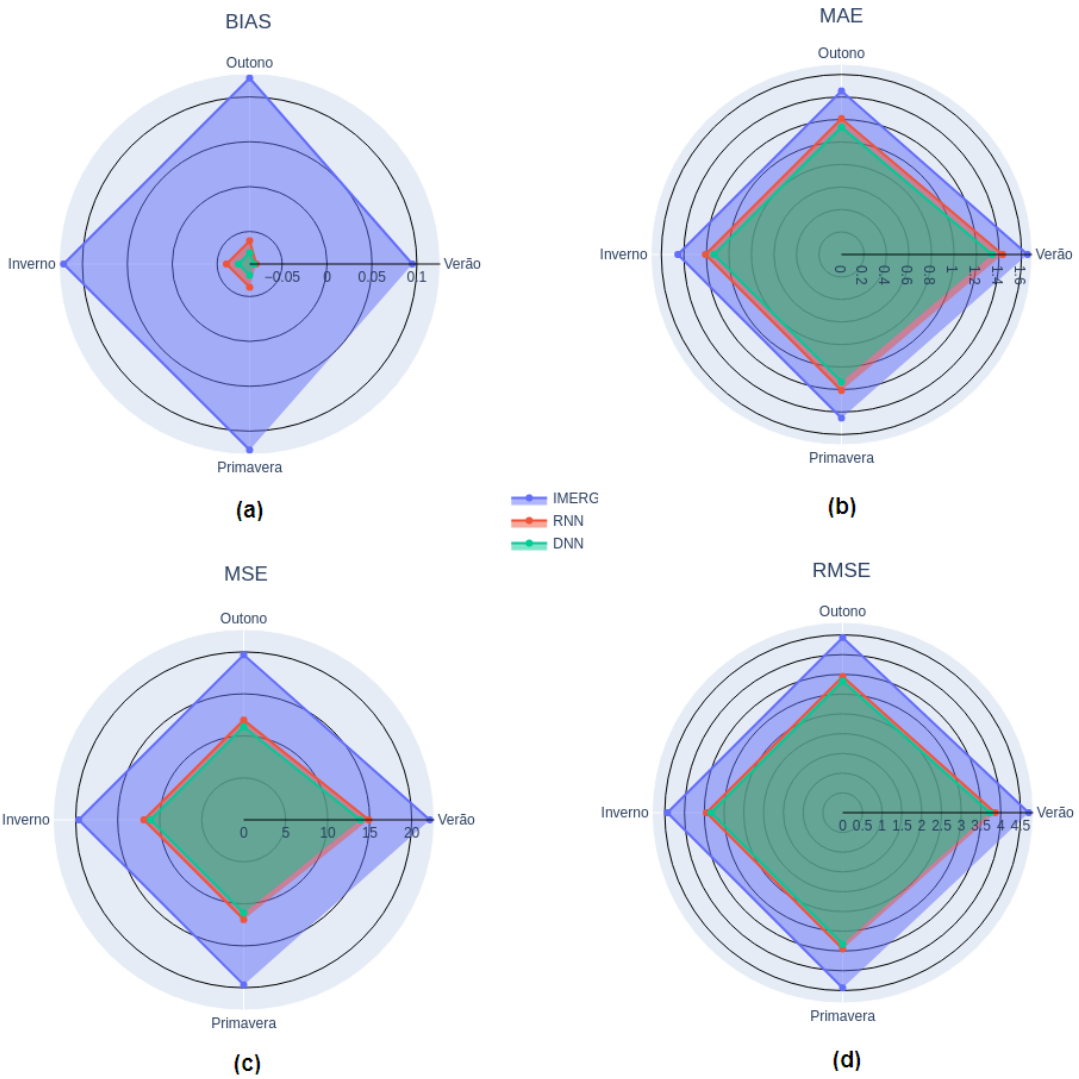

Fonte: Produção do Autor.

Figura B2: Métricas para a DNN-Brasil por região.




Fonte: Produção do Autor.

Figura B3: Métricas para a DNN-Brasil por estação.




Fonte: Produção do Autor.