UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Aplicação da Rede Neural LSTM para Predição de Nível do Rio Caí/RS

Luciana da Silva Mieres

Trabalho de Conclusão de Curso

MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Aplicação da Rede Neural LSTM para Predição de Nível do Rio Caí/RS

USP – São Carlos

2024

Aplicação da Rede Neural LSTM para Predição de Nível do Rio Caí/RS

Trabalho de conclusão de curso apresentado ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientadora: Profa. Dra. Elaine Souza
Co-orientador: Prof. Dr. Arthur Tschiedel

USP - São Carlos
2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

d633a

da Silva Mieres, Luciana
Aplicação da rede neural LSTM para predição de nível do rio Caí/RS / Luciana da Silva Mieres; orientadora Elaine Parros Machado de Sousa; coorientador Arthur da Fontoura Tschiedel. -- São Carlos, 2024.
74 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. Rede neural. 2. LSTM. 3. Inundação. 4. Nível. 5. hidrologia. I. Parros Machado de Sousa, Elaine , orient. II. da Fontoura Tschiedel, Arthur, coorient. III. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

AGRADECIMENTOS

Primeiramente gostaria de agradecer à minha rede de apoio que é formada por pessoas incríveis e fundamentais não somente na trajetória deste período do MBA, mas na vida. Essa rede é formada pelo meu marido, meu maior incentivador em todos os projetos que decido “meter a cara”, responsável pelos meus momentos de leveza, de força, de alegria e superação na hora do cansaço, além de me proporcionar as melhores discussões técnicas que sempre enriquecem meu conhecimento. Minha mãe por ser minha base, meu exemplo e meu ídolo. Minhas irmãs por serem minha base que me fortalece nas conversas e risadas. Minha sobrinha Camila por nossas conversas filosóficas que sempre me inspiram a querer seguir aprendendo sempre mais. Minha sobrinha Antonella, recém chegada a esse mundo e que já é força motivadora para que eu siga acreditando que através da pesquisa é possível mudar um pouquinho mais o mundo que ela vai crescer. Ao meu pai que não está mais conosco, mas plantou e cultivou em mim o desejo de estar sempre em movimento e em desenvolvimento.

Gostaria de agradecer à minha orientadora Elaine Sousa por todos os nossos encontros e orientação, fundamentais para o desenvolvimento e finalização desta pesquisa. Por fim, um agradecimento especial à Professora Solange Rezende pela compreensão e apoio em relação às circunstâncias que levaram ao pedido de prorrogação de prazo para a defesa.

RESUMO

A ocorrência de eventos naturais extremos vem ao longo dos anos resultando em desastres que geram danos socioeconômicos de grande relevância. O Rio Grande do Sul, que já têm sofrido com a ocorrência de desastres, vivenciou neste ano o mais expressivo de sua história tendo, nesse contexto, os maiores danos humanos ocasionados pelas inundações. Diante desse cenário e acreditando que o fortalecimento da fase de prevenção dentro do ciclo dos desastres, como aquela que tem maior potencial para impactar na redução dos danos socioeconômicos, vislumbra-se a adoção de Inteligência Artificial como ferramenta para a previsão de níveis de inundação de forma a possibilitar a emissão de alertas com antecipação capaz de salvar muitas vidas. Com esse propósito o presente estudo avaliou o uso da rede neural recorrente *Long Short-Term Memory* (LSTM) para a predição de nível do rio Caí especificamente no trecho localizado no município de São Sebastião do Caí/RS. Para seu treinamento e predição, utilizou-se uma série temporal composta por dados de 32 anos de precipitação e nível provenientes, das estações pluviométricas e fluviométrica da Agência Nacional de Águas (ANA), ambas localizadas na bacia hidrográfica do referido rio. Visando o refinamento do conjunto de dados, realizou-se uma análise exploratória identificando *outliers* e ausências de valores. Realizou-se também o pré-processamento do conjunto de dados que contemplou a sua divisão em conjunto de treinamento e de teste a normalização desses dois conjuntos e a transformação para o padrão exigido pela rede neural. A etapa seguinte consistiu no treinamento do modelo LSTM utilizando o conjunto de treino, dessa maneira, identificando os melhores hiperparâmetros a serem utilizados na predição, sendo esta a etapa final aplicada tanto ao conjunto de treino, quanto ao de teste. Os resultados encontrados foram avaliados com base nas medida Erro Médio Quadrático (MSE), no Raiz do Erro Médio Quadrático *(RMSE)* e Coeficiente de Nash-Sutcliffe (NSE), cujos valores encontrados apresentaram-se bastantes satisfatórios, sendo MSE (0,0035), RMSE (0,059) e NSE (0,87). A rede neural LSTM gerou resultados muito bons para valores recorrentes de nível, contudo valores extemos máximos ficaram um pouco subestimados, sendo esse o ponto a ser melhor desenvolvido. Como sugestão de trabalhos futuros está a adoção de uma série temporal mais volumosa e que seja composta por dados oriundos de fontes indiretas como sensoriamento remoto, como forma de contornar a restrição de amostras de dados de valores extremos, consequentemente ampliando a série temporal.

Palavras-Chave: rede neural recorrente, RNN, LSTM, predição de nível, precipitação, desastre natural, inundação

ABSTRACT

The occurrence of extreme natural events has, over the years, resulted in disasters that cause significant socioeconomic damage. Rio Grande do Sul, which has already suffered from the occurrence of disasters, experienced this year the most significant disaster in its history, with the greatest human losses caused by flooding. In this context, and believing that strengthening the prevention phase within the disaster cycle is the approach with the greatest potential to reduce socioeconomic damage, the adoption of Artificial Intelligence is envisioned as a tool for forecasting flood levels to enable the issuance of early warnings capable of saving many lives. With this purpose, the present study evaluated the use of the Long Short-Term Memory (LSTM) recurrent neural network to predict the level of the Caí River, specifically in the section located in the municipality of São Sebastião do Caí/RS. For its training and prediction, a time series comprising 32 years of precipitation and water level data was used, sourced from the rainfall and river gauge stations of the National Water Agency (ANA), both located in the river's hydrographic basin. To refine the dataset, an exploratory analysis was conducted to identify outliers and missing values. Subsequently, the dataset underwent preprocessing, which included dividing it into a training set and a test set, normalizing both sets, and transforming them into the format required by the neural network. The next step involved training the LSTM model using the training set, thereby identifying the best hyperparameters to be used for prediction. This final step was applied to both the training and test sets. The results were evaluated based on the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Nash-Sutcliffe Efficiency Coefficient (NSE). The values obtained were highly satisfactory: MSE (0.0035), RMSE (0.059), and NSE (0.87). The LSTM neural network produced very good results for recurring level values. Nevertheless, maximum extreme values were slightly underestimated. This decline in model performance is attributed to the limited sample size of maximum extreme values, which may not have been sufficient for the network's learning process. As a suggestion for future work, adopting a larger time series that includes data from indirect sources, such as remote sensing, is proposed. This approach could address the limitation of extreme data samples and consequently expand the time series.

Key-Words: Recorrent Network Neural, RNN, Long Short-Term Memory, LSTM,forecasting flood level, precipitation, natural desaster, flood

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE EQUAÇÕES

SUMÁRIO

1 INTRODUÇÃO

Países em todo o mundo são assolados por desastres, sejam de origem antropogênica ou natural. Ao considerar-se, na avaliação de risco, a multiplicidade de diferentes fatores socioeconômicos, políticos, de infraestrutura institucional, além dos ambientais, constata-se maior vulnerabilidade associada aos países pobres e em desenvolvimento, conforme a publicação *World Risk Report* (2023).

Os desastres podem ser classificados como naturais, tecnológicos e híbridos, sendo, respectivamente, resultantes da ocorrência de um fenômeno natural extremo, das ações antrópicas ou da relação entre desastres naturais e tecnológicos, conforme Monte (2022). Em relação aos híbridos, são definidos pela ocorrência de um desastre natural que leva à ocorrência de um tecnológico. Salienta-se, ainda, que a compreensão da temática que envolve os desastres necessita também do entendimento de outros conceitos igualmente importantes como perigo, vulnerabilidade e risco, uma vez que os fenômenos naturais e antrópicos por si só não resultam em um desastre, mas sua ocorrência sobre um sistema social vulnerável sim. Segundo Monte (2022), a premissa de perigo é a relação do fenômeno natural sobre um sistema social, podendo causar um potencial dano ao bem-estar da comunidade, sendo medido e definido por sua natureza, localização, extensão, magnitude intensidade, frequência e duração. Já a vulnerabilidade relaciona-se ao sistema social, ou seja, à população de uma determinada área e a relação que desenvolve com o território, estando, desse modo, diretamente relacionada à capacidade de resposta, à capacidade de enfrentamento, à resiliência, à capacidade de adaptação e susceptibilidade, definindo, assim, quais serão os possíveis danos que um fenômeno poderá ocasionar (MONTE, MICHEL E GOLDEFUN, 2018). No que tange ao risco, pode-se citar que decorre da relação entre perigo e vulnerabilidade, uma vez que está associado à exposição de uma comunidade a um fenômeno natural ou tecnológico, gerador do perigo, e à situação de vulnerabilidade em que a população se encontra (MONTE, 2022).

No Brasil a Secretaria Nacional de Proteção e Defesa Civil, órgão vinculado ao Ministério da Integração e do Desenvolvimento, busca universalizar o conhecimento em

proteção e defesa civil e, para tanto, estabeleceu a Classificação e Codificação Brasileira de Desastres (COBRADE) que segue a divisão entre naturais e tecnológicos (BRASIL, 2012|). Diante do exposto e considerando a extensão de conteúdo que envolve essa temática, o presente estudo está direcionado aos desastres naturais. Conforme a COBRADE, esses podem ser subdivididos em cinco grupos: geológicos, hidrológicos, meteorológicos, climatológicos e biológicos, os quais se diferenciam conforme a natureza do fenômeno de origem. Visando delinear ainda mais o tema, os esforços desta pesquisa são direcionados aos desastres hidrológicos, especificamente as inundações. Cabe salientar que esse grupo é dividido em três subgrupos: inundação, enxurrada e alagamento, e a motivação pela escolha do estudo sobre inundação baseia-se no quantitativo de danos humanos que desses eventos decorrem.

Segundo o *World Risk Report* (2023), que considera a vulnerabilidade obtida a partir de indicadores sociais, políticos, econômicos, e também a partir da exposição ao perigo natural relacionado a diversos fenômenos, o Brasil apresenta um índice de risco de desastres classificado como muito alto, ocupando a 40º posição em um *ranking* de 193 países. De acordo com o Atlas Digital de Desastres Naturais, no período de 1991 a 2022, o Brasil registrou um total de 57.581 ocorrências de desastres, sendo 28,4% referente ao grupo hidrológico. As inundações foram responsáveis pelo maior número de desalojados[1] e desabrigados[2] totalizando 3,62 milhões de pessoas, sendo que na contabilização total do número de pessoas que de algum modo foram atingidas pelos eventos, registra-se o valor de 20,15 milhões de afetados. Em âmbito nacional, o Rio Grande do Sul foi um dos estados mais afetados por desastres naturais nas últimas décadas, principalmente por eventos de inundação, conforme detalhado no Capítulo 2 deste documento. Em particular, o município de São Sebastião do Caí foi um dos municípios com maior número de registros de danos humanos ocasionados por inundações associadas à bacia hidrográfica do Rio Caí. Esse cenário, somado à disponibilidade de dados hidrometeorológicos históricos da região, motivaram a realização deste trabalho e a definição do município São Sebastião do Caí como área de interesse.

[1] Desalojado – pessoa que precisou deixar sua residência, mas foi hospedada na casa de familiares ou amigos.

[2] Desabrigados – pessoa que precisou deixar sua residência e ser alocada em abrigo fornecido pelo poder público.

No contexto do ciclo de proteção e defesa civil relacionado a desastres naturais, que compreende a prevenção, a mitigação, a preparação, a resposta e a recuperação, percebe-se que, nos últimos anos, estudos foram desenvolvidos, sobretudo no que tange à prevenção, etapa do ciclo capaz de reduzir danos humanos e prejuízos nas comunidades que possuem elevado risco, tendo ou não sido atingidas por algum desastre. O governo do Estado do Rio Grande do Sul lançou a publicação denominada Desastres Naturais do Rio Grande do Sul, cujo foco foi o de mapear as ocorrências de desastres naturais, além de analisar sua distribuição e frequência no território como forma de subsidiar o planejamento de ações da defesa civil estadual (RIO GRANDE DO SUL, 2022). Já o governo federal lançou em 2023 a plataforma denominada Atlas de Desastres no Brasil, a qual possibilita a visualização de dados sobre desastres de forma estruturada, contribuindo para o diagnóstico e planejamento de ações preventivas (BRASIL, 2023). Quando se avaliam estudos voltados à previsão de inundações, encontra-se soluções baseadas em modelos hidrológicos que auxiliam o provimento de informações importantes na gestão de recursos hídricos de forma integrada. É o caso de Fagundes (2021) que buscou desenvolver um sistema de previsão do aumento de nível do Rio do Boi, localizado em Santa Catarina, baseando-se em um modelo hidrológico-hidrodinâmico acoplado a dados de previsão por conjunto. Brunner et al (2021), discutiram os desafios do uso desses modelos para previsão de inundações e de secas. Modelos hidrológicos são uma representação matemática do ciclo hidrológico utilizada para compreender as relações existentes entre forçantes climatológicas e processos ambientais e hidrológicos (TSCHIEDEL, 2022). Ainda, segundo o autor, esses modelos podem ser classificados quanto a sua variabilidade espacial e representação de processos de escala da bacia hidrográfica, fazendo-se necessário o uso de variáveis como precipitação, evapotranspiração, infiltração, armazenamento de água e escoamento que podem ser provenientes tanto de medições diretas , quanto de fontes secundárias como sensoriamento remoto, além da utilização de dados para representação da topografia da área da bacia hidrográfica estudada, obtidos a partir de modelos digitais de elevação (MDE).

Contudo, a realização de pesquisas embasadas em modelos matemáticos mais robustos, como os modelos hidrológicos de base física, requer grande esforço e capacitação para compreensão da estrutura do modelo, de sua dinâmica de funcionamento, além das variáveis condicionantes utilizadas para previsão de inundações. Visando o entendimento de abordagens mais simplificadas, quando comparadas aos modelos hidrológicos, constatou-se

alguns estudos que aplicaram conceitos de inteligência artificial (IA) por meio de técnicas de aprendizado de máquina (*machine learning*) e aprendizado profundo (*deep learning*), para a obtenção de variáveis de interesse no contexto de desastres.

Sambati et al. (2019), desenvolveram uma aplicação destinada à previsão de risco de alagamento e inundação na Região Metropolitana de São Paulo (RMSP) a partir do uso de uma das técnicas de aprendizado de máquina, conhecida como KNN (K-*Nearest Neighbor*). Segundo os autores, para o desenvolvimento da aplicação, além do conhecimento relacionado à técnica de IA, também foi necessária a validação e qualificação dos dados utilizados como variáveis de entrada no modelo de previsão, utilizando, para tanto, dados de precipitação, de descargas elétricas, mapas de susceptibilidade de inundação e alagamento, além de adoção de um modelo baseado em linguagem natural para extração de informações das redes sociais acerca de pontos de alagamento e inundação, utilizando essas informações para o treinamento do modelo desenvolvido. Ainda nessa temática, porém utilizando aprendizado profundo, destaca-se o estudo desenvolvido por Schimdt et al. (2021), que desenvolveram uma aplicação baseada na técnica de redes neurais adversárias para criação de imagens de inundações em qualquer ponto de uma cidade, utilizando como base imagens do Google Street View[3], com o objetivo de promover a conscientização da população quanto à indução de desastres desse tipo em decorrência das mudanças climáticas. De acordo com os autores, a aplicação denominada *ClimateGam* permite que o usuário consulte um local através do endereço e então verifique como essa localização ficaria em caso de inundação. Cabe, ainda, destacar um sistema de alerta de inundação baseado em inteligência artificial, com técnica de aprendizado de máquina, elaborado pela empresa Google em parceria com o Serviço Geológico do Brasil (PHEBO, 2022). Segundo a autora, tal sistema emitirá alertas sobre previsão de inundação em tempo real a partir de consultas do usuário às plataformas Google[4] e Maps[5] para mais de 60 localidades no Brasil.

Embora a estrutura dos modelos de inteligência artificial não seja elementar, acredita-se que sua adoção possibilite o desenvolvimento mais simplificado de sistemas de

[3] Google Street View – disponível em: https://www.google.com.br/maps

[4] Google – disponível em: https://www.google.com.br

[5] Maps – disponível em: https://www.google.com.br/maps

alertas, quando comparado aos modelos hidrológicos existentes. Em particular, alguns estudos têm explorado técnicas de aprendizado profundo baseadas em redes neurais, tais como os trabalhos de Alberton et al. (2021) e Liang et al. (2018), apresentados em mais detalhes no Capítulo 2. Os resultados apresentados por esses estudos indicam que soluções baseadas em Redes Neurais Recorrentes (RRN, do inglês *Recurrent Neural Networks*) são abordagens promissoras, o que motivou a elaboração da seguinte questão de pesquisa, relacionada à área de interesse deste estudo:

> *"É possível prever cotas de inundação, no município gaúcho de São Sebastião do Caí, a partir da modelagem de variáveis de precipitação e nível fluviométrico utilizando Rede Neural Recorrente, especificamente a arquitetura LSTM (Long-Short Term Memory)?*

Os resultados confirmam que a utilização dessa técnica pode representar a simplificação do trabalho de previsão de eventos de inundação, uma vez que não necessita a calibração de variáveis para representação de processos de escala da bacia hidrográfica, como necessitam os modelos hidrológicos. Logo, este trabalho teve como objetivo geral a avaliação do uso de técnica de aprendizado profundo baseada em RNN, utilizando o modelo LSTM para modelagem de variáveis hidrometeorológicas, visando a previsão de cotas de inundação dentro de limiares tipicamente obtidos a partir do uso de técnicas de simulação tradicionais.

2 FUNDAMENTAÇÃO TEÓRICA E REVISÃO DE LITERATURA

Neste capítulo serão abordados tópicos relacionados à contextualização dos desastres naturais, trazendo um breve diagnóstico do Rio Grande do Sul no que tange ao tema, além da abordagem de modelos hidrológicos utilizados para previsão de inundações e por fim o uso de técnicas de redes neurais aplicadas à hidrologia, em especial previsão de cotas de inundação.

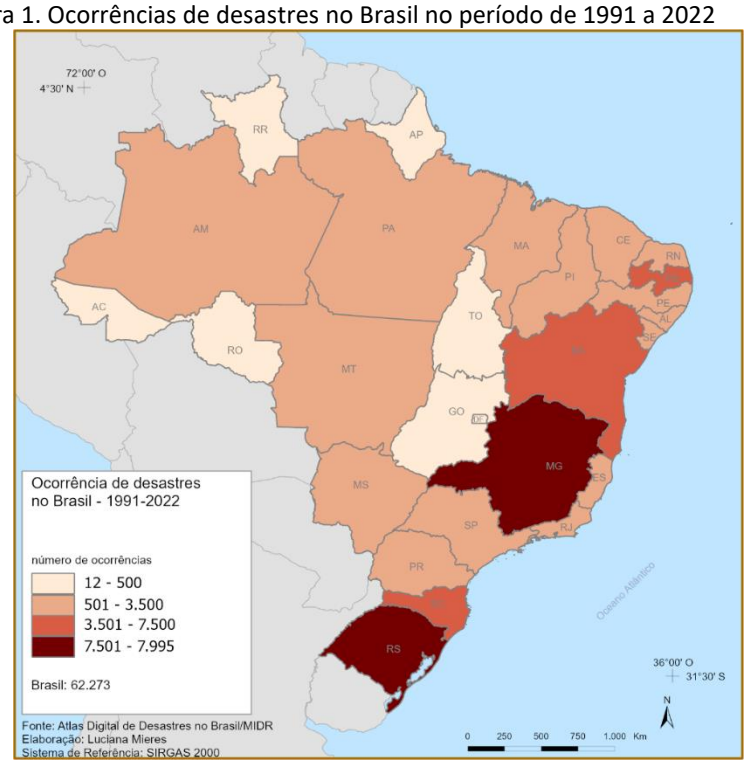
2.1 Contextualização

Em muitas partes do mundo, a ocorrência de eventos naturais extremos vem ao longo do tempo resultando em desastres que geram danos socioeconômicos de grande relevância. Ao mesmo tempo, os efeitos de médio e longo prazo das alterações climáticas não só resultarão no aumento da frequência como também no aumento da intensidade desses eventos, implicando em um maior número de pessoas potencialmente atingidas e susceptíveis a desastres no futuro (*World Risk Index,* 2023). Segundo BRASIL (1999), o termo "desastre" pode ser interpretado como o resultado de "eventos adversos, naturais ou provocados pelo homem, sobre um ecossistema vulnerável, causando danos humanos, materiais e ambientais e consequentes prejuízos econômicos e sociais". Considerando-se apenas os desastres naturais Tominaga, Santoro e Amaral (2015) os definem como "o resultado do impacto de fenômenos naturais extremos ou intensos sobre um sistema social, causando sérios danos e prejuízos que excedem a capacidade da comunidade ou sociedade atingida em conviver com o impacto".

A mensuração da intensidade de um desastre se dá pela quantificação dos danos humanos, materiais, ambientais e também pelos prejuízos econômicos públicos e privados associados. Conforme BRASIL (1999) a intensidade de um desastre vai depender da interação entre a magnitude do evento e o grau de vulnerabilidade do sistema social atingido por esse evento. A vulnerabilidade está associada à capacidade que uma população possui de se recuperar após a afetação por um desastre natural, estando diretamente relacionada à organização nas esferas política, institucional, econômica, social e ambiental. (*World Risk Index,* 2023)

Seguindo na avaliação do relatório *World Risk Index* (2023), verifica-se que os países que possuem o maior índice de risco de desastres naturais são Filipinas, Indonésia e Índia, sendo esse índice uma medida que considera a exposição da população ao fenômeno natural adverso e a vulnerabilidade desse grupo. O Brasil, segundo a publicação, ocupa a 40º posição entre os 193 países avaliados, classificando-se como um país de risco muito alto para desastres naturais. Isso fica evidenciado quando se avalia um período de 31 anos (1991 a 2022) de reconhecimento federal de decretos de situação de emergência[6] ou de estado de calamidade pública[7], totalizando 62.273 decretos de desastres que atingiram 10.462.600 pessoas, de acordo com o Atlas Digital de Desastres no Brasil. Nesse período, destacam-se os estados de Minas Gerais e Rio Grande do Sul como os de maiores quantitativos de decretos reconhecidos, sendo respectivamente 7.995 e 7.565. A Figura 1 ilustra a distribuição do número de decretos reconhecidos no período citado.

Figura 1. Ocorrências de desastres no Brasil no período de 1991 a 2022



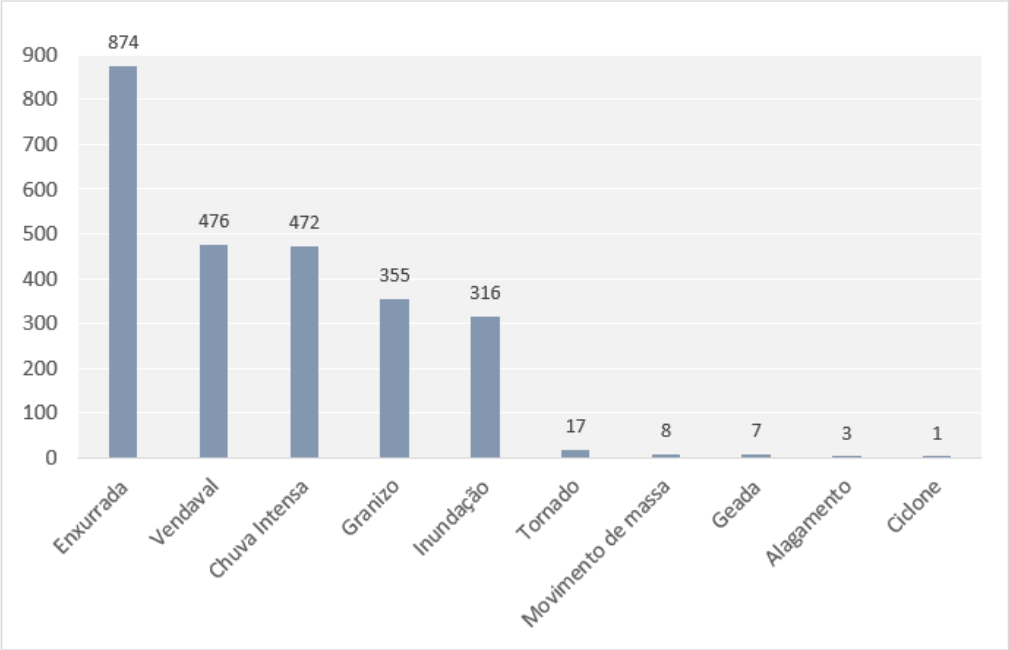
Fonte: Atlas Digital de Desastres no Brasil/MIDR – Elaborado pela autora

[6] Situação de Emergência: Reconhecimento legal pelo poder público de situação anormal provocada por desastres, causando danos suportáveis e superáveis pela comunidade afetada (BRASIL, 1999)

[7] Estado de Calamidade Pública: Reconhecimento legal pelo poder público de situação anormal provocada por desastre, causando sérios danos à comunidade afetada, inclusive à incolumidade e à vida de seus integrantes. (BRASIL, 1999)

O Rio Grande do Sul é um dos estados mais atingidos por desastres naturais conforme pode ser observado na Figura 1. Quando se avaliam as informações pertinentes ao estado, disponíveis no Sistema Integrado de Informações sobre Desastres (S2iD), referentes ao período de 2003 a 2023, verifica-se o reconhecimento federal de 5.124 decretos decorrentes da situação de urgência ocasionada especificamente por desastres naturais. Aprofundando-se a análise para verificação dos danos[8] ocasionados por esses eventos, constata-se que 235.783 pessoas diretamente afetadas, resultando em 67 mortes, 1.533 feridos, 1.599 enfermos, 93 desaparecidos, 23.755 desabrigados e 208.736 desalojados, além disso, um total de 248.699 habitações danificadas ou destruídas. Analisando-se os tipos de desastres naturais que acometeram o estado no período supracitado, observa-se o maior número de ocorrências de enxurrada, vendaval, chuva intensa, granizo e inundação (Figura 2).

Figura 2. Número de ocorrências, por tipo de desastre natural, no período de 2003 a 2023.



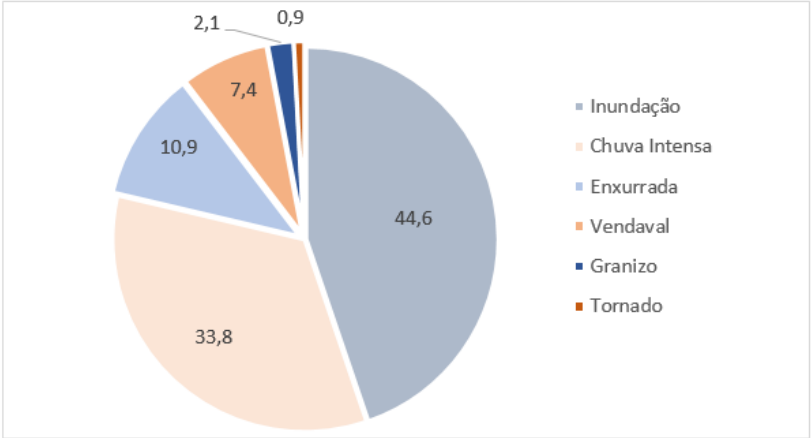
Fonte: S2iD/MIDR

Em uma avaliação dos danos humanos ocasionados pelos diferentes tipos de desastres naturais no Rio Grande do Sul, constata-se que os eventos de inundação e chuvas intensas são responsáveis por 78,4% dos danos humanos (Figura 3). Adicionalmente, na

[8] Dados de danos humanos e materiais e de prejuízos econômicos são restritos ao período de 2017-2023, conforme disponibilização do Sistema Integrado de Informações sobre Desastres.

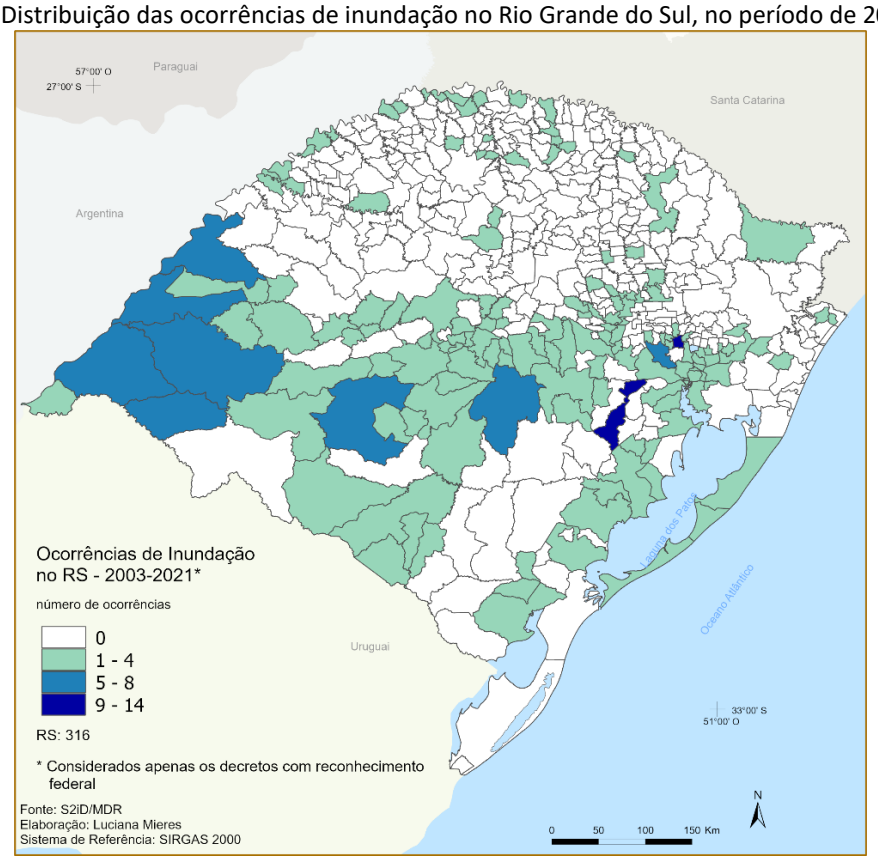
Figura 4 é apresentada a distribuição das ocorrências dos desastres de inundação no Estado, em que é possível observar que a região do município de São Sebastião do Caí é uma das que mais registrou eventos de inundações nos últimos anos.

Figura 3: Proporção de danos humanos conforme os tipos de desastres naturais



Fonte: S2iD/MIDR

Figura 4: Distribuição das ocorrências de inundação no Rio Grande do Sul, no período de 2003-2023

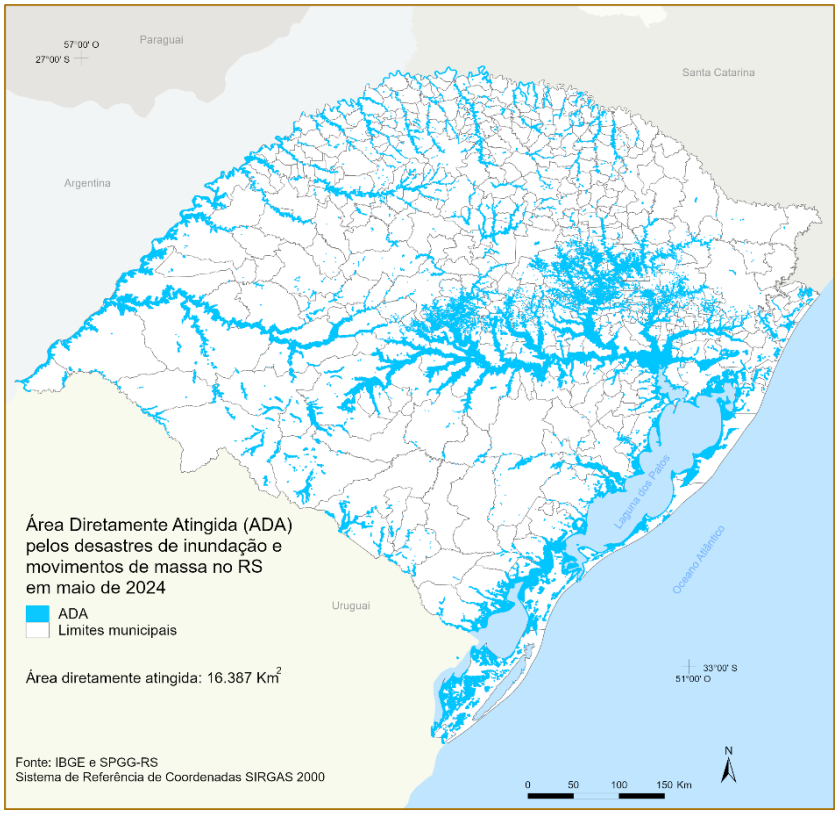


Fonte: Atlas Digital de Desastres no Brasil/MIDR – Elaborado pela autora

Dessa forma, a motivação do presente estudo parte da conexão existente entre o elevado percentual de danos humanos ocasionados por eventos de inundação e a relação direta que este processo tem com eventos de chuva intensa, em um contexto tecnológico cada vez mais abrangente que permite o uso de técnicas de IA para beneficiar diretamente populações vulneráveis a partir da previsão de níveis.

Neste sentido, salienta-se que essa motivação se mostrou ainda mais fortificada considerando os eventos catastróficos sofridos pelo Estado do Rio Grande do Sul ao final do mês de abril e durante o mês de maio de 2024. Segundo Paiva et al (2024) em algumas regiões foram registrados acumulados superiores a 900 mm, distribuídos por um período de 35 dias, sendo esse valor 10 vezes superior à precipitação média esperada. O resultado desse evento de precipitação intenso foi o maior desastre natural já vivenciado pelo estado. No que tange ao total de municípios atingidos, segundo o Mapa Único do Plano Rio Grande (MUP-RS) do total de 497 municípios, apenas 45 (9,1%) não tiveram a decretação de calamidade pública ou de situação emergência. O painel destaca ainda que 970.788 pessoas foram diretamente atingidas, o que 8,9% da população total e salientando, ainda, que a área atingida foi de 16.387 km_2 representando 6,1% da área total do estado. A Figura 5 ilustra a mancha decorrente dos desastres de inundação e movimentos de massa referentes ao evento acima citado.

Figura 5. Área diretamente atingida pelos desastres de inundação e movimentos de massa.



Fonte. IBGE, SPGG-RS – Elaborado pela autora

2.2 Modelos Hidrológicos

Modelos hidrológicos são tipicamente divididos em dois módulos: o módulo de balanço hídrico, o qual transforma chuva em vazão e o módulo de propagação que transforma vazão em nível (cota) em uma dada seção de interesse. O módulo de balanço hídrico tem como principal base a representação de fenômenos como a precipitação, a evapotranspiração e o balanço de água subsuperficial, bem como armazenamento subterrâneo. Já o módulo de propagação tem o intuito de verificar características associadas ao deslocamento das ondas de cheia para jusante podendo serem utilizadas abordagens mais complexas que envolvem o uso das equações de *Saint-Venant* ou mais simplificadas (TSCHIEDEL, 2022).

Nesse contexto, a consolidação de um modelo hidrológico para uma dada área de interesse exige considerável esforço, além de dados de entrada consistentes. Uma vez estabelecido o modelo hidrológico, também se destaca a etapa de calibração que exige um conhecimento substancial do analista para avaliação das diferenças entre o resultado

simulado e o dado observado. Portanto, a análise dos resultados simulados necessita conhecimento prévio das relações existentes nos módulos e submódulos presentes em um modelo hidrológico (TSCHIEDEL, 2022).

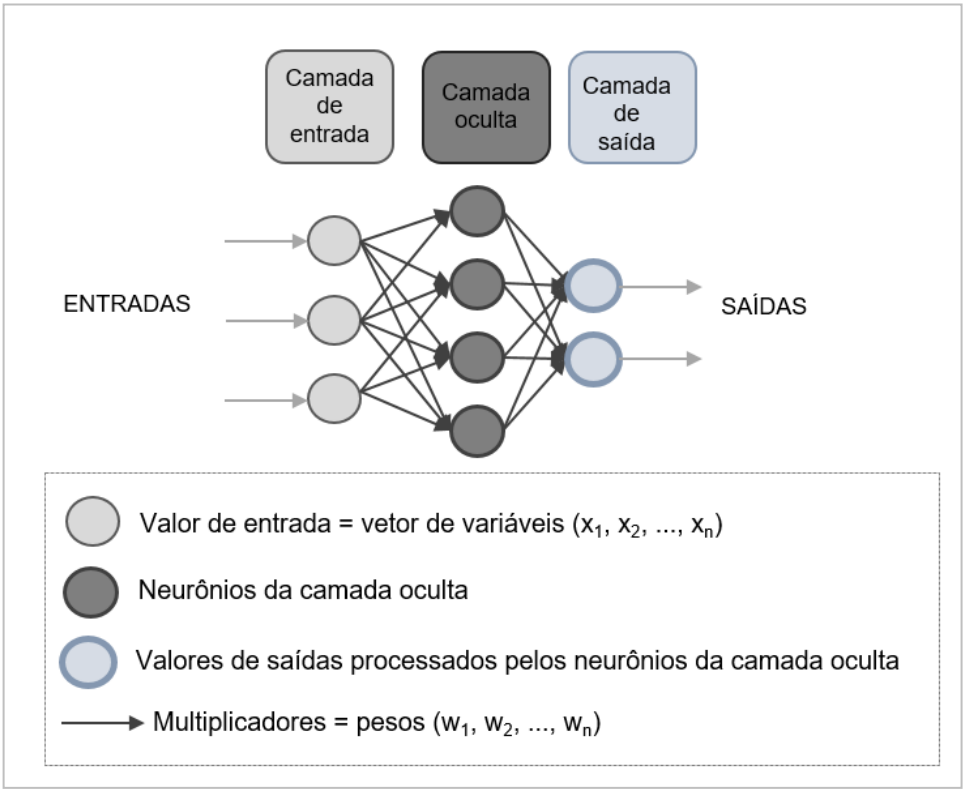
Em adição, aponta-se a dificuldade em obter-se longas séries históricas completas de dados observacionais, o que pode comprometer diretamente o resultado da simulação (KIM e KIM, 2021). A necessidade de calibração de um grande número de parâmetros, conjuntamente com a carência de dados hidrológicos tem sido apontados como fatores que complexificam a previsão de cheias a partir de modelos hidrológicos. (ALBERTON, SEVERO e MELO, 2021).

2.3 Redes Neurais

A adoção de redes neurais na hidrologia tem sido recorrente em estudos aplicados à previsão de inundações, ao gerenciamento de águas subterrâneas, à qualidade da água, e à previsão de precipitações, entre outros, desde o início dos anos 2000, conforme observado nos estudos de SPERB et al. (1999) e de GOVINDARAJU (2000). Segundo Cruz, Rodrigues e Versani (2010) as técnicas de redes neurais são promissoras nos estudos hidrológicos, pois utilizam dados de entrada mais simplificados, diferentemente dos modelos hidrológicos que necessitam de uma série de parâmetros como topografia do terreno, tipo de solo, estabelecimento de coeficientes, entre outros, tornando o processo de previsão mais trabalhoso. Neste sentido, no presente observa-se o uso de redes neurais como uma alternativa viável e consolidada de predição de variáveis hidrológicas como vazões e níveis (MEDEIROS et al., 2023; BOUIX, C.P. 2024). Contudo, apesar de ser apontada como uma promissora alternativa para estudos desse domínio, o processamento de redes neurais artificiais também necessita de uma base de dados rica em amostras. Shen (2018), afirma que as vantagens do aprendizado profundo aumentam à medida que aumenta o número de exemplos que são processados. No entanto, Gama e Pedrollo (2018) salientam que existe um número ideal de variáveis para que a rede obtenha um ponto ótimo de desenvolvimento, ressaltando que quantidades superiores a esse número ideal podem comprometer o desempenho.

Redes neurais são modelos matemáticos que possibilitam a modelagem de comportamentos temporais complexos e não lineares. Essas técnicas foram consolidadas diante da evolução de *hardwares* disponíveis comercialmente, além da consolidação de grandes bases de dados (NIELSEN, 2021). Sua origem foi inspirada nas ciências biológicas, sendo comparadas ao funcionamento do cérebro humano, conforme destaca Rohn e Mine (2003). A estrutura básica de uma rede neural pode ser observada na Figura 6 que ilustra de forma simplificada seu funcionamento.

Figura 6.: Estrutura genérica de uma rede neural



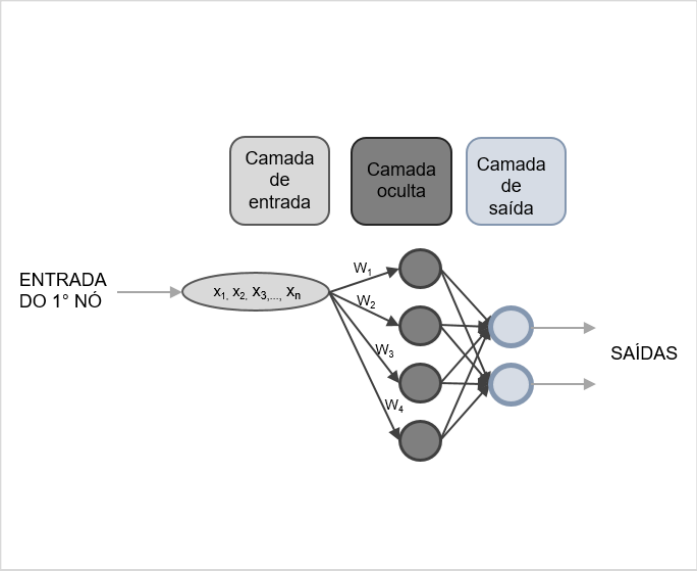
Fonte: Adaptado de Rohn e Mine (2003)

Os nós da Camada de Entrada representam os dados de entrada (vetores de valores), os quais são processados pelos neurônios localizados nas camadas ocultas. Antes deste processamento, os valores de entrada são multiplicados por pesos (vetor de pesos), representados pelas arestas. É importante destacar que a figura ilustra apenas uma camada oculta, porém as redes neurais profundas utilizadas em aprendizado profundo possuem múltiplas camadas intermediárias, as quais são compostas por n neurônios cada uma. Os valores de todas as entradas em um único nó (neurônio) são somados e passados para uma

função de ativação não linear (NIESLEN, 2021). Em resumo, o neurônio recebe as variáveis de entrada, já com a aplicação dos pesos e aplica sobre esse conjunto uma função de ativação.

A Figura 7 apresenta um recorte visual referente ao primeiro nó de entrada, ilustrando a multiplicação dos valores de entrada pelos pesos que os conectam aos neurônios. Desse modo, observa-se que o que ocorre é uma multiplicação de matrizes. Especificamente nesse exemplo, de uma matriz de 4x1, representando a multiplicação do vetor de entrada por 4 pesos, uma vez que se tem quatro neurônios na camada oculta. Ao final, aplica-se uma camada que combinará quatro entradas (oriundas dos 4 neurônios) em duas saídas (NIELSEN, 2021).

Figura 7: Esquema de representação do 1° nó de entrada na rede neural



Fonte: Adaptado de Rohn e Mine (2003)

Rohn e Mine (2003) compararam as entradas recebidas por um neurônio artificial aos estímulos que um neurônio natural recebe, destacando que o sinal que chegará aos núcleos do neurônio equivalerá ao número de entradas recebidas. Além disso, os autores destacaram que os pesos agregados aos atributos poderiam corresponder aos dendritos realizando suas sinapses em uma comparação ao sistema natural. Tais pesos refletem a importância de determinada entrada a um neurônio.

As redes neurais possuem distintas arquiteturas que variam conforme o modelo teórico sob os quais foram embasadas. Em particular, quando se busca um modelo para

previsões de eventos hidrológicos, é importante considerar abordagem mais adequada à análise de dados temporais, dentre as arquiteturas utilizadas de aprendizado profundo. Nesse sentido, Nielsen (2021) cita a rede *Multilayer Perceptron* (MLP), a Rede Neural Convolucional (CNN do inglês *Convolutional Neural Network*) e a Rede Neural Recorrente (RNN do inglês *Recurrent Neural Network*), salientando que as duas últimas, por serem mais recentes em relação à primeira são mais comumente utilizadas. Corroborando com a autora, Brownlee (2020) também utiliza as três redes para demonstrações de modelos aplicados a séries temporais.

Conforme anteriormente destacado, o uso dessa ferramenta aplicado ao domínio hidrológico não é recente. Rohn e Mine (2003), utilizaram uma rede neural recorrente para previsão de chuvas em um horizonte de curtíssimo prazo, apontando como aspecto positivo o fato de que não é necessário conhecimento detalhado sobre as relações entre as variáveis envolvidas e o problema. Rocha, Mine e Kavisky (2015), avaliaram o potencial de uma rede neural artificial do tipo *Perceptron* para obtenção de dados de vazão mensal a partir do processamento de variáveis geradas por um modelo climático regional, recomendando, diante dos resultados positivos, o uso da ferramenta para descrição do processo chuva-vazão com cenários climáticos, ressaltando ainda a vantagem econômica e operacional de implementação. Contudo, Nielsen (2021) reforça que apesar de a técnica demonstrar grande eficiência para processamento de séries temporais, ainda é imprescindível que se realize o pré-processamento dos dados, antes de sua aplicação.

A RNN, de interesse neste estudo dado o objetivo geral definido no Capítulo 1, constitui-se como um modelo em que os mesmos parâmetros são aplicados repetidamente, mesmo quando as entradas são alteradas com a passagem do tempo (NIELSEN, 2021). A autora destaca ainda as arquiteturas *Gated Recurrent Unit* (GRU) e a *Long Short-Term Memory* (LSTM), cujas diferenças básicas apontam para a maior rapidez de processamento característico à GRU, mas melhor desempenho da LSTM por dispor de mais parâmetros.

Estudos aplicados à hidrologia têm demonstrado bons resultados com uso de RNN, sobretudo com a adoção da arquitetura LSTM. Segundo KIM (2021), LSTM pode ser indicada para as situações em que se necessita estimar a vazão de um rio localizado em uma determinada região, para a qual não se dispõe de um conjunto massivo de dados exigidos por modelos hidrológicos. Em sua análise, a rede neural demonstrou resultados promissores ao estimar dados de vazão, quando comparados aos mesmos dados obtidos a partir de um

modelo hidrológico. Huang et.al (2023), avaliando desafios na previsão da dinâmica de recarga de água subterrânea, propuseram o uso do modelo LSTM, comparando-o com outros modelos de aprendizado de máquina, relatando a superioridade dos resultados processados em relação aos demais.

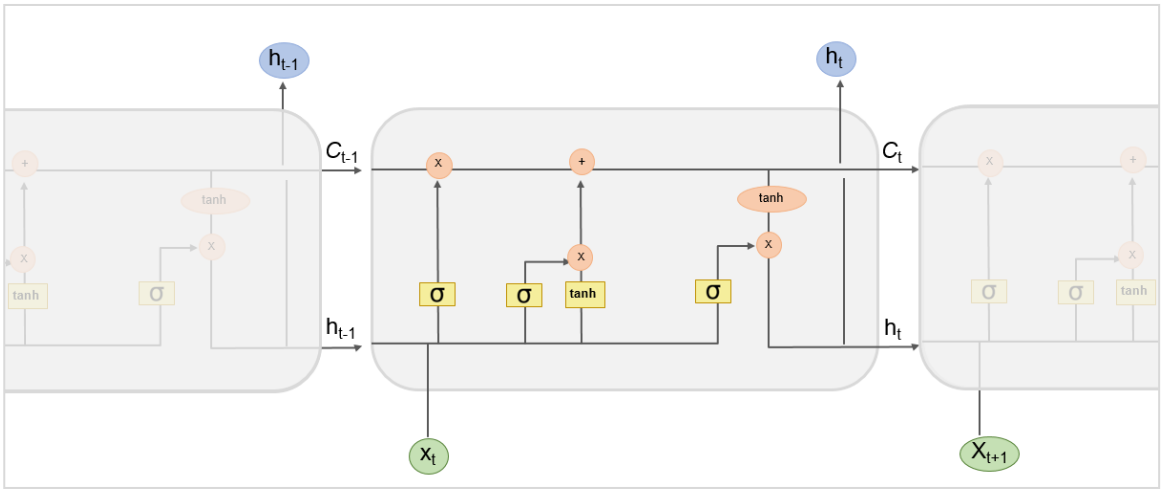
Em outro estudo, apresentado por Liang et al. (2018), o modelo LSTM foi utilizado para predizer a variação diária do nível de água do lago Dongting localizado na China, utilizando como dado de entrada níveis históricos e dados observados de vazão, apresentando resultados semelhantes aos obtidos a partir de modelagem hidrodinâmica, com a vantagem de ser um processo mais otimizado uma vez que, diferentemente dos modelos hidrodinâmicos, não exige uma quantidade substancial de parâmetros, calibração do modelo e recursos computacionais. Por fim, destaca-se o estudo de Alberton et al.(2021), no qual os autores avaliaram dois modelos de redes neurais, LSTM e MLP, para predição do nível do rio Itajaí-Açu localizado no município de Blumenau em Santa Catarina, considerando como dados de entrada informações de precipitação e nível ao longo de pontos da bacia hidrográfica. O estudo concluiu que ambos os modelos podem ser aplicados à finalidade proposta, mas destacou o modelo LSTM que, com simples pré-processamento dos dados, foi capaz de prever com alta precisão o nível da água do rio durante eventos de cheia, apresentando melhores resultados quando comparado ao MLP.

Os exemplos supracitados fundamentam a escolha da rede LSTM na aplicação do presente estudo. Perante o exposto, o entendimento de sua arquitetura é de extrema relevância uma vez que, em etapas posteriores, será realizado o desenvolvimento das análises temporais utilizando-a como modelo de processamento. Assim sendo, cabe destacar que as RNN são redes que contêm loops que permitem que uma memória anterior de entrada persista influenciando na saída. Contudo, muitas vezes a informação anterior acaba decaindo à medida que o treinamento vai ocorrendo, o que acontece quando o gradiente usado para atualização da rede se torna pequeno à medida que os resultados do processamento vão sendo propagados, refletindo na camada de saída. Visando solucionar este problema, denominado dissipação do gradiente, foi desenvolvida a arquitetura LSTM (LIANG et al, 2018).

Segundo Hochreiter e Schmidhuber (1997) apud Migliato (2021), a LSTM surgiu com a intenção de resolver diversos problemas relacionados ao aprendizado utilizando dados sequenciais, de modo a obter-se um modelo que fosse efetivo e escalável. Os autores

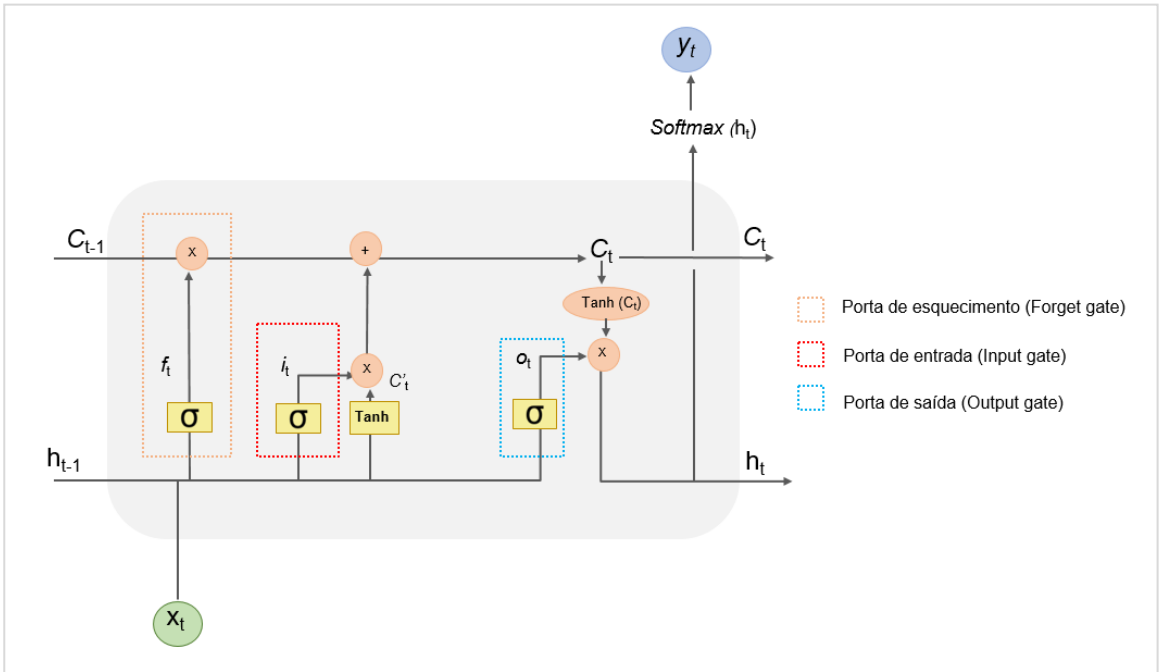
ressaltaram ainda que diferentemente de outros tipos de redes neurais que possuem núcleos de processamento (neurônios), as LSTM são compostas por blocos de memória conectados por meio de camadas, tendo como ponto preponderante o estado da célula (*cell state*). A Figura 8 ilustra essa estrutura, enquanto que a Figura 9 exemplifica o funcionamento de um bloco de memória LSTM. Após estas figuras, as explicações das variáveis associadas são realizadas a partir do exposto em Migliato (2021), Liang et al. (2018) e Brownlee (2020).

Figura 8: Esquema ilustrativo da estrutura de blocos de memória da rede LSTM



Fonte: Adaptado de Liang et al. (2018)

Figura 9. Esquema de funcionamento de um bloco de memória de uma rede LSTM



Fonte: Adaptado de Liang et al. (2018)

Na Figura 9, "x_t" representa o vetor de entrada no tempo "t" e "h_{t-1}" representa a camada escondida anterior no tempo t-1. Os blocos de memória possuem três portões que são o portão de esquecimento (*forget gate*), o portão de entrada (*input gate*) e o portão de saída (*output gate*).

A primeira etapa é dedicada à seleção de qual informação será repassada ao estado da célula, por meio de uma função logística, ou seja, o portão de entrada recebe x_t e h_{t-1} que são multiplicados pela função logística, a qual consiste em uma função sigmoide (σ). Desse modo, produz-se um sinal (f_t) que possui como saída valores entre 0 e 1, em que 0 significa que a informação será totalmente descartada e 1 que será totalmente considerada. Esse sinal é multiplicado por cada valor do estado da célula anterior (Ct-1), determinando o quanto do estado da célula anterior será considerado no estado atual (Ct). A seguir a equação do portão de esquecimento (*forget gate*) (1), onde wf é o peso de entrada e bf é o peso do viés, e a equação da função logística (2).

$$f_t = \sigma\left(w_{f[h_{t-1},x_t]} + b_f\right) \qquad \text{Equação 1}$$

$$\sigma(x) = \frac{1}{1 + e^{-x}} \qquad \text{Equação 2}$$

A segunda etapa do processamento consiste na decisão do quanto a nova informação será mantida no estado da célula (Ct). Essa etapa é dividida em dois momentos. No primeiro momento, o portão de entrada recebe xt no tempo t e ht-1 multiplicando-os por matrizes de peso W_i as quais são somadas a um viés bi e processadas por outra função sigmoide, novamente produzindo uma saída entre 0 e 1. A seguinte equação é dada no portão de entrada (*input gate*):

$$\sigma i_t = \sigma\left(w_{i[h_{t-1},x_t]} + b_i\right) \qquad \text{Equação 3}$$

No segundo momento é gerado um novo estado (C'_t) a partir da aplicação de uma função tangente hiperbólica sobre x_t no tempo t e $ht-1$ sendo o resultado multiplicado pelo resultado do portão de entrada. O resultado desta multiplicação é passado ao estado atual (C_t), já considerando o processamento realizado na primeira etapa ($f_t * C_{t-1}$). A seguir a equação que gera o novo estado (4) e a equação que calcula a tangente hiperbólica (5):

$$C'_t = tanh\left(w_{c'[h_{t-1},x_t]} + c'\right) \quad \text{Equação 4}$$

$$tanh(x) = \frac{e^{2x} - 1}{e^{2x} + 1} \quad \text{Equação 5}$$

Por fim, a terceira e última etapa dedica-se a calcular a saída do modelo. Nessa etapa, primeiramente é passado o estado da célula novamente por meio de uma função tangente hiperbólica gerando um valor. Esse valor é multiplicado pelo resultado do portão de saída, o qual é obtido aplicando a função sigmoide em xt no tempo t e $ht-1$, desse modo indicando o quanto do estado atual estará presente na saída. A seguir apresenta-se a equação do portão de saída (Ot) (*output gate*) (6) e a equação de obtenção da saída (7).

$$tO_t = \sigma\left(w_{o[h_{t-1},x_t]} + b_o\right) \quad \text{Equação 6}$$

$$h_t = O_t * \tanh\left(C_t\right) \quad \text{Equação 7}$$

3 METODOLOGIA E PROPOSTA DE DESENVOLVIMENTO

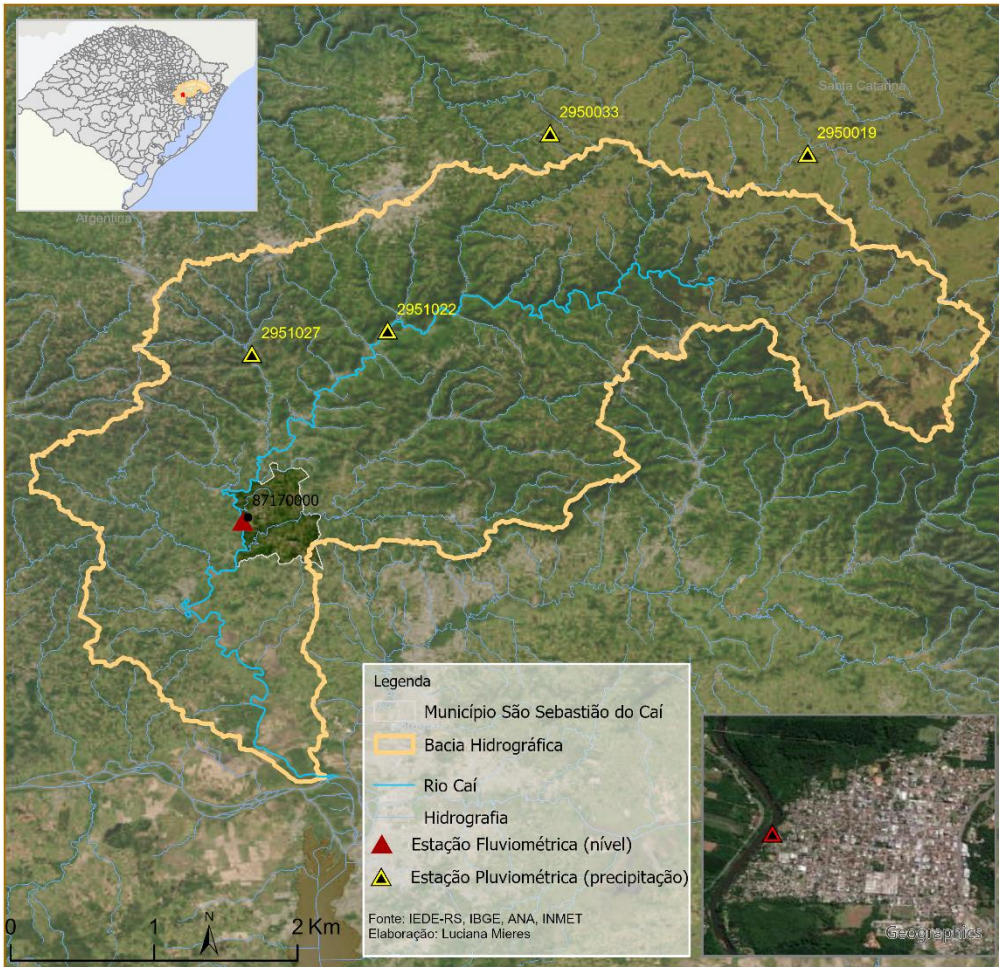
O presente capítulo apresenta as etapas de desenvolvimento deste estudo que objetiva analisar a aplicação de uma rede neural recorrente, modelo LSTM, para obtenção de dados de nível fluviométrico a partir do processamento de variáveis hidrometeorológicas, além de apresentar a área de estudo.

3.1 Área de Estudo

Conforme relatado no capítulo 2, o Rio Grande do Sul é um dos estados brasileiros mais atingidos por desastres naturais, sobretudo os de natureza hidrometeorológica. Recentemente, o Estado foi acometido pelo maior desastre dessa natureza e de sua totalidade de municípios, 415 decretaram situação de calamidade pública ou de emergência, representando 83,5% desse total. Um município que se destaca em relação aos demais por seu histórico em ocorrência de desastres do tipo inundação é São Sebastião do Caí, localizado na porção leste do território estadual, especificamente na bacia hidrográfica do Rio Caí. Ainda, cabe ressaltar que o município, ao longo do período analisado, foi um dos que mais registrou danos humanos devido a inundações. Diante disso e considerando o monitoramento dessa bacia hidrográfica com disponibilidade de dados hidrometeorológicos, o presente estudo concentrará sua análise nesse município.

A Figura 10 ilustra a localização do município e sua inserção na bacia hidrográfica do rio Caí, bem como a identificação das estações pluviométricas e fluviométrica utilizadas para consolidação da base de dados.

Figura 10. Localização da área de estudo

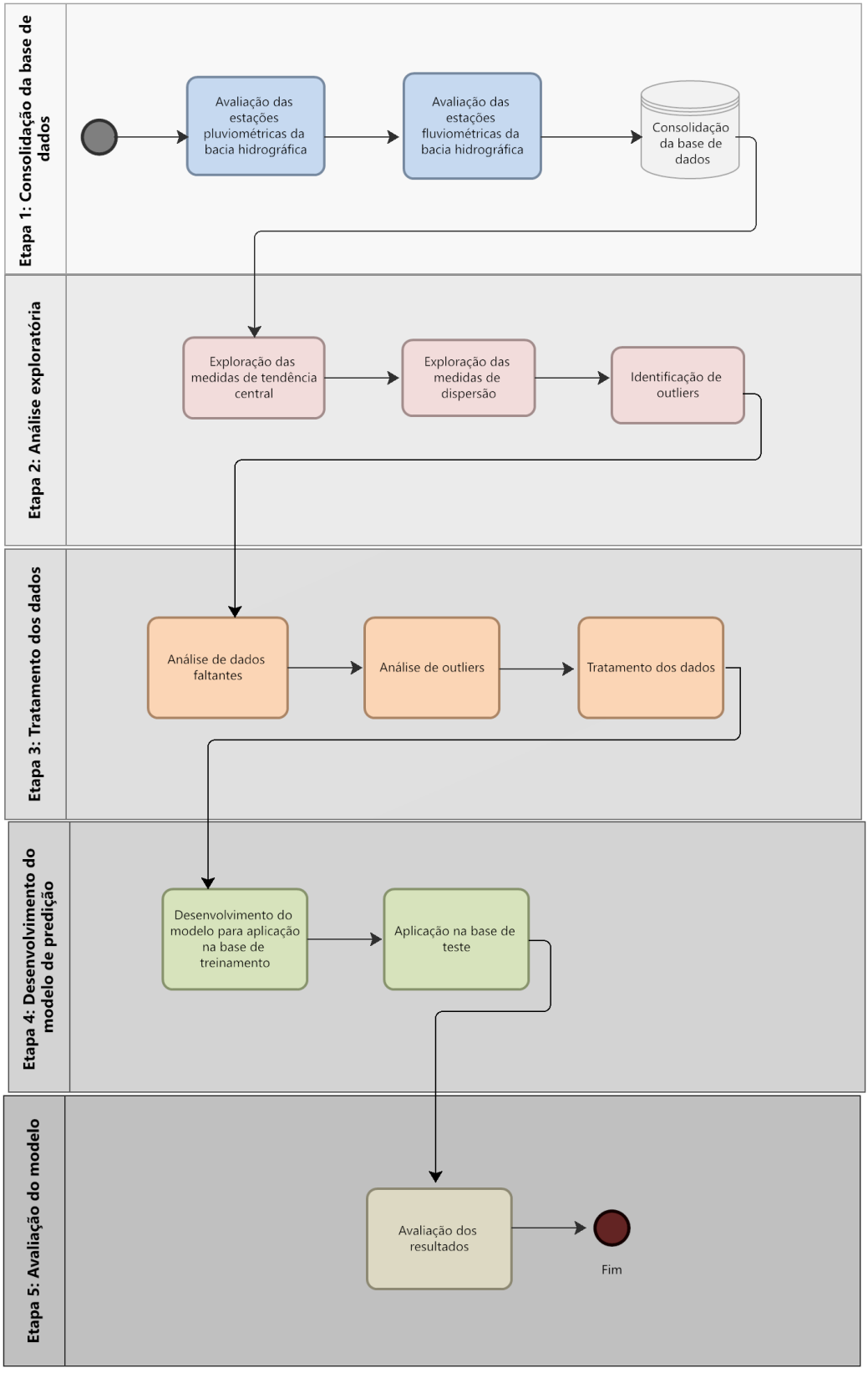


Fonte: ANA, IBGE e SEMA-RS – Elaborado pelo autora

3.2 Etapas de desenvolvimento

Destacam-se como principais atividades deste trabalho: 1) a avaliação das estações pluviométricas e fluviométrica, consolidando a base de dados; 2) a análise exploratória dos dados, buscando principalmente a identificação de *outliers*; 3) o tratamento e transformação desses dados; 4) o desenvolvimento do modelo de predição baseado na arquitetura de rede neural recorrente LSTM e sua aplicação ao grupo de treinamento e de teste; 5) avaliação final dos resultados a partir do emprego de medidas de acurácia. A Figura 11ilustra as etapas de desenvolvimento do estudo, representando um resumo da metodologia utilizada.

Figura 11: Etapas de desenvolvimento do estudo



Fonte: Elaborado pela autora

No que tange às etapas de desenvolvimento destaca-se a importância da primeira dedicada à consolidação da base de dados, consistindo na avaliação das estações pluviométricas, das quais obteve-se a série de dados de precipitação, e da estação fluviométrica que forneceu dados de nível do rio em estudo. Cabe ressaltar que as estações pluviométricas e a fluviométrica, pertencentes à Agência Nacional de Águas (ANA), apresentam momentos de falhas operacionais dos equipamentos repercutindo na ausência de dados e configurando algumas lacunas nas séries temporais. Diante dessa constatação, analisa-se as séries de dados dessas estações para definir um recorte temporal que contemple o menor período de falhas, ou seja, que contenha o menor número de dados ausentes proporcionalmente ao total da amostra. Tendo isso em vista, optou-se por consolidar uma base de dados diários de precipitação e nível que representam o recorte temporal de 1970 a 2004. Apesar de não representar uma série recente de dados, a escolha foi também fundamentada no objetivo desse estudo, o de avaliar o uso da rede neural LSTM para predição de nível fluviométrico, e não o de aplicar técnicas de Inteligência Artificial no tratamento dos dados ausentes em si.

A segunda etapa consiste na análise exploratória do conjunto de dados como forma de compreender suas características principais. Essa análise é fundamental para identificar o comportamento do conjunto de dados em relação a um ponto central, como se caracteriza sua distribuição e principalmente, possibilita a identificação de *outliers* que em estudos hidrológicos são bastante relevantes para indicar os períodos de maiores níveis que podem resultar em inundações expressivas. A etapa seguinte consiste no tratamento do conjunto de dados, buscando identificar valores faltantes, analisando-se os *outliers* com intuito de compreender se decorrem de falhas dos equipamentos ou se de fato representam o fenômeno meteorológico que repercute no hidrológico e, por fim, dedica-se ao tratamento do conjunto de dados. A quarta etapa dedica-se à criação do modelo de predição baseado na rede neural LSTM com a configuração de seus parâmetros e aplicação no conjunto de treinamento. Entende-se a grande relevância dessa etapa para esse trabalho, por ser fundamental o ajuste de parâmetros do modelo visando os melhores resultados de treinamento, para que posteriormente, o modelo devidamente ajustado possa ser aplicado o conjunto de teste. Por fim, a etapa final consiste na análise dos resultados de predição do nível fluviométrico obtido, resultado diretamente dependente do correto desenvolvimento das etapas anteriores.

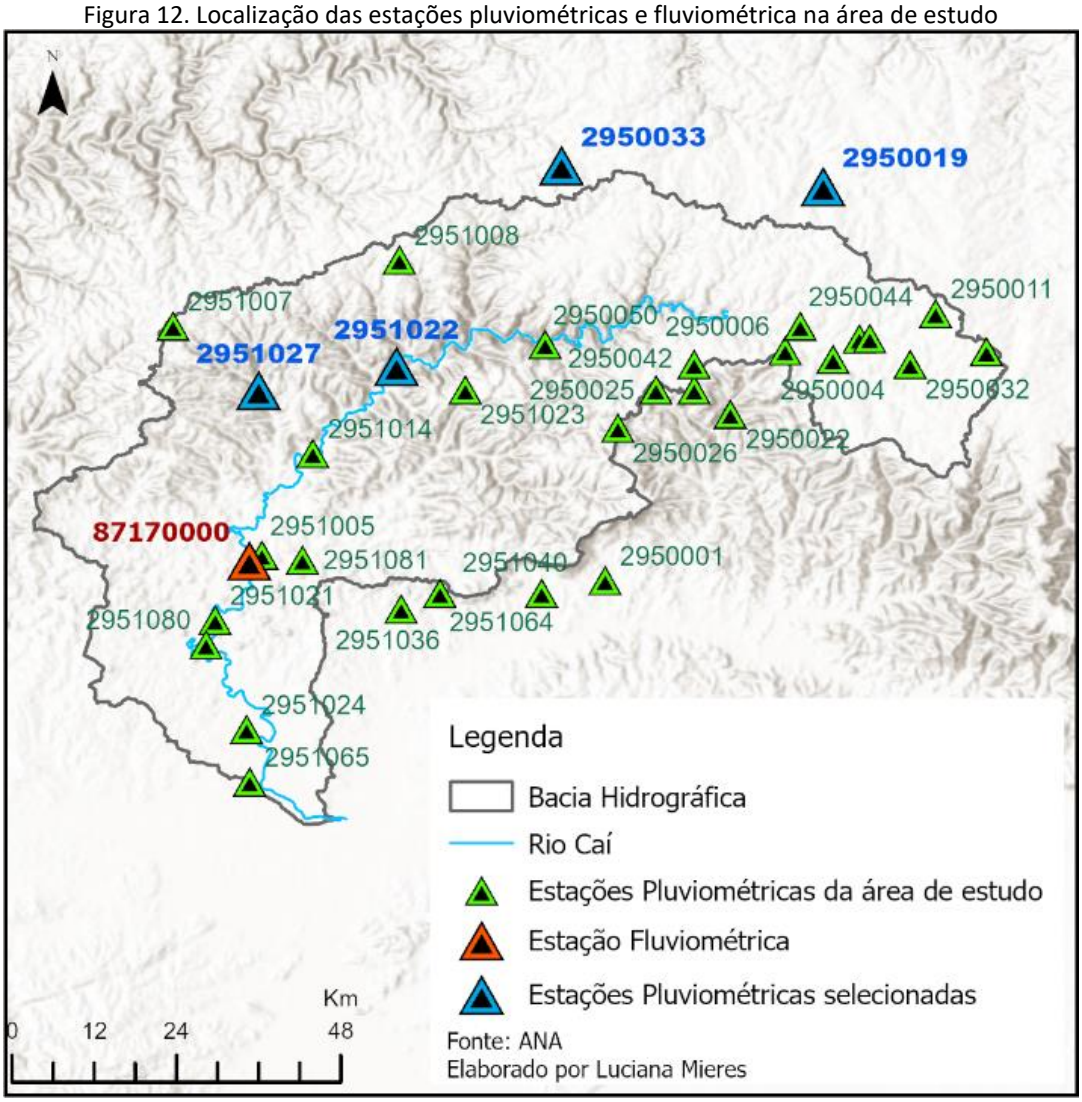
4 RESULTADOS E DISCUSSÃO

O presente capítulo dedica-se à exploração do conjunto de dados sob o viés das medidas de tendência central e de dispersão, além da identificação de *outliers* e adoção de outras estatísticas descritivas, referentes à etapa de análise exploratória. Posteriormente, será apresenta a etapa de tratamento dos dados, na qual serão abordadas a análise dos dados faltantes e suas transformações, bem como a análise dos *outliers* identificados anteriormente. Em seguida será apresentado o processo de trabalho relacionado à etapa de desenvolvimento do modelo de predição, sua aplicação ao conjunto de treinamento e de teste e finalmente será abordada a análise dos resultados com base em medidas estatísticas indicadas tanto para avaliação de modelos de regressão, quando de dados hidrológicos.

4.1 Consolidação da base de dados

Para a composição da base foram considerados dados de precipitação de quatro estações pluviométricas e dados de nível do rio Caí da estação fluviométrica, todos obtidos diretamente a partir do acesso ao portal Hidroweb da Agência Nacional de Águas e Saneamento Básico (ANA). A série temporal dessas estações varia entre os anos de 1950 a 2024, entretanto a ocorrência de falhas operacionais ao longo do período, as quais são identificadas pelo código “-1”, ocasionou diversas lacunas tendo como consequência a ausência de dados. A Figura 12 traz o mapa ilustrando a localização da estação fluviométrica e das pluviométricas existentes na área de estudo, as quais foram analisadas com base em suas séries temporais para então serem definidas aquelas que seriam utilizadas na presente pesquisa. A Figura 13 traz o gráfico de Gantt ilustrando a disponibilidade temporal de dados de todas essas estações, onde amarelo indica que a estação possui menos de 90% de dados disponíveis e azul representa que possui 90% ou mais. Observando a Figura 13 é possível identificar a existência de falhas nos dados de precipitação e nível ao longo do período, indicando ainda o percentual de dados existentes.

Figura 12. Localização das estações pluviométricas e fluviométrica na área de estudo



Fonte. ANA – Elaborado pela autora

Figura 13. Disponibilidade temporal das estações pluviométricas e fluviométrica com indicação do percentual de dados existente para o ano

	Estação	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974
Precipitação	2950001	0	0	0	0	0	0	0	0	0	0	0	0	17	100	100	100	100	100	100	100	100	100	100	100	100
	2950002	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2950004	0	0	0	0	0	0	0	0	83	100	98	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	2950005	0	0	0	0	0	0	0	0	0	0	0	100	100	100	100	100	100	100	100	100	100	100	95	100	98
	2950006	0	0	0	0	0	0	0	0	0	0	0	99	98	100	100	100	100	99	100	100	99	100	100	100	100
	2950009	100	100	37	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	2950011	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	2950019	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	2950022	100	100	100	100	100	100	100	100	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2950025	0	75	100	100	100	100	92	100	100	100	100	100	100	87	92	100	100	100	92	50	100	100	92	100	100
	2950026	100	100	100	100	100	100	100	100	100	100	100	100	92	100	100	50	100	100	100	100	100	100	99	100	100
	2950032	100	100	100	100	100	100	100	100	100	100	100	92	0	0	0	0	0	0	0	0	0	0	0	0	0
	2950033	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	2950041	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	98	100	100	100
	2950042	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2950044	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2950050	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2951005	0	0	0	0	0	0	0	0	0	0	0	0	9	100	100	100	100	100	100	100	100	100	100	100	99
	2951007	0	0	75	100	100	100	100	100	100	100	100	100	92	100	100	100	100	0	0	0	0	0	0	0	0
	2951008	100	100	100	100	100	100	100	92	100	92	92	92	100	100	100	100	100	92	100	100	100	0	0	0	0
	2951014	52	100	100	100	100	75	100	100	100	100	100	92	100	100	100	100	100	92	0	0	40	92	100	70	100
	2951021	100	100	100	100	100	100	100	100	100	100	100	92	100	100	100	100	100	100	100	100	100	100	100	100	100
	2951022	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	2951023	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2951024	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	88	92	100	100	100
	2951027	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	100	100	100	100
	2951036	0	0	0	0	0	0	0	0	0	0	0	76	100	100	100	100	100	66	100	100	100	100	100	100	100
	2951040	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2951064	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2951065	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2951080	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2951081	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Nível	**87170000**	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

Estação	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
2950001	100	100	100	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2950002	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2950004	100	100	100	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2950005	100	100	100	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2950006	100	98	100	100	98	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2950009	100	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2950011	100	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2950019	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
2950022	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2950025	100	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2950026	100	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2950032	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2950033	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
2950041	100	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2950042	75	0	0	58	50	48	67	24	92	0	0	0	0	59	0	0	0	0	0	0	0	0	0	0	0
2950044	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2950050	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2951005	100	100	100	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2951007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2951008	97	81	71	89	50	64	70	65	100	99	25	0	0	52	75	0	9	92	92	95	92	75	92	92	0
2951014	100	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2951021	100	100	100	100	100	100	100	100	100	100	84	76	100	100	84	0	0	0	0	0	0	0	0	0	0
2951022	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	93	100	100
2951023	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
2951024	100	100	100	100	100	100	100	100	100	100	92	100	100	92	100	100	100	100	100	92	100	100	100	100	100
2951027	100	100	100	100	100	100	100	100	100	100	100	100	100	92	100	100	72	100	100	100	100	100	100	100	100
2951036	100	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2951040	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2951064	0	0	0	0	0	0	0	0	0	0	0	0	0	75	74	0	0	0	0	95	98	73	0	62	0
2951065	0	0	0	0	35	50	62	67	100	0	0	0	0	78	0	0	0	0	0	99	98	99	99	99	0
2951080	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2951081	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
87170000	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

Estação	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
2950001	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2950002	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2950004	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2950005	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2950006	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2950009	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2950011	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2950019	100	100	100	100	100	100	100	100	100	100	100	100	100	0	0	0	84	100	100	100	100	100	100	0
2950022	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2950025	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2950026	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2950032	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2950033	100	100	100	100	100	100	100	100	100	100	100	100	100	0	0	0	84	100	100	100	100	100	100	0
2950041	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2950042	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2950044	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2950050	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	89	100	100	100	100	100	0
2951005	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2951007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2951008	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2951014	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2951021	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2951022	100	100	100	100	100	100	100	75	100	72	66	100	100	100	100	100	100	100	100	100	100	100	100	0
2951023	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2951024	100	100	100	100	100	100	100	67	100	83	67	100	100	100	100	100	97	100	98	98	100	100	100	0
2951027	100	100	100	98	99	100	98	62	99	82	67	100	99	100	100	100	92	95	100	100	100	97	100	0
2951036	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2951040	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2951064	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2951065	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2951080	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	100	58	100	100	100	0
2951081	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	99	100	99	79	99	0
87170000	100	100	100	100	100	100	92	75	100	96	100	100	100	100	100	100	100	100	100	100	100	100	100	0

Fonte. ANA – Elaborado pela autora

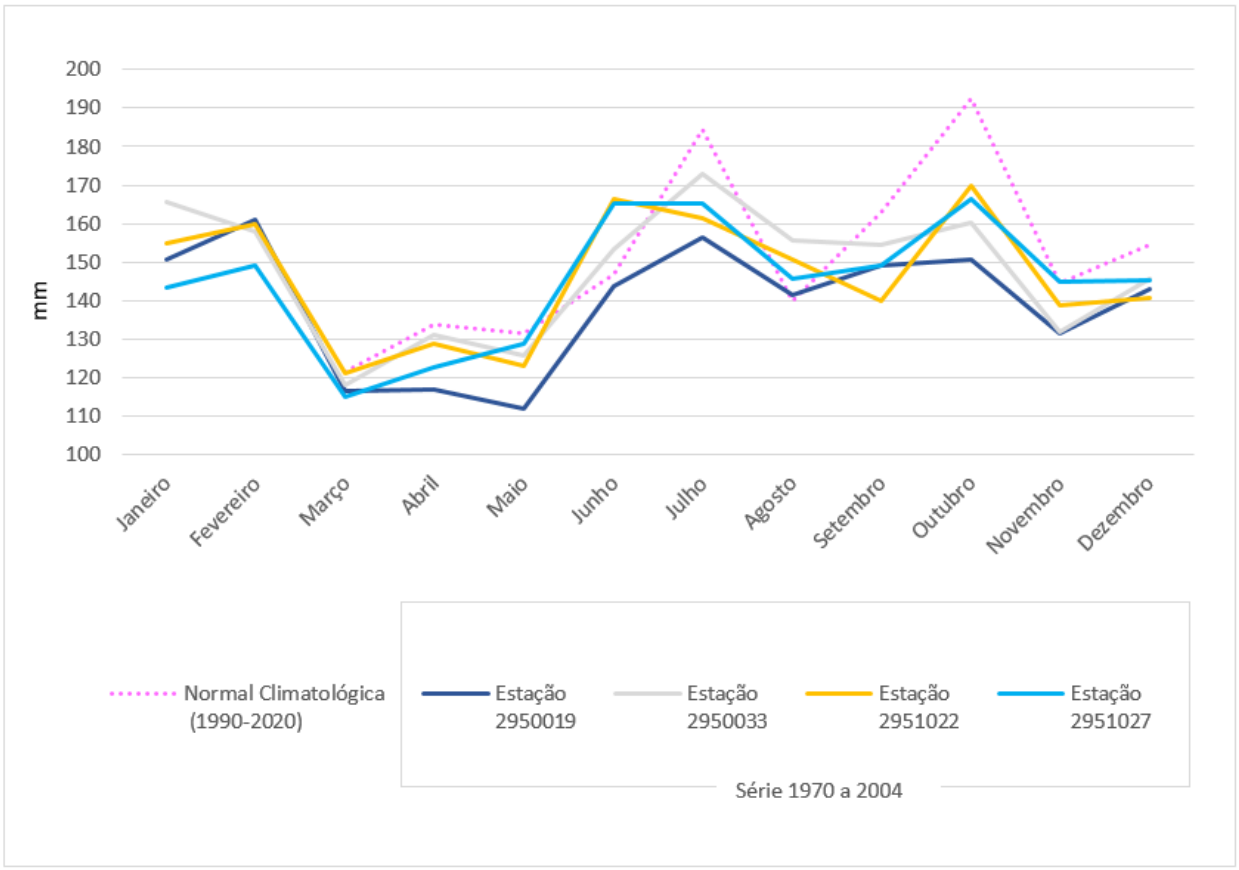
A Figura 13 ilustra a disponibilidade temporal dos dados das estações indicando, sobretudo, o quanto existem falhas nos dados de precipitação. Segundo Abreu et al (2018), para estudos hidrológicos a indicação é considerar estações com um percentual máximo de falhas de 10% por ano. Sendo assim, para a análise em questão, foram consideradas as estações com percentual mínimo de disponibilidade de dados de 90% buscando, dessa forma, diminuir o impacto na fase do tratamento de dados ausentes. Sobre o tratamento, salienta-se brevemente que as metodologias utilizadas consideram o uso de outras estações próximas para que possam ser inferidos dados naquelas que possuem falhas com base nas estações vizinhas. Neste contexto, a partir da interpretação do gráfico de Gantt (Figura 13), é possível destacar as estações pluviométricas 2950019, 2950033, 2951022 e 2951027 como aquelas que possuem as séries mais completas ao longo do tempo e por essa razão foram as estações selecionadas para comporem a base de dados deste trabalho. No que tange à estação fluviométrica percebe-se que até o ano de 2004 não houve registro de falhas na leitura dos níveis diários do rio Caí, contudo, nos anos de 2006 e 2007 ocorreram falhas e considerando a metodologia de tratamento dos dados, a qual será melhor explanada na próxima seção, entende-se que para a série de níveis a escolha do melhor período seria até o ano de 2004.

Diante da constatação dos períodos de falha e considerando-se ainda que o objetivo deste estudo é avaliar o uso da rede neural LSTM na predição de níveis com base em dados de precipitação, entende-se que a calibração do modelo em questão necessita de uma série temporal com o mínimo de falhas e tratamento de dados. Visando a redução de viés no resultado do modelo, portanto, optou-se por estabelecer um recorte temporal que representasse o período mais completo. Sendo assim, foi definido o conjunto de dados diários de precipitação e nível referente ao período de 1970 a 2004. Apesar da série temporal finalizar em 2004, o que representa uma defasagem de 20 anos em relação ao período atual, ainda assim está de acordo com a indicação da Organização Mundial de Meteorologia (OMM) que sugere a adoção de períodos de 30 anos para estudos climáticos, uma vez que a série temporal definida totaliza 35 anos de dados diários.

A fim de ilustrar a representatividade do período selecionado, a Figura 14 traz o comparativo entre as médias mensais de precipitação para a série definida (1970-2004) com o período oficial mais recente de normal climatológica (análise de 30 anos de dados

meteorológicos) determinados pelo INMET (1991-2020)[9]. É importante ressaltar que os dados da normal climatológica se referem à estação localizada no município de Caxias do Sul, o qual está próximo da área de estudo (São Sebastião do Caí) e representa o clima da região. Sua análise possibilita verificar que o comportamento médio mensal para o período do estudo acompanha a tendência observada pela normal.

Figura 14.Precipitação média mensal



Fonte: Dados da Pesquisa – Elaborado pela autora

Finalizando a definição do período escolhido apresenta-se a Figura 15 que ilustra a completude do conjunto de dados tanto de precipitação quanto de nível, atendendo aos critérios indicados para estudos hidrológicos, conforme citados anteriormente. Após a finalização deste estudo, quando o modelo LSTM estiver calibrado para a predição de dados de nível a partir de variáveis de precipitação e nível, a sugestão é testá-lo, em trabalho futuro, com séries temporais mais recentes obtidas a partir de fontes de medição indireta, caso dos dados de precipitação

[9] Normais Climatológicas disponível em: https://portal.inmet.gov.br/normais

obtidos a partir de satélites ou derivados da combinação de diferentes fontes de medição indireta (*ensemble)*, como forma de contornar o problema da ausência de dados apresentado.

A seção a seguir dedica-se à análise exploratória dos conjuntos de dados, tendo sido desenvolvido considerando a série temporal de 1970 a 2004, conforme definido acima.

Figura 15. Disponibilidade temporal de dados das estações pluviométricas e fluviométrica (% ano) para o período de 1970 a 2004 – 35 anos de dados diários de precipitação e de nível

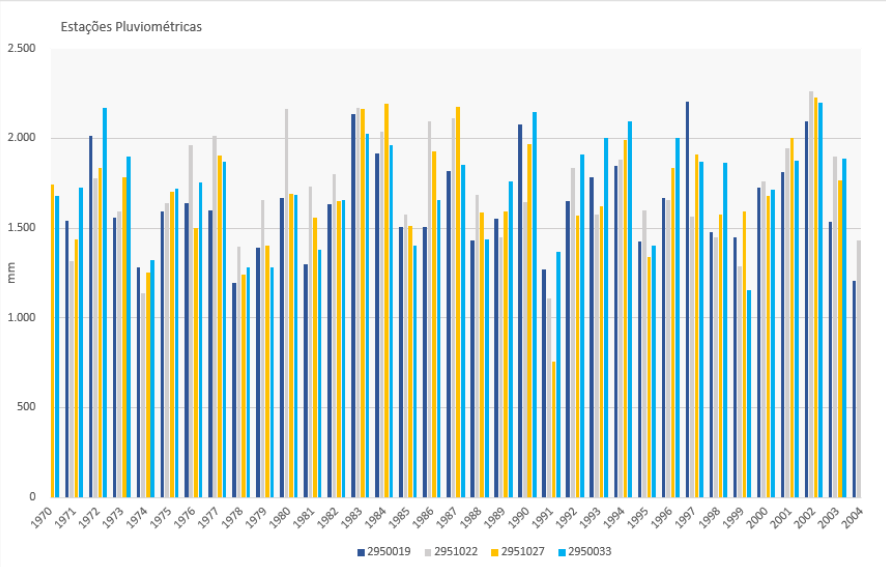
	Estação	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Precitação	2950019	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	2950033	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	2951022	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	**93**	100	100	100	100	100	100	100
	2951027	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	**92**	100	100	72	100	100	100	100	100	100	100	100	100	100	100	**98**	**99**
Nível	87170000	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

Fonte: ANA – Elaborado pela autora

4.2 Análise Exploratória

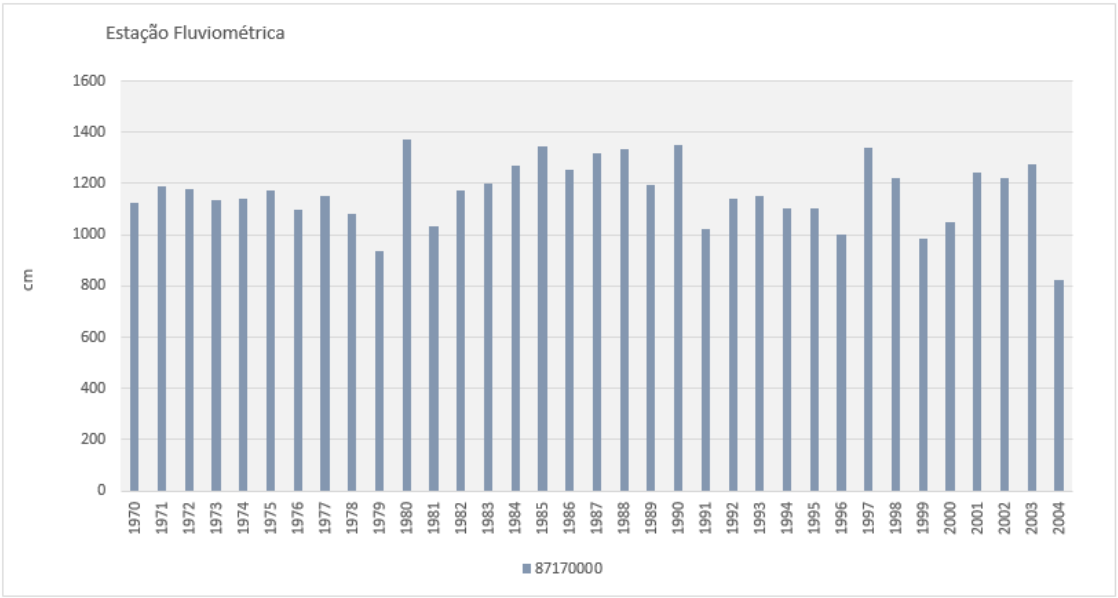
Compreender o comportamento da precipitação na bacia hidrográfica é fundamental para o entendimento da relação entre as variáveis em análise. Por essa razão, foram elaborados gráficos referentes à precipitação e ao nível fluviométrico de acordo com o recorte temporal anteriormente citado. A Figura 16 ilustra a precipitação anual acumulada nas quatro estações, enquanto a Figura 17 ilustra o nível fluviométrico máximo anual. Observando-as é possível verificar que a bacia hidrográfica do Rio Caí se caracteriza por apresentar chuvas acumuladas anualmente com valores acima de 1.000 mm, com exceção para a estação 2951027 (em amarelo na Figura 15) no ano 1991, que apresentou acumulados inferiores em razão ausência de dados no período. Também se destaca que é comum a observação de níveis máximos anuais do rio acima da cota de 10 metros, uma vez que em 32 anos, de um período de 35 anos, o rio Caí atingiu cotas máximas acima do valor referido.

Figura 16. Precipitação acumulada anual nas estações pluviométricas



Fonte: Dados da Pesquisa – Elaborado pela autora

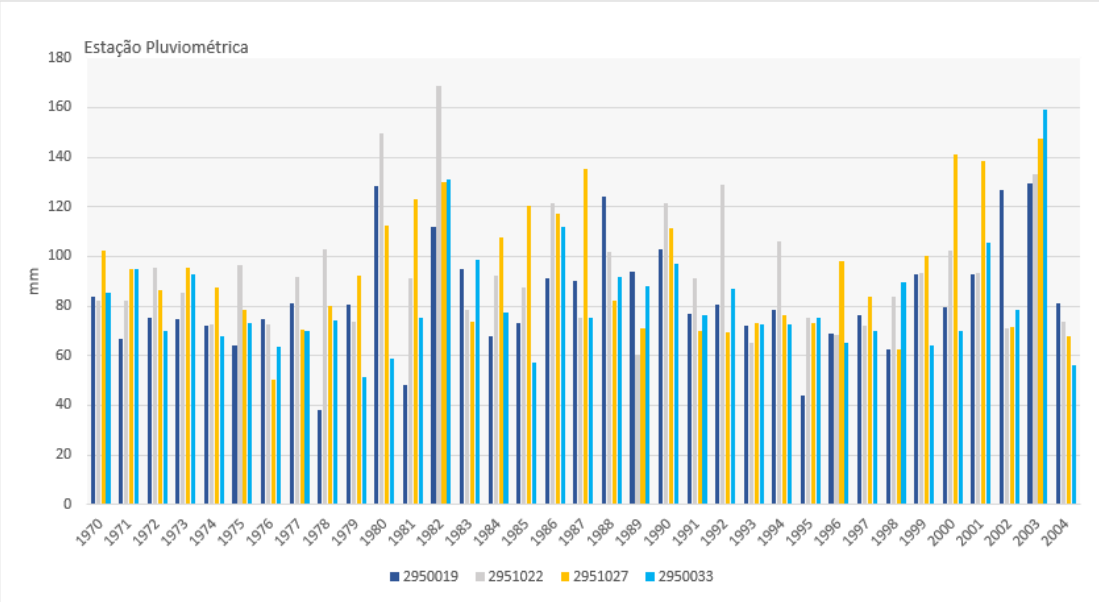
Figura 17. Nível do Rio Caí máximo anual na estação fluviométrica



Fonte: Dados da Pesquisa – Elaborado pela autora

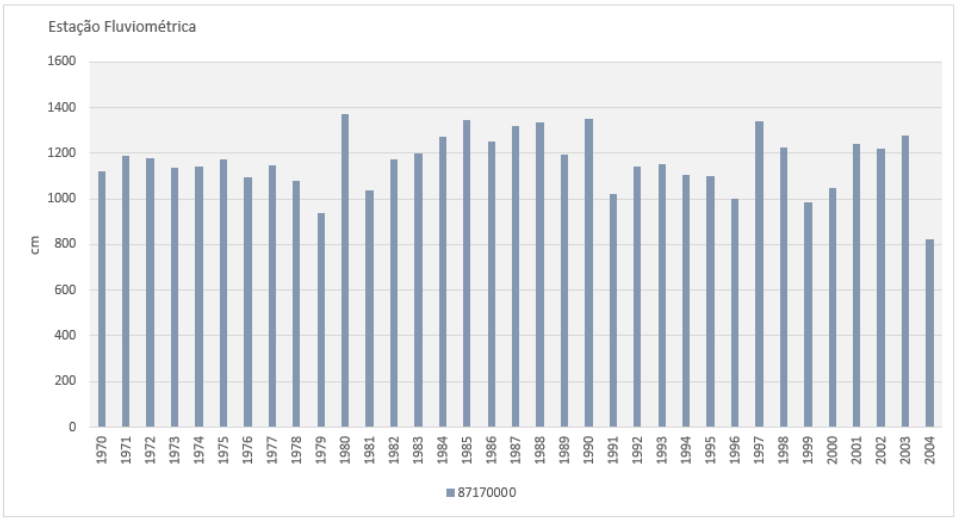
Quando avalia-se os valores máximos diários de chuva anuais, verifica-se o predomínio de registros acima de 80 mm em todas as estações, com destaque para as estações 2951022 e 2951027 com recorrência de registros máximos diários anuais superiores a 100 mm. No que tange ao nível da água no Rio Caí, observa-se valores predominantes acima de 10 metros, conforme as Figura 18 e Figura 19 respectivamente.

Figura 18. Precipitação máxima diária anual nas estações pluviométricas



Fonte: Dados da Pesquisa – Elaborado pela autora

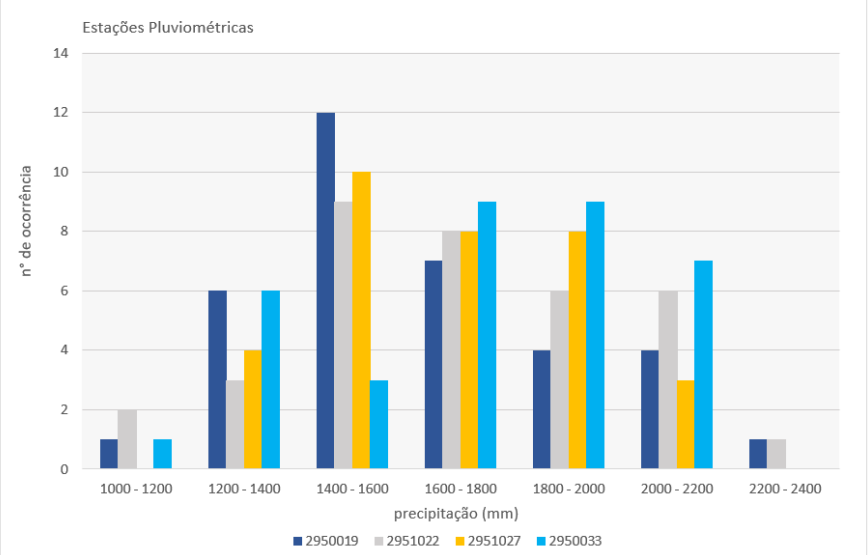
Figura 19. Nível máximo diário anual do Rio Caí na estação fluviométrica



Fonte: Dados da Pesquisa – Elaborado pela autora

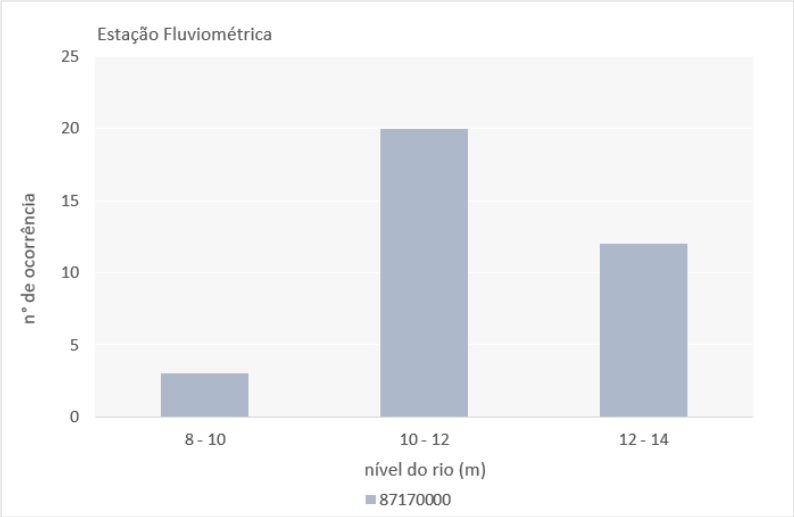
Observando-se a precipitação a partir da ocorrência de volumes anuais acumulados distribuídos em classes, verifica-se que em três estações ocorre o predomínio de volumes em torno de 1400 a 1600 mm anuais, enquanto na estação 2950033 o predomínio é caracterizado por volumes superiores, entre 1600 a 2000 mm. Em relação à cota do rio Caí, constata-se o predomínio entre 10 e 12 metros de subida máxima do nível da água. A Figura 20 e a Figura 21 ilustram, respectivamente, a distribuição de ocorrência de volumes anuais acumulados e a distribuição de ocorrência de níveis anuais máximos, ambos divididos em intervalos.

Figura 20: Distribuição de ocorrência de volumes anuais acumulados



Fonte: Dados da Pesquisa – Elaborado pela autora

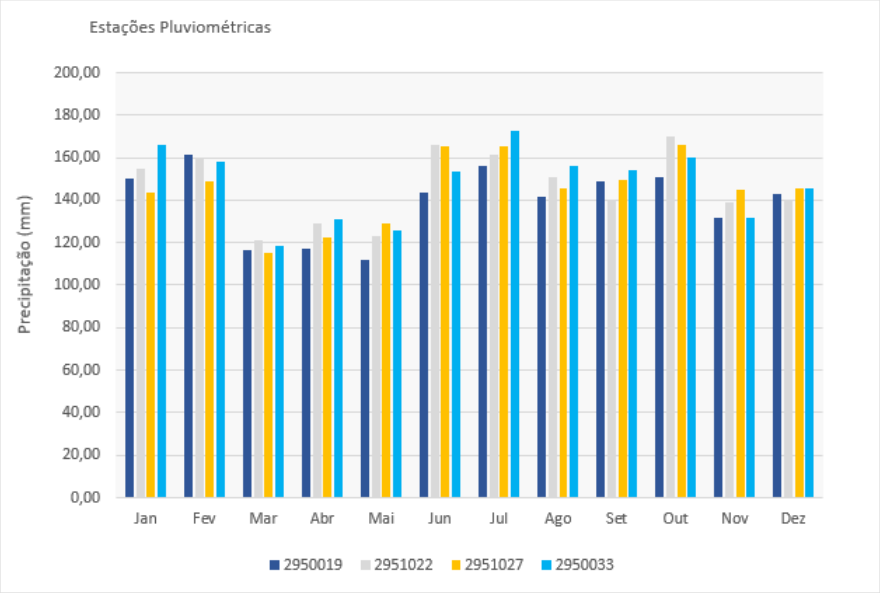
Figura 21: Distribuição de ocorrência de níveis anuais máximos



Fonte: Dados da Pesquisa – Elaborado pela autora

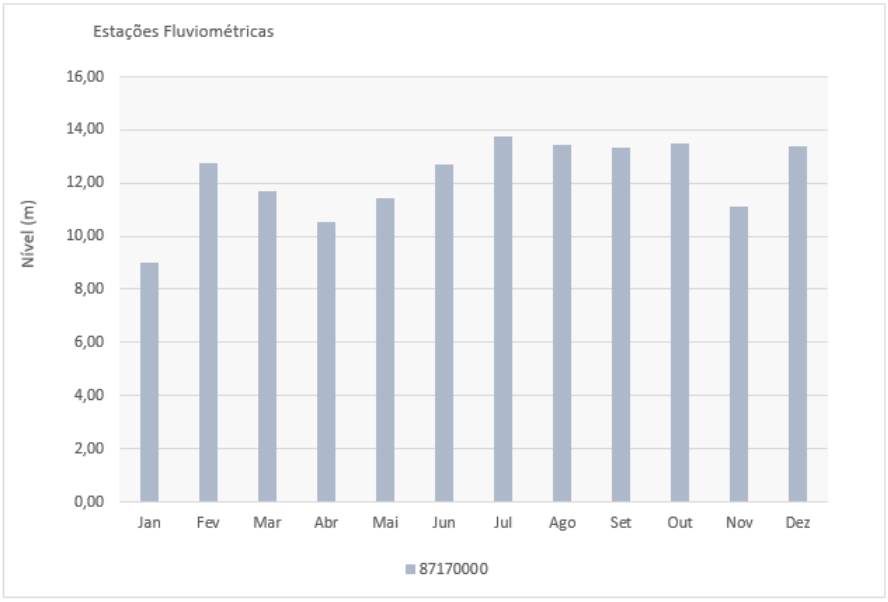
Observando-se o comportamento da precipitação ao longo dos meses, percebe-se que junho, julho e outubro apresentam os maiores em volumes de precipitação mensal, seguidos por janeiro e fevereiro e, na análise do nível do rio Caí, observa-se comportamento semelhante, maiores cotas registradas entre os meses de julho a outubro, além de dezembro e fevereiro, conforme a Figura 22 e a Figura 23.

Figura 22: Precipitação média mensal



Fonte: Dados da Pesquisa – Elaborado pela autora

Figura 23: Nível máximo mensal



Fonte: Dados da Pesquisa – Elaborado pela autora

A fim de compreender a distribuição dos dados de cada uma das estações foram obtidas medidas estatísticas, aplicadas às precipitações máximas anuais e ao nível máximo anual, utilizando o *software* R (versão 4.4.1) e seu ambiente de desenvolvimento integrado R Studio. Nesse sentido, foram geradas medidas descritivas de tendência central, como a média e a mediana, medidas de dispersão, como a obliquidade, curtose, coeficiente de variação e desvio

padrão, além da análise de distribuição de probabilidade dos conjuntos de dados (indicado pelo *p-valor),* sendo o resultado dessas medidas[10] apresentados na Tabela 1.

Analisando-se os valores da Tabela 1, nota-se que os dados apresentam certo desequilíbrio em relação à média, demonstrando tratar-se de distribuições assimétricas, o que também é indicado pelos valores das medianas diferentes das médias, já que em uma distribuição normal (simétrica) esses valores estão ambos localizados no centro da distribuição (FARIAS, SOARES e CÉSAR, 1998). A medida de obliquidade comprova essa afirmação e indica que os dados de precipitação máxima apresentam uma assimetria enviesada à direita, também indicado pelos valores de média maiores que as medianas, enquanto os dados de nível possuem assimetria à esquerda, representado pelo valor negativo da obliquidade e corroborado pelo valor da mediana superior ao da média. Buscando seguir na análise quanto às suas distribuições, calculou-se a curtose, cujos valores evidenciam que os dados de precipitação apresentam menor grau de achatamento (leptocúrtica) em relação à curva normal, estando mais concentrados em relação à média (três é o valor de referência para dados que seguem uma distribuição normal[11]), sendo que os conjuntos de dados das estações 2951027 e 2950033 são os que apresentam menor grau de achatamento de suas curvas.

Tabela 1. Resumo da análise exploratória dos dados de precipitação e nível máximos do rio Caí

Estação	Mínimo	1° Quartil	Mediana	Média	3° Quartil	Máximo	Desvio Padrão	Coeficiente de variação (%)	p-valor	Obliquidade (skewness)	Curtose
2950019	38,00	72,10	79,60	82,85	93,00	129,40	22,04	26,61	0,0460	0,4646	3,1882
2951022	60,50	74,55	91,00	93,16	102,25	168,70	24,29	26,08	0,0016	1,3031	4,4917
2951027	50,20	73,35	92,40	98,22	114,70	217,30	32,54	33,13	0,0015	1,4863	6,1855
2950033	51,50	69,75	75,30	81,32	90,50	159,30	21,57	26,53	0,0004	1,6645	6,5286
87170000	825,00	1.099,00	1.171,00	1.163,77	1.248,00	1.374,00	125,96	10,52	0,5821	-0,4235	3,0594

Fonte: Dados da Pesquisa – Elaborado pela autora

[10] Para a geração das medidas descritivas de tendência central (média e mediana) e de distribuição (máximos, mínimos, 1° quartil e 3° quartil) utilizou-se a função *summary* disponível em: https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/summary. - Para a obtenção da medida de obliquidade e curtose foram utilizadas respectivamente a função *skewness* e *kurtosis (*pacote *moments)* disponível em: https://cran.r-project.org/web/packages/moments/moments.pdf - Para obtenção do desvio padrão foi obtida a variância dos dados, aplicando-se a função *var* e posteriormente a função *sqrt* (ambas nativas do R) para extração da raiz quadrada da variância. Para obtenção do p-valor foi utilizado o teste de *Shapiro-Wilk* disponível em https://search.r-project.org/CRAN/refmans/psyntur/html/shapiro_test.html

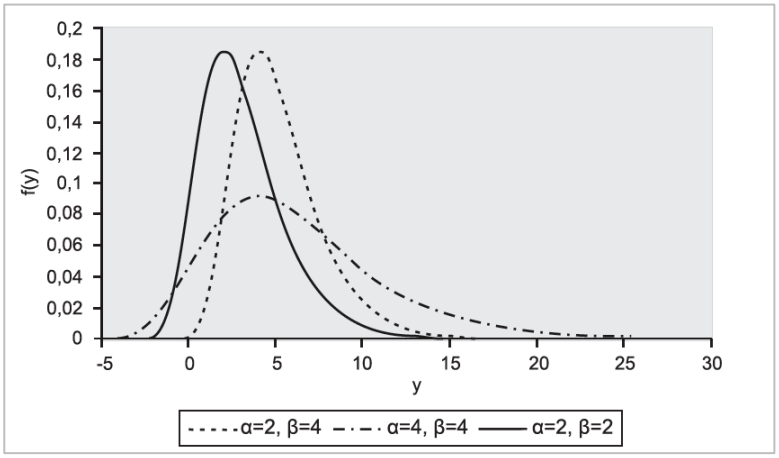
[11] Uma distribuição normal tem curtose igual a 3. Documentação sobre curtose disponível em: https://search.r-project.org/CRAN/refmans/AMR/html/kurtosis.html

A curtose aplicada aos dados máximos de nível, apresenta um valor muito próximo a três e, apesar de assimétrico, esse conjunto de dados, quando comparado com as chuvas máximas, apresenta um grau de achatamento mais próximo da curva normal, sugerindo uma tendência à normalidade da distribuição. Essas dispersões em relação à média são também indicadas pelo coeficiente de variação, a partir do qual é possível observar que os dados de nível têm a menor variação em relação à média, em torno de 10,5%, enquanto os dados de precipitação máximas apresentam uma variação superior a 20%.

Apesar das indicações de assimetria das distribuições, aplicou-se o teste de *Shapiro-Wilk* com o objetivo de confirmar se os dados seguiriam uma distribuição normal. Conforme os p-valores apresentados na Tabela 1, apenas os dados de nível seguem uma distribuição normal, uma vez que o p-valor resultante foi superior a 0,05 enquanto os resultados para os dados de precipitação foram todos inferiores a esse valor de referência, indicando que não seguem uma distribuição normal. Cabe salientar que essa característica para precipitações máximas é esperada, pois valores extremos tendem a seguir distribuição de Gumbel, sendo esta citada como a mais utilizada para análise de frequência de variáveis hidrológicas (NAGHETTINI e PINTO, 2007). A Figura 24 ilustra o gráfico da função de densidade de Gumbel para máximos[12] e a Figura 25 traz os gráficos de densidade referente aos dados de cada uma das estações, a partir da qual é possível avaliar o quanto as precipitações apresentam comportamento semelhante à função densidade de Gumbel. Os gráficos apresentados na Figura 25

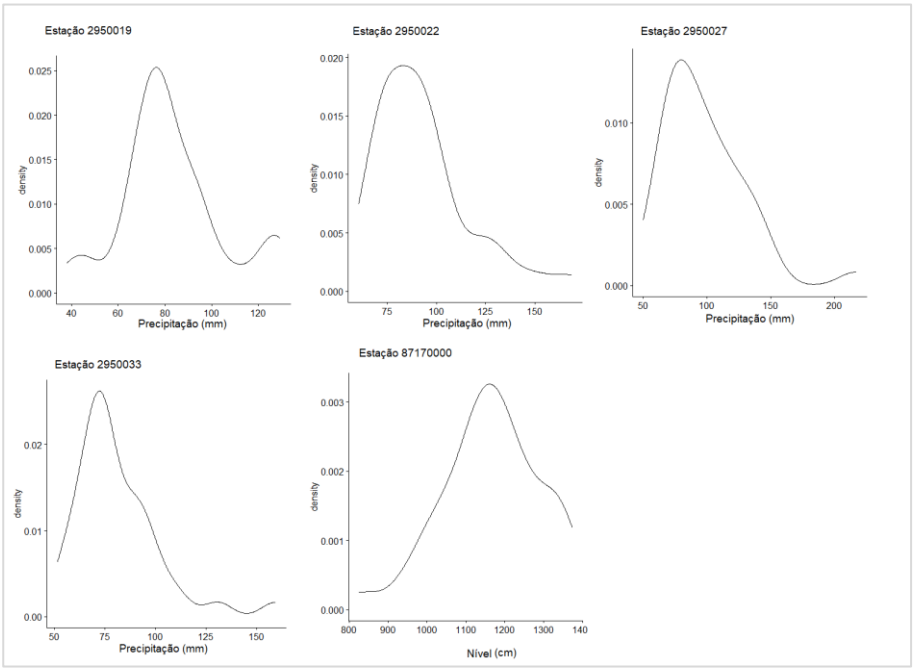
[12] Gráficos elaborados utilizando a função *ggdensity* do pacote *ggpubr.* Documentação disponível em: https://www.rdocumentation.org/packages/ggpubr/versions/0.6.0/topics/ggdensity

Figura 24. Função de densidade de Gumbel para valores máximos



Fonte: NAGHETTINI e PINTO, 2007

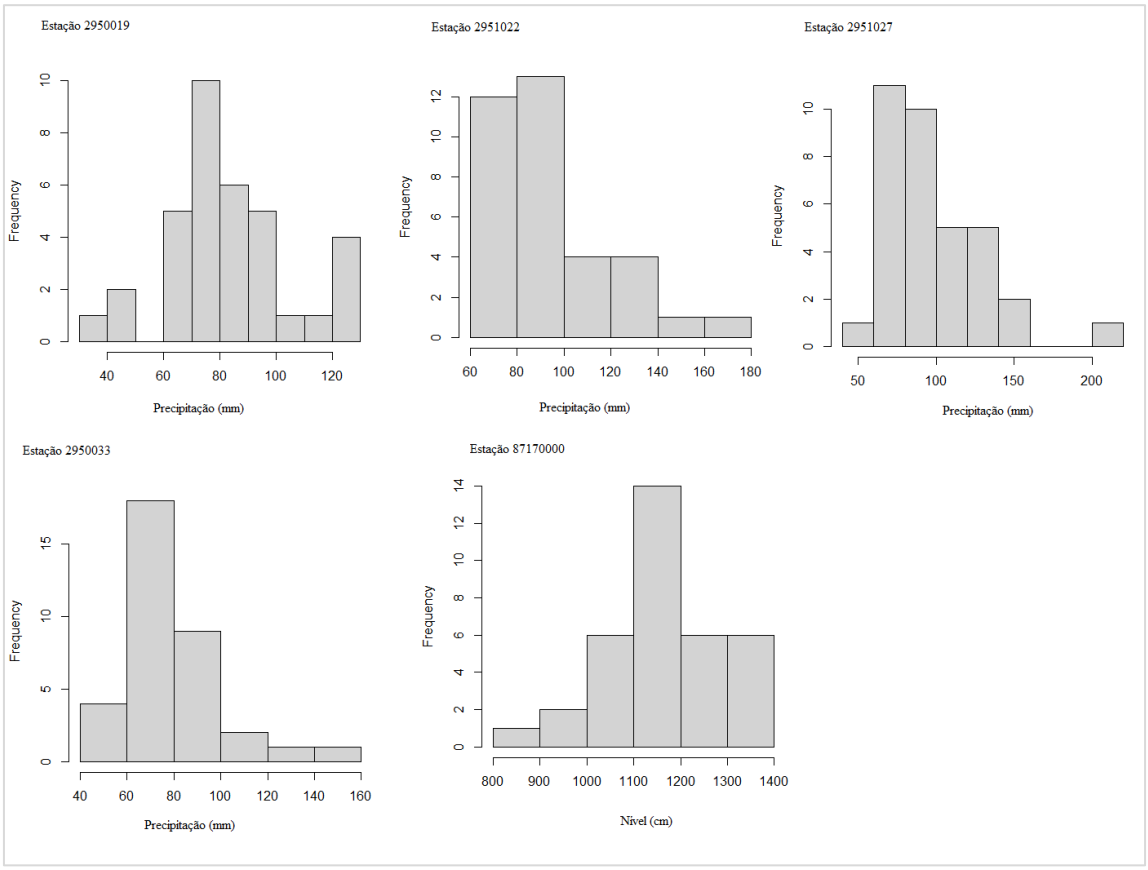
Figura 25. Gráfico de densidade referente às precipitações e nível máximos anuais



Fonte: Dados da Pesquisa – Elaborado pela autora

Ainda, complementar a análise estatística a partir de gráficos foram elaborados os histogramas apresentados na Figura 26, os quais corroboram com a constatação da assimetria dos dados e sua dispersão em relação ao valor central. Para a elaboração dos histogramas foi utilizada a função *hist.* nativa do R.

Figura 26. Histogramas de frequência das precipitações e do nível fluviométrico

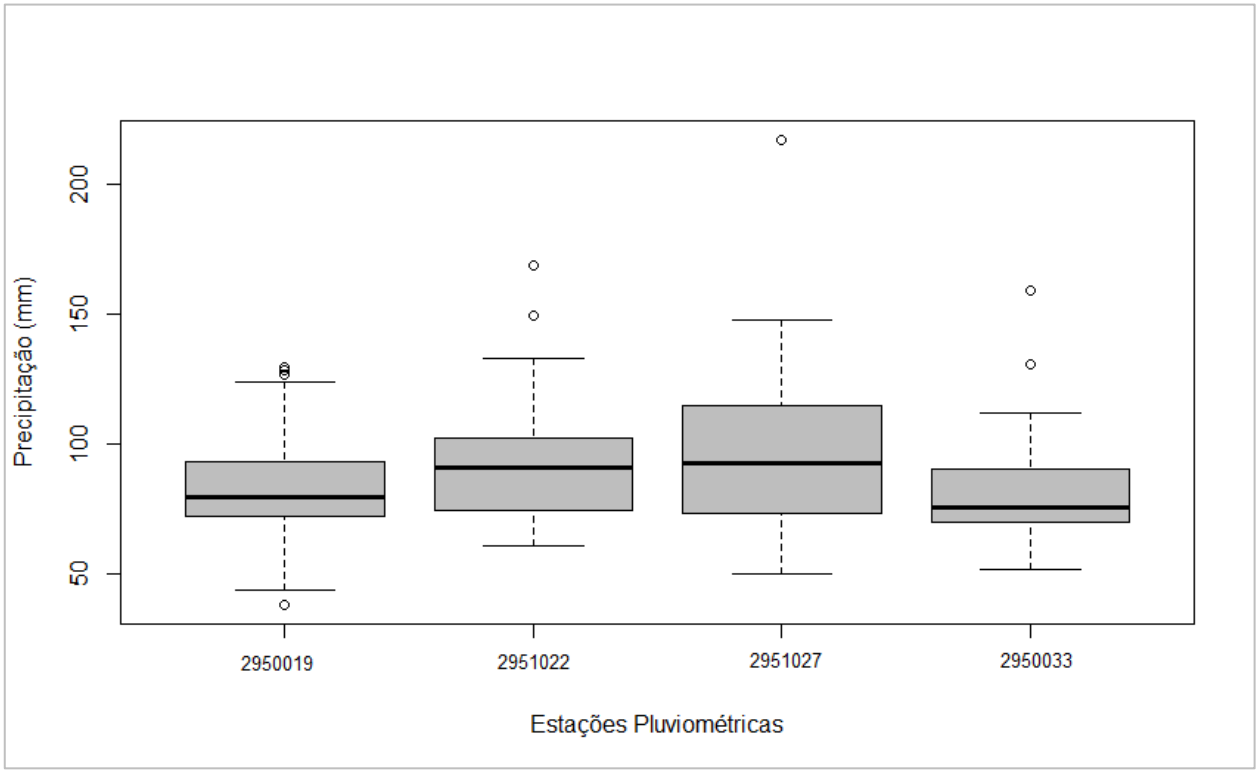


Fonte: Dados da Pesquisa – Elaborado pela autora

Além da análise das medidas de tendência central e de dispersão, foi realizada a identificação de *outliers* que podem ser analisados a partir dos gráficos do tipo Boxplot[13], os quais são ilustrados nas Figura 27 e Figura 28. Verifica-se que todos os conjuntos de dados das estações apresentam valores extremos, sendo que apenas o conjunto de dados de nível indicou existência de *outliers* abaixo do limite inferior. A análise desses valores será tratada na próxima seção, referente à terceira etapa deste trabalho e que representa a terceira etapa, na qual serão abordados, além dos *outliers,* os dados ausentes e seu tratamento.

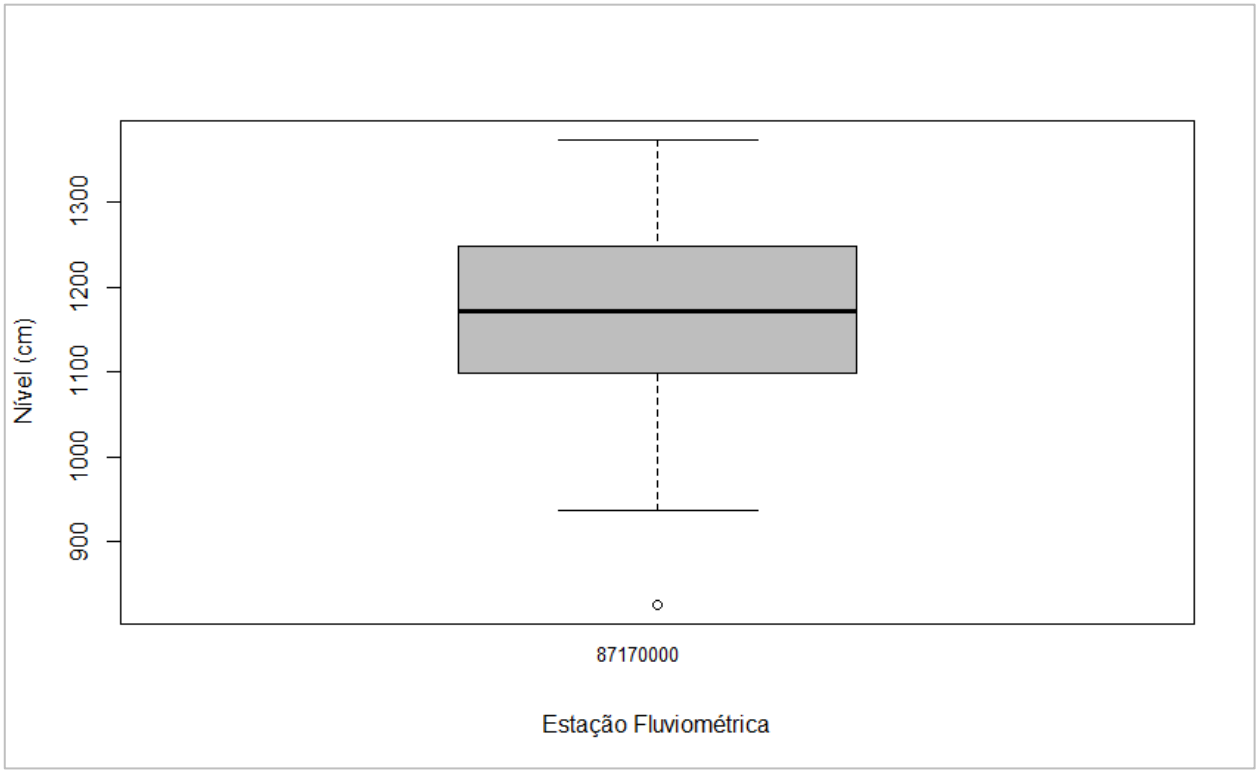
[13] Foi utilizada a função *geom_boxplot* da biblioteca *ggplot* do R. Documentação disponível em: https://www.rdocumentation.org/packages/ggplot2/versions/3.5.0/topics/geom_boxplot

Figura 27. Boxplot referente aos dados de precipitações



Fonte: Dados da Pesquisa – Elaborado pela autora

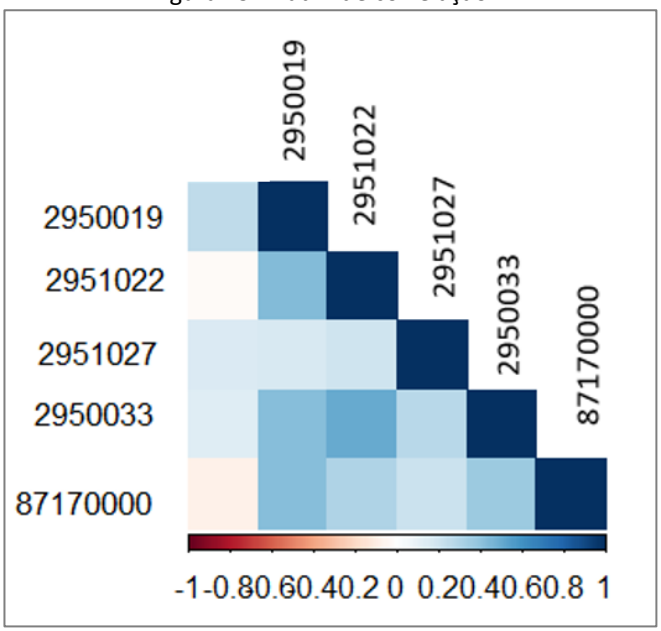
Figura 28. Boxplot referente aos dados de nível do rio Caí



Fonte: Dados da Pesquisa – Elaborado pela autora

Por fim, a Figura 29 traz a matriz de correlação das variáveis precipitação e nível indicando que inexiste correlação negativa, ou seja, à medida que uma variável aumenta seus valores a outra também aumentará. Avaliando a referida Figura, quanto mais forte a correlação, mais escuro o tom de azul. Os dados de precipitação da estação 2950019 foram os que apresentaram correlação mais forte com os dados de nível. Para elaboração dessa matriz foram utilizadas as funções *cor* e *corrplot* do R.

Figura 29. Matriz de correlação



Fonte: Dados da Pesquisa – Elaborado pela autora

4.3 Tratamento dos dados

A presente seção irá abordar a etapa de tratamento dos dados, dedicando-se à análise de dados ausentes, à análise de *outliers* e à transformação dos dados. Primeiramente foram analisados os conjuntos de dados de cada uma das estações pluviométricas e da estação fluviométrica buscando identificar quais possuem falhas nas séries temporais, conforme apresentado na Tabela 2. Observando-a, constata-se que apenas as estações pluviométricas 2951022 e 2951027 apresentam ausência de dados. O número de dados faltantes tem baixa representatividade em relação ao total de cada um dos conjuntos, representando menos de 0,5% na primeira e 1,1% na segunda. Na estação 2951022 os dados ausentes correspondem a dias com falhas no registro de chuva no mês de março de 1997. Já na estação 2951027 apresentou maior número de dias com falhas, ocorridas em dezembro (30 dias) de 1988, em abril (30), maio (30),

setembro (30) e outubro (14) de 1991, setembro (3) e outubro (4) de 2003 e janeiro (2) e fevereiro (1) de 2004.

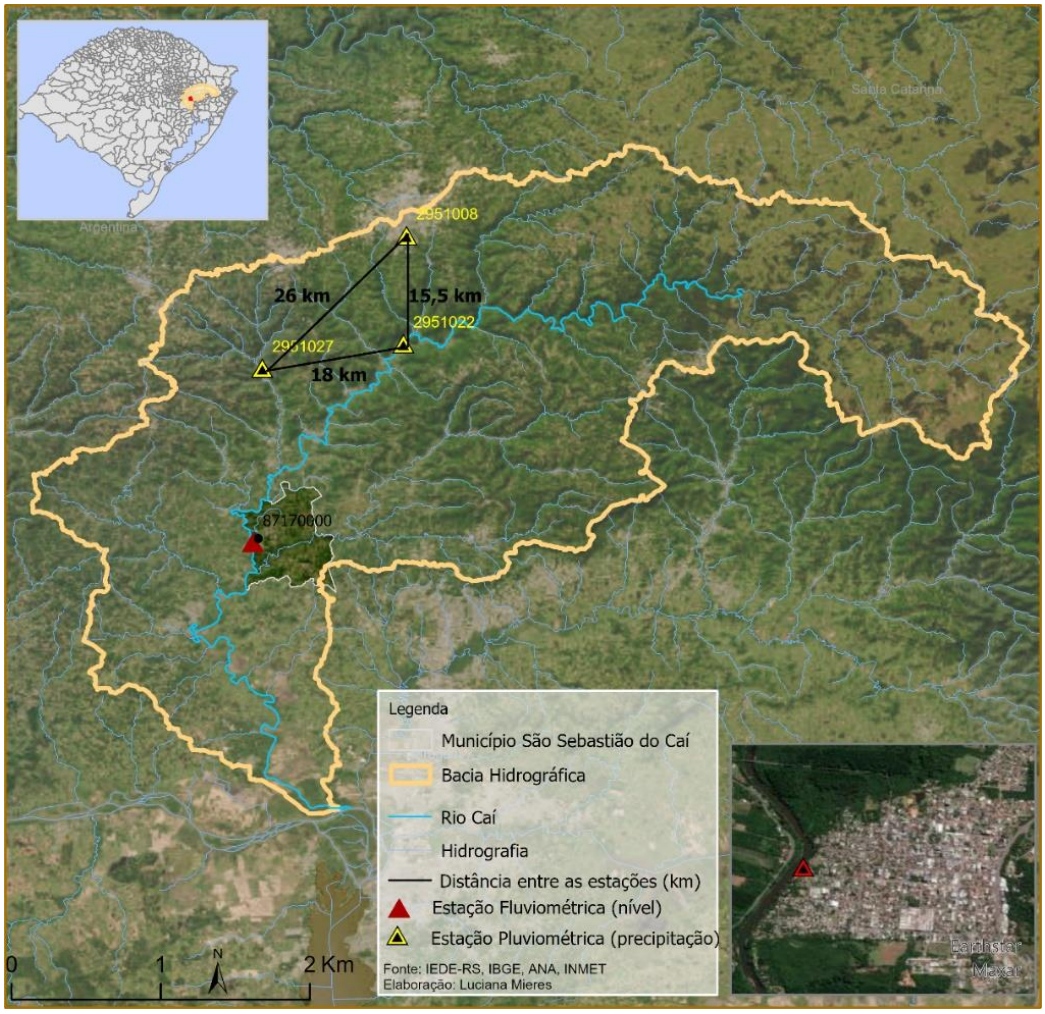
Tabela 2: Resumo dos dados faltantes nas séries temporais por estação

Estação	Categoria do dado	Total de dias referente ao período de análise	Total de dias com ausência de dado (Código -1)	Proporção de dias sem dados em relação ao total de dias (%)
2950019	Precipitação (mm)	12.784	0	0,0
2950033	Precipitação (mm)	12.784	0	0,0
2951022	Precipitação (mm)	12.784	25	0,2
2951027	Precipitação (mm)	12.784	144	1,1
87170000	Nível do rio (cm)	12.784	0	0,0

Fonte: Dados da Pesquisa – Elaborado pela autora

Buscando estabelecer uma série temporal mais completa, realizou-se a atribuição dos valores ausentes por valores que pudessem representar a precipitação provável no dia. Para tanto, aplicou-se uma técnica de interpolação utilizada na hidrologia, a qual é baseada no cálculo da média ponderada pelo inverso da distância, em que se considera que a precipitação em uma estação pluviométrica pode ser calculada como a média ponderada das precipitações registradas em estações próximas (COLLISCHONN e DORNELLES, 2013). Para tanto foi necessário acessar os dados de uma nova estação, próxima tanto à estação 2951022 quanto à 2951027. A Figura 30 ilustra a localização dessa estação (código 2951008) utilizada para auxiliar no tratamento dos dados faltantes. A escolha dessa estação foi baseada na completude de dados no período necessário para a correção dos dados faltantes.

Figura 30. Localização da estação complementar utilizada para aplicação da técnica de atribuição de valores.



Fonte: IEDE-RS, IBGE e ANA – Elaborado pela autora

Para a atribuição dos valores de precipitação aplicou-se a Equação 8, baseada em Colischon e Dorneles (2013), onde *Pm* é a chuva no dia, *NP* é o número de postos pluviométricos com dados disponíveis, P_j é a chuva observada na estação *j* e *b* um expoente igual a 2, o que identifica o método como interpolação ponderada pelo inverso da distância ao quadrado.

$$Pm = \frac{\sum_{j=1}^{NP} \frac{P_j}{(d_{ij})^b}}{\sum_{j=1}^{NP} \frac{1}{(d_{ij})^b}} \qquad \text{Equação 8}$$

Como os períodos de falhas não coincidem entre as estações em questão, para a proposição de valores para o período de falha da estação 2951022 aplicou-se a Equação 8, utilizando-se os dados diários de chuva das estações 2951027 e 2951008, além de considerar as

distâncias destas estações em relação à 2951022. O mesmo procedimento foi aplicado para a proposição de valores para a estação 2951027, sendo tudo realizado no *Microsoft Excel*. A partir da aplicação da técnica de interpolação, uma nova base de dados foi obtida, sem valores ausentes para dados de chuva.

Como apresentado na Figura 11, além da etapa de Tratamento dos dados, estava prevista a análise dos *outliers.* Conforme citado na seção anterior, foram identificados *outliers* em todos os conjuntos de dados, porém algumas estações de precipitação não apresentaram valores mínimos de *outlier,* enquanto a fluviométrica não apresentou valor máximo. A Tabela 3 traz a especificação dos *outiers* encontrados e respectivos anos de observação. Para identificação desses valores calculou-se os limites inferiores e superiores de cada conjunto de dados utilizando o método da amplitude interquartil (IQR) e definidos esses limites, todos os valores acima do superior ou abaixo do inferior são considerados *outliers.* Em uma breve análise, identifica-se que todos os valores de precipitação que superaram o limite máximo ocorreram em anos de *El Niño*, variando apenas entre anos em que o fenômeno foi classificado como fraco (1993), moderado (1980, 2002 e 2003), ou forte (1982). Já os valores mínimos identificados, ambos ocorreram em anos normais sem influência de fenômenos ENOS[14]climáticos.

Tabela 3. Identificação dos *outliers* presentes nas estações

Estação	Categoria	*Outlier* mínimo	Ano (mínimo)	*Outlier* máximo	Ano (máximo)
2950019	Precipitação	38,00	1978	126,8 128,4 129,4	2002 1980 2003
2951022	Precipitação	Não apresenta	-	149,6 168,7	1980 1982
2951027	Precipitação	Não apresenta	-	217,30	1993
2950033	Precipitação	Não apresenta	-	130,9 159,3	1982 2003
87170000	Nível	825,00	2004	Não apresenta	-

Fonte: Dados da Pesquisa – Elaborado pela autora

[14] Segundo INPE, o fenômeno El Niño Oscilação Sul (ENOS) refere-se às situações nas quais o oceano Pacífico Equatorial está mais quente (El Niño) ou mais frio (La Niña) do que a média normal histórica. A mudança na temperatura do oceano Pacífico Equatorial acarreta efeitos globais na temperatura e precipitação. Anos de El Niño apresentam anomalias no aumento da precipitação e temperaturas, enquanto nos anos de La Niña, observa-se a diminuição de ambos. Disponível em: http://enos.cptec.inpe.br/ Consulta em 23 de setembro de 2024.

Considerando o objetivo do presente estudo, de utilizar variáveis de chuva e nível para predição da variável nível do rio, a ideia central que inspirou este trabalho, que é avaliar a aplicação do modelo LSTM para predição de nível como ferramenta auxiliar na antecipação dos níveis críticos responsáveis por eventos de inundação, destaca-se que os valores mínimos extremos não colaboram com esse objetivo, uma vez que eventos de inundação decorrem justamente de valores máximos de níveis do rio, sendo aqueles extremos máximos, por vezes, responsáveis por desastres de inundação. Por fim, tendo em vista que se tratam de poucos casos de *outliers* com valores mínimos, optou-se por mantê-los nos respectivos conjuntos de dados, assim como *outliers* com valores máximos, mantidos sobretudo por sua importância quando avalia-se o potencial de impactarem em níveis mais elevados do rio que possam resultar em inundação.

4.4 Desenvolvimento do modelo de predição

Nesta seção será abordada a etapa de desenvolvimento do modelo de predição, na qual buscou-se estabelecer os melhores valores para os hiperparâmetros ajustando-se, assim, o melhor modelo. Os resultados foram avaliados a partir de medidas indicadas para problemas de regressão e também indicadas para análise de variáveis hidrológicas e serão apresentados na última etapa do processo, que consiste na avaliação do modelo (a ser abordado na próxima sseção). Cabe relembrar que o modelo LSTM consiste em uma rede neural recorrente, sendo bastante utilizado no processamento de séries temporais. A abordagem teórica acerca da rede neural foi tratada no Capítulo 2.

Para o desenvolvimento do modelo utilizado nesta pesquisa, optou-se por adotar o ambiente virtual *google colaboratory*[15], disponibilizado através do navegador Chrome da empresa *Google.* A escolha por esse ambiente embasou-se na facilidade do acesso à máquina virtual com disponibilização do processamento baseado na GPU[16] e a linguagem de programação

[15] Google Colaboratory disponível em: https://colab.google/

[16] O processamento foi realizado utilizando a opção T4 GPU, disponibilizada em máquina virtual da Google. Disponível em: https://cloud.google.com/compute/docs/gpus?hl=pt-br#:~:text=GPUs%20NVIDIA%20T4,-No%20entanto%2C%20as&text=*-,A%20mem%C3%B3ria%20da%20GPU%20%C3%A9%20a%20mem%C3%B3ria%20dispon%C3%ADvel%20em%20um,com%20uso%20intensivo%20de%20gr%C3%A1ficos.

utilizada foi *Python*[17]. Ressalta-se que antes de partir para o desenvolvimento do modelo, é necessário realizar o pré-processamento dos dados, etapa fundamental para estabelecer-se os conjuntos de treino e teste, a normalização dos dados e, por fim, a adequação à estrutura esperada pela rede LSTM. A seguir serão detalhadas cada uma dessas atividades, destacando-se que foram considerados como dado de entrada tanto os dados de chuva quanto de nível para o treinamento da rede.

Primeiramente, organizou-se os conjuntos de dados em um arquivo único do tipo .CSV (valores separados por vírgula), composto por cinco colunas, das quais quatro referiam-se aos dados de precipitação (expressos em milímetros) e uma aos dados de nível (expressos em centímetros). O arquivo, composto por 12.784 linhas, contempla um total de 63.920 valores. Esse arquivo foi salvo em uma pasta de trabalho criada diretamente no *google drive*, possibilitando assim, o acesso através do ambiente *Colaboratory*.

O referido arquivo foi carregado em uma variável denominada *dataset*, a qual foi submetida à divisão em conjunto[18] de treinamento (denominado *train*) contendo 70% dos dados do conjunto original e de teste (denominado *test*) contendo os 30% restantes. Salienta-se que a opção por formar conjuntos de dados de treino e teste de forma sequencial e não utilizando alguma técnica randômica que estabelecesse os melhores conjuntos com base em seus valores, decorreu da necessidade de manter preservada a sequência dos dados considerando tratar-se de uma série temporal.

Posteriormente, foi realizada a normalização[19] de cada um dos conjuntos que consiste em transformar os valores para que fiquem em um intervalo pré-definido, definindo-se o intervalo entre zero e um já que não existem valores negativos. Esse processo é fundamental considerando que os dados de chuva e de nível possuem unidades de medidas diferentes. Normalizando-os, passam os valores a estar na mesma escala para o processamento da rede neural preservando os *outliers*. A supressão da etapa de normalização impactaria na determinação dos pesos

[17] Documentação disponível em: https://www.python.org/downloads/

[18] Para determinação dos conjuntos, definiu-se que o conjunto de treinamento (*train)* iniciaria na linha 0 e finalizaria na linha que representasse os 70% do conjunto total (*dataset*) e o conjunto de teste (*test*) iniciaria a partir da primeira linha posterior à linha final do conjunto de treino, estendendo-se até o final do *dataset.*

[19] Para normalização dos conjuntos de treino e de teste foi utilizada a função *MinMaxScaler* da biblioteca *Scikit learn.* A documentação pode ser acessada em: https://scikit-learn.org/1.5/modules/generated/sklearn.preprocessing.MinMaxScaler.html

efetuado pelo modelo durante o processo de treinamento, uma vez que variáveis que possuam valores maiores por sua natureza (caso dos dados de nível) poderiam indicar para o modelo que possuem maior relevância e assim, receberiam maior atribuição de peso. A normalização coloca todos os valores na mesma ordem de grandeza e a função utilizada nessa etapa pode ser expressa pela Equação 9 a seguir.

$$x_{(normalizado)} = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad \text{Equação 9}$$

Estabelecida a normalização dos conjuntos de treino e de teste, foi necessário adequar a matriz de entrada ao formato exigido pela rede neural LSTM. Até o momento esses dados estavam organizados em uma matriz de duas dimensões (2D), composta por amostras (*sample)* nas linhas, e por colunas (*features)* que representam os atributos de chuva e nível. O conjunto de treinamento consistia em uma matriz composta por um total de 8.946 linhas e cinco colunas, enquanto o conjunto de teste continha 3.828 linhas e cinco colunas. A transformação da estrutura 2D em uma matriz com três dimensões (3D) consiste basicamente em definir, além do número de amostras e colunas, o número de passos por amostra (*time step*), sendo que esse valor irá definir o número de amostras (*samples)* que serão consideradas para realizar a previsão do nível. Desse modo, a estrutura 3D dos conjuntos de treinamento e de teste variou conforme os passos definidos para avaliação da rede neural. Salienta-se que a definição desses passos foi embasada no tempo de concentração da bacia hidrográfica em estudo, cujo tempo máximo é de 7 dias. Portanto, utilizou-se os passos de 2, 5 e 7 dias para avaliação do LSTM na obtenção de dados de nível. Para a reestruturação dos conjuntos em uma matriz 3D aplicou-se uma função[20] que recebe como parâmetros o conjunto de dados (tanto de treino, quanto de teste) e o número de passos (*time step*), sendo utilizado o mesmo número de passos para cada conjunto de dados.

Com os dados de entrada organizados em uma matriz 3D, o passo seguinte foi a estruturação do modelo LSTM objetivando a avaliação de seus hiperparâmetros com o objetivo de encontrar o melhor modelo a ser posteriormente utilizado para a predição dos dados de nível.

[20] Para essa função foi utilizada a biblioteca *Numpy*, especificamente as funções *array, ndarray, append.* A estrutura da função foi adaptada ao contexto em estudo, a partir do script utilizado pela Professora Roseli Francelin Romero na disciplina de Redes Neurais e Deep Learning - Fundamentos que faz parte da estrutura curricular do MBA em Inteligência Artificial e Big Data vinculado ao Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (USP). As bibliotecas citadas podem ser encontradas em: https://numpy.org/doc/stable/index.html

Os hiperparâmetros considerados foram: 1) *Units:* número de neurônios da camada LSTM; 2) *Batch size:* número de amostras processadas antes que os pesos fossem atualizados; e 3) *Epochs:* número de passagens completas por todo o conjunto de dados de treinamento. Além disso, foi considerada na camada densa do modelo a função de ativação *sigmoide* aplicada para garantir não-linearidade, característica dos dados em estudo. Para agilidade na determinação do melhor modelo foi desenvolvida uma classe[21] composta por funções que implementam o modelo em si e testam os hiperparâmetros mencionados variando seus valores de acordo com os intervalos apresentados na Tabela 4, sendo o melhor modelo aquele cuja combinação de valores resultou em um menor valor para a função de perda para a validação (*val_loss*[22]).

Tabela 4: Hiperparâmetros e respectivos valores testados para identificar o melhor modelo a ser utilizado na predição dos valores de nível

Hiperparâmetro	Valores testados	Melhores valores (utilizados no modelo de predição)
Units	4 a 78	20
Batch Size	8, 16, 32	8
Epochs	50 - 100	50

Fonte: Dados da Pesquisa – Elaborado pela autora

Como camada de entrada (*input_shape)* foram utilizadas as matrizes 3D de treinamento (denominada *X_train*) e de teste (*X_test*) sendo adotado o otimizador *Andam*[23]. Como medida de avaliação da função de perda foi definido o Erro Quadrático Médio (*Mean Squared Error - MSE),* indicado para avaliar a acurácia de modelos que buscam resolver problemas de regressão[24], ressaltando que sua avaliação se baseia na diferença entre os valores reais (observados) e os preditos (simulados), conforme apresenta a Equação 10.

[21] Para o desenvolvimento da classe e suas respectivas funções foram consideradas as funções *layer* e *turner* ambas do *Keras.* Documentação disponível em: https://keras.io/guides/keras_tuner/getting_started/ e https://keras.io/api/keras_tuner/hypermodels/

[22] *Val_loss:* é a função de perda da validação tem o objetivo de monitorar como o modelo está generalizando a predição para dados que ele não viu durante o treinamento. Ela é calculada do mesmo modo que a *Loss.* Esta por sua vez é a métrica que avalia o quão bem o modelo está se ajustando aos dados de treinamento. Nesse estudo a medida adotada é a *mean squared error (mse).* A documentação pode ser acessada em: https://keras.io/api/losses/regression_losses/#meansquarederror-class

[23] *Adam (Adaptive Moment Estimation)* é um otimizador baseado em gradiente que busca ajustar os parâmetros peso e *bias* durante o processo de treinamento, de modo a minimizar a função de perda (*loss).* A documentação pode ser encontrada em https://keras.io/api/optimizers/adam/

$$MSE = \frac{1}{n}\sum_{I=0}^{n-1}(y_{obs.} - y_{predito})^2$$ Equação 10

Findado o treinamento da rede e definido o melhor modelo, o próximo passo foi sua aplicação à predição dos conjuntos de treino e de teste a fim de serem obtidos os valores de nível relacionados ao treino (denominado como *train_pred*) e ao teste (*test_pred*). Por fim, para a análise dos resultados procedeu-se a inversão da normalização trazendo os resultados preditos para a escala real dos dados de nível (em centímetros), facilitando a comparação com os dados observados para o período em análise. Outra medida utilizada na avaliação dos resultados foi a Raiz do Erro Quadrático Médio (*Root Mean Square Error - RMSE),* que consiste na raiz quadrada do MSE, conforme Equação 12.

$$RMSE = \sqrt{\frac{1}{n}\sum_{I=0}^{n-1}(y_{obs.} - y_{predito})^2}$$ Equação 11

Ressalta-se, ainda, que além das medidas supracitadas, foi também utilizado o Coeficiente de Nash-Sutcliffe (*Nash-Sutcliffe Efficiency - NSE*), medida comumente adotada na avaliação previsões de variáveis hidrológicas. Esse coeficiente também avalia a diferença entre os valores observados e os preditos, conforme exemplifica a Equação 12.

$$NSE = 1 - \frac{\sum_{i=1}^{n}(y_{obs.} - y_{predito})^2}{\sum_{i=1}^{n}(y_{obs.} - y_{média\ obs.})^2}$$ Equação 12

Convém destacar que os valores e as respectivas interpretações para o coeficiente de Nash-Sutcliffe são apresentados na Tabela 5 (LUFI e RISPININGTATI, 2020).

[24] A aplicação da rede neural LSTM também foi baseada na biblioteca *Keras.* Documentação disponível e: https://keras.io/api/layers/recurrent_layers/lstm/

Tabela 5. Critério de valores para Coeficiente de Nash-Sutcliffe (NSE)

Valor NSE	Interpretação
NSE > 0,75	Bom
0,36 < NSE < 0,75	Qualificado
NSE < 0,36	Não qualificado

Fonte. LUFI e RISPININGTATI, 2020 – Elaborado pela autora

A próxima seção abordará a etapa final, que consiste na avaliação do modelo com base nas medidas MSE, RMSE e NSE.

4.5 Avaliação do modelo

A rede neural foi treinada com base no conjunto de treinamento que corresponde a 70% do total de dados, referindo-se ao período 01/01/1970 a 01/07/1994, enquanto os dados de teste correspondem ao período de 02/07/1994 a 31/12/2004. Para avaliação do desempenho do modelo na obtenção de dados de nível preditos, foram calculadas as seguintes medidas: 1) Erro médio quadrático (MSE); 2) Raiz do erro médio quadrático (RMSE) e; 3) Coeficiente de Nash-Sutcliffe (NSE), descritas e detalhadas na seção anterior. Além dessas medidas, foram analisados os erros percentuais ao longo do período, considerando os resultados em cada um dos passos (*time step*) e assim possibilitando a avaliação de qual conjunto apresentou melhor desempenho. Por fim, foram analisados os valores de nível predito e observado referente ao passo cujos resultados foram melhores.

A Tabela 6 apresenta os valores referentes a cada uma das medidas de erro relacionadas aos distintos passos (*time step)* avaliados. É possível observar que dos três passos utilizados nas simulações, o que apresentou pior resultado foi o *time step* = 2, com maior erro entre o nível predito e o observado, indicado pelo maior valor do MSE. Quando os conjuntos de treino e teste são analisados separadamente, constata-se leve piora da medida relacionada aos dados de teste. Já os resultados referentes aos passos 5 e 7 não apresentaram tanta diferença entre si em relação ao MSE, contudo, ainda assim os resultados referentes ao *time step = 7* apresentaram os menores valores para essa medida.

Outra medida utilizada foi o RMSE, que consiste na raiz do MSE, e da mesma forma que a avaliação anterior, o pior resultado encontrado referiu-se ao passo 2 e o melhor ao passo 7. A Tabela 6 apresenta o valor dessa medida para os dados normalizados, mas, avaliando-se o RMSE

para o dado de nível na escala original (em centímetros), constata-se que em relação ao passo 2 a referida medida calculada para o conjunto de dados de forma geral foi de 167cm, identificando-se que 28,7% dos dias os níveis simulados apresentaram esse valor de erro ou valor superior. Analisando-se o passo 5 identifica-se considerável redução desse erro, cujo valor calculado para o conjunto foi 70,1 cm e 14,2% dos dias apresentaram valores de erro iguais ou superiores à medida. Por fim, o passo 7 se destaca por apresentar erro menor ainda, RMSE calculado em 67,7 cm e 12,6% dos dias apresentando valor de erro igual ou superior.

Essas características referentes ao passo 2 repercutiram diretamente no baixo coeficiente de Nash-Sutcliffe, classificando-o como não suficiente, conforme Tabela 5, uma vez que quanto maior for a diferença entre os dados observados e os preditos, pior é o resultado desse coeficiente (quanto mais próximo o resultado for de 1, melhor é o NSE, indicando maior proximidade entre o dado observado e o simulado). Entretanto, quando se avaliam os resultados decorrentes dos passos 5 e 7, percebe-se expressiva melhora do coeficiente, sobretudo o *time step = 7,* cujos valores de Nash foram os mais elevados tanto na análise geral quanto na análise por conjunto de treino e de teste.

A Tabela 6. Resumo das medidas de erro conforme os diferentes passos (*time steps*) adotados

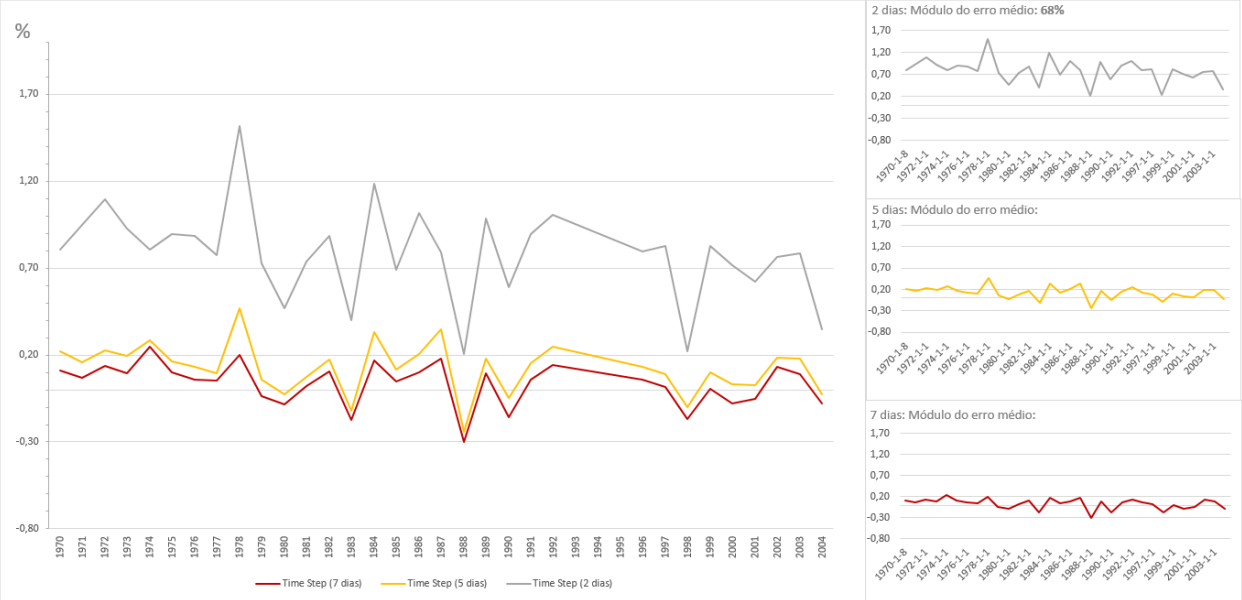
	Geral			Treino			Teste		
Medida de erro	**Predições com *Time Step* 7 dias**	**Predições com *Time Step* 5 dias**	**Predições com *Time Step* 2 dias**	**Predições com *Time Step* 7 dias**	**Predições com *Time Step* 5 dias**	**Predições com *Time Step* 2 dias**	**Predições com *Time Step* 7 dias**	**Predições com *Time Step* 5 dias**	**Predições com *Time Step* 2 dias**
MSE	0,0035	0,0037	0,017	0,0033	0,0035	0,017	0,0039	0,0042	0,018
RMSE	0,059	0,061	0,13	0,057	0,059	0,13	0,062	0,065	0,13
NSE	0,87	0,86	0,26	0,86	0,85	0,26	0,91	0,90	0,25
	Destacados em azul os melhores resultados das medidas analisadas								

Fonte: Dados da Pesquisa – Elaborado pela autora

Corroborando com a análise, a Figura 31 apresenta o percentual de erros encontrados ao longo do período para cada um dos passos adotados e novamente o passo 2 apresentou os piores resultados. Analisando-se o módulo da média do erro para esse passo encontrou-se o valor de 68%, indicando pouca coerência entre os dados observados e preditos. No que tange ao passo 5 o módulo da média do erro foi de 14%, enquanto que para o passo 7 foi de 10%. Analisando-se a Figura 31, observa-se que os resultados obtidos com o passo 2 apresentaram maior

concentração de níveis preditos superiores aos níveis observados, o que é caracterizado pelos valores percentuais positivos, apresentando elevados percentuais de erros. Já os passos 5 e 7 apresentaram resultados mais concentrados entre -20% e + 20%.

Figura 31. Percentual de erro referente à diferença do nível observado e nível predito

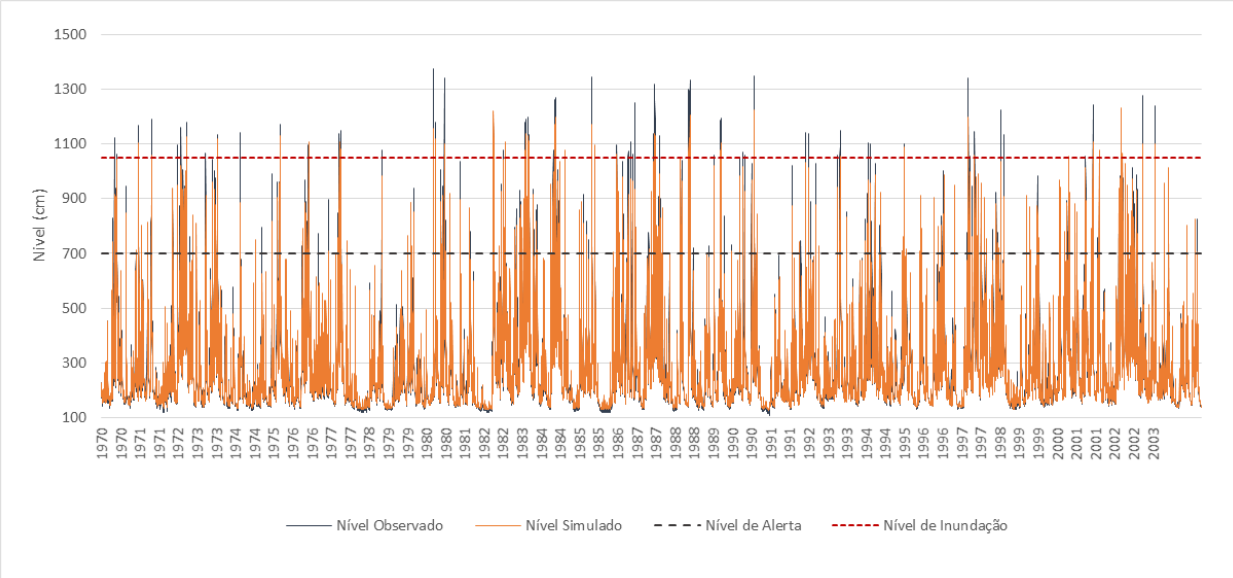


Fonte: Dados da Pesquisa – Elaborado pela autora

Considerando os apontamentos anteriores, definiu-se que os melhores resultados foram obtidos com base nos conjuntos de treinamento e teste embasados no passo de 7 dias. Diante disso, as análises a seguir serão restritas a esse passo. Para auxiliar na análise foram considerados os níveis de alerta e de inundação, ambos calculados pelo Serviço Geológico do Brasil (SGB) para a estação fluviométrica em estudo, considerando a régua de medição (BRASIL, 2022). O nível de alerta, calculado em 700 cm, indica a cota (nível) a partir da qual existe uma grande possibilidade de inundação, enquanto a cota de inundação calculada foi de 1.050 cm.

A Figura 32 ilustra os níveis observados e preditos ao longo dos anos, além de indicar as cotas de alerta e de inundação. Avaliando-se os níveis observados ao longo da série temporal, em relação aos níveis preditos pela rede neural, verifica-se que de modo geral os resultados são promissores, o que é corroborado pelo coeficiente de *Nash-Sutcliffe*, cujo valor calculado foi de 0,87. Contudo, cabe destacar que valores extremos de nível (aqueles acima do nível de alerta e de inundação) apresentam-se de modo subestimado em relação aos observados, algo que deve ser avaliado em estudos futuros.

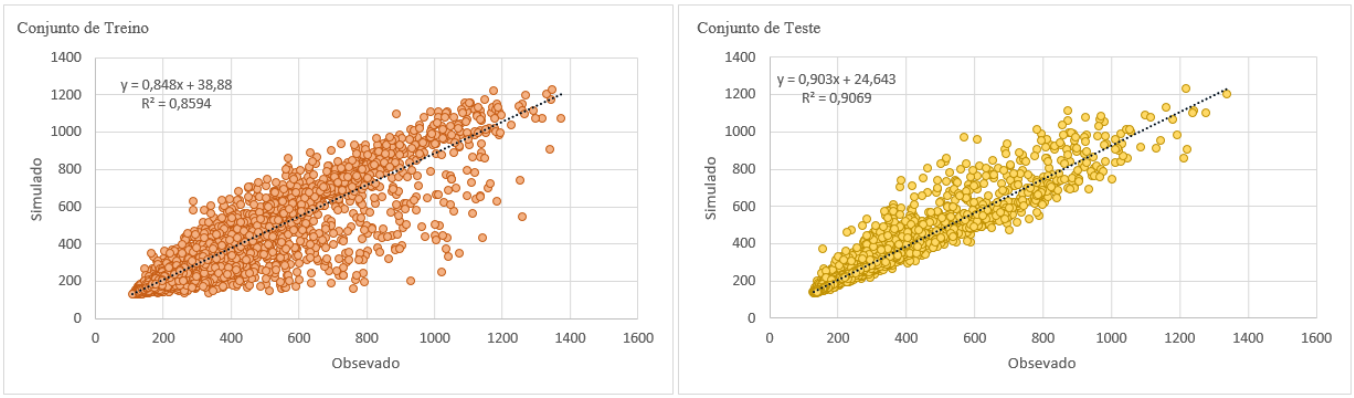
Figura 32. Níveis observados e preditos ao longo da série temporal



Fonte: Dados da Pesquisa – Elaborado pela autora

Com vistas ao entendimento do desempenho do modelo em cada um dos conjuntos, a Figura 33 representa os gráficos de dispersão dos valores de nível preditos em relação aos observados. Percebe-se que o conjunto de teste apresentou melhor desempenho que o de treinamento, descartando-se assim a possibilidade de ter ocorrido *overfitting* ou *underfitting.*

Figura 33. Dispersão dos dados de nível observado e predito para os conjuntos de treinamento e de teste

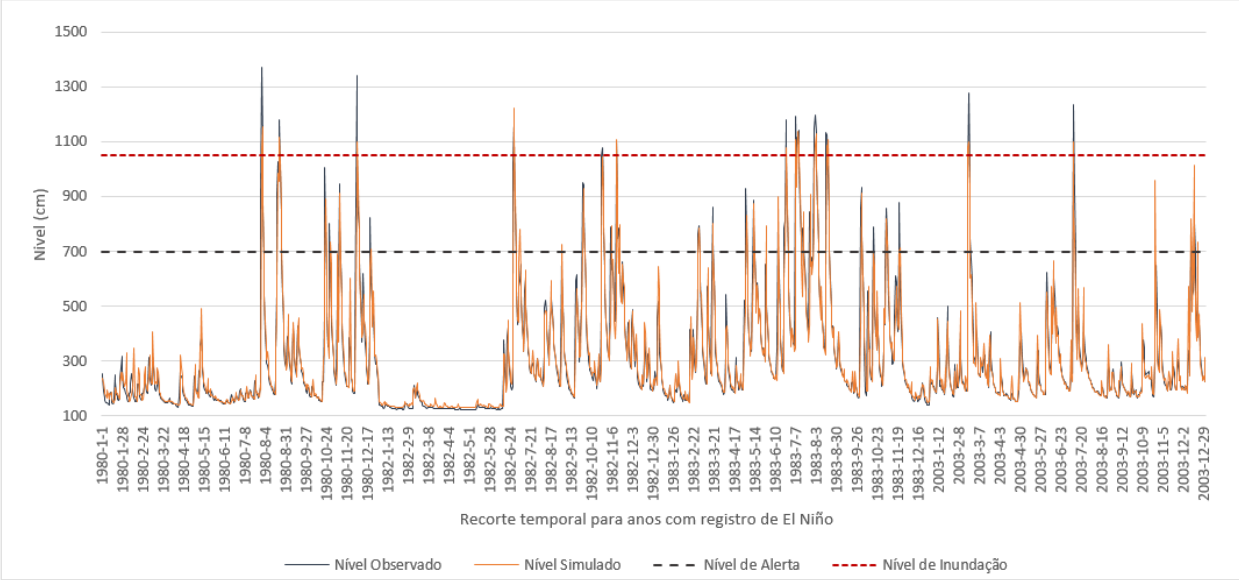


Fonte: Dados da Pesquisa – Elaborado pela autora

Tendo em vista o objetivo deste estudo, realizou-se a análise dos anos em que houve a ocorrência do fenômeno ENOS relacionado ao El Niño, uma vez os *outliers* máximos ocorreram todos nesses anos. De modo geral verifica-se um bom desempenho do modelo para os anos em

questão, contudo, identifica-se que segue subestimando os níveis acima da cota de inundação (Figura 34).

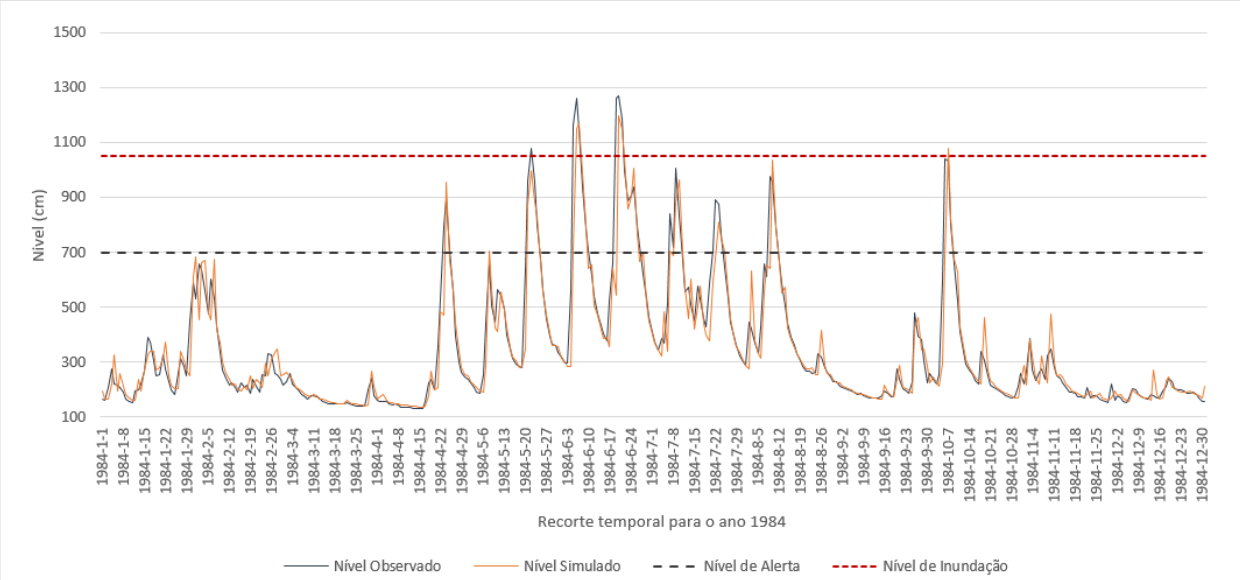
Figura 34. Recorte temporal referente aos anos de El Niño.



Fonte: Dados da Pesquisa – Elaborado pela autora

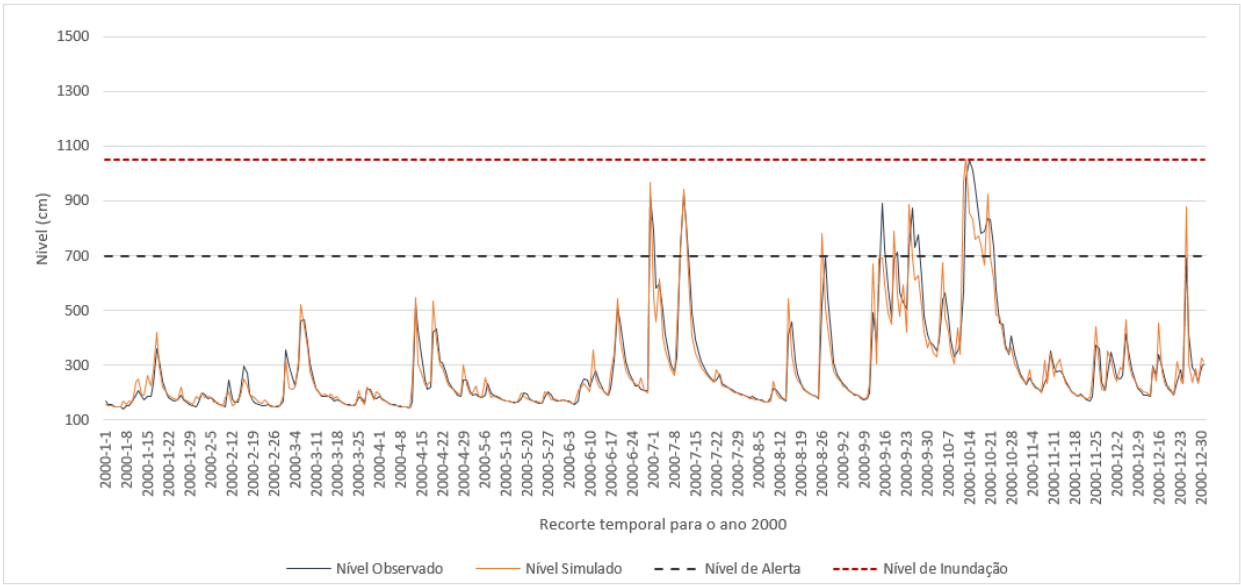
Quando se analisa apenas os dias em que houve registro de nível de alerta (acima de 700 cm), verifica-se que o modelo teve baixo desempenho, apresentando valores simulados inferiores ao nível de alerta, embora tenha acertado o nível de alerta em 87% das vezes em que houve nível de alerta para os dados observados. Esse bom acoplamento pode ser verificado no recorte temporal a seguir, referente ao ano de 1984 (Figura 35) e 2000 (Figura 36)

Figura 35. Nível observado versus predito para o ano completo de 1984



Fonte: Dados da Pesquisa – Elaborado pela autora

Figura 36. Nível observado versus predito para o ano completo de 2000

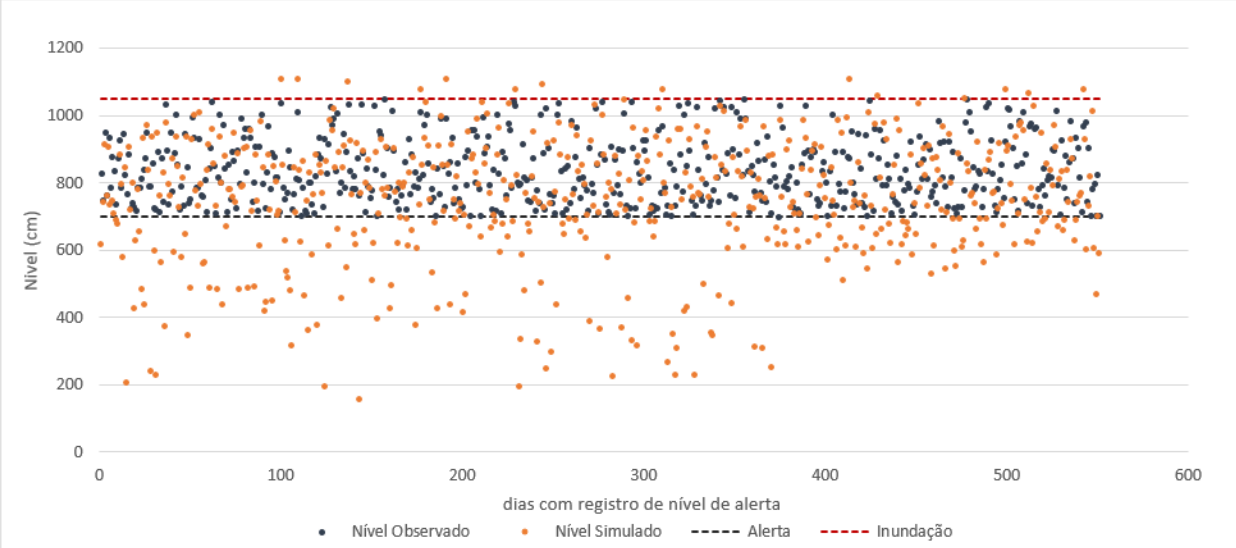


Fonte: Dados da Pesquisa – Elaborado pela autora

O mesmo comportamento é observado quando analisa-se os dias em que foram registrados níveis de inundação, nesse caso apresentando um desempenho ainda pior. Considerando que, do total de dias da série temporal, apenas 551 apresentaram níveis de alerta (o que representa 4,3% do total de dias) e, nos casos de inundação esse valor diminui ainda mais uma vez que apenas em 112 dias da série temporal houve registro de níveis de inundação (representando 0,9%), atribui-se a redução do desempenho do modelo à escassez de amostras

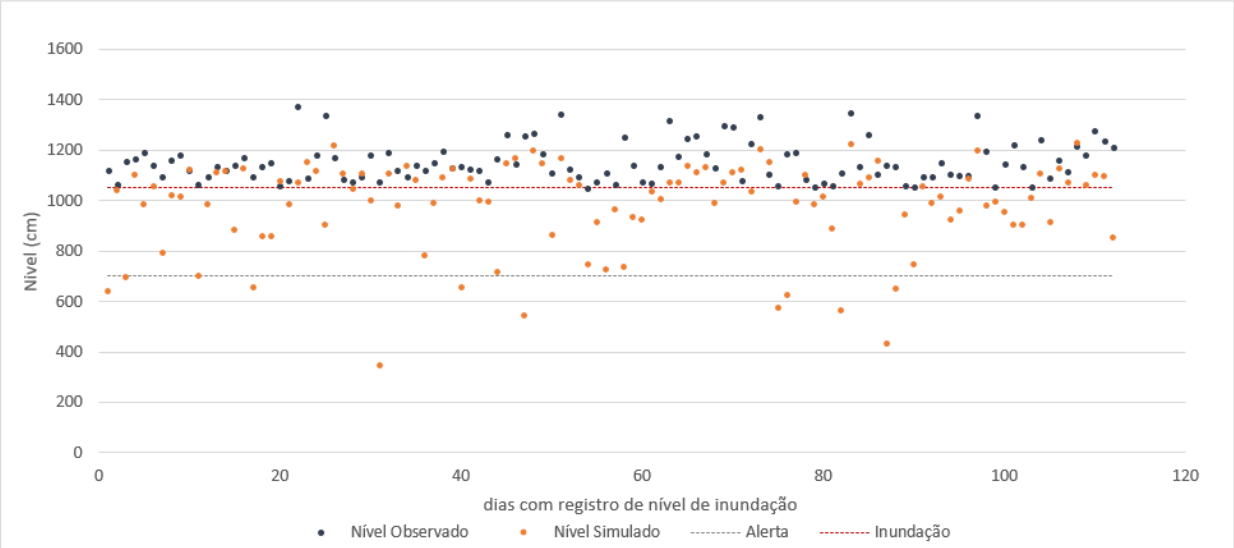
de dias com níveis máximos. Acredita-se que o enriquecimento da série temporal, estendendo-a para um período ainda maior, mas sobretudo que englobe uma quantidade maior de registros de níveis de inundação, repercutirá em uma melhora nos resultados referentes aos níveis de inundação e de alerta preditos. Uma possibilidade é a utilização de dados indiretos, obtidos a partir de sensoriamento remoto, para composição de uma série temporal mais rica em termos de amostras de níveis de alerta e de inundação. As simulações apresentadas na Figura 37, cotas de alerta observadas e respectivas predições, apresentam melhores resultados quando comparado às cotas de inundação preditas (Figura 38), corroborando com a suposição de que maior quantidade de amostras possibilitarão o refinamento do treinamento da rede neural, repercutindo em melhores resultados preditos.

Figura 37. Resultados referentes ao nível de alerta (níveis entre 700 cm até 1.050 cm)



Fonte: Dados da Pesquisa – Elaborado pela autora

Figura 38 Resultados referentes ao nível de inundação (nível acima de 1.050 cm)



Fonte: Dados da Pesquisa – Elaborado pela autora

4.6 Conclusões

Com base no exposto, avalia-se o bom desempenho da rede neural LSTM para a predição de dados de nível, acertando a predição de ocorrência de níveis associados a alerta em 87 % das vezes que isso ocorreu ao longo do período testado. Os resultados obtidos apresentam bom acoplamento da série simulada em relação à observada, obtendo-se melhores resultados, entretanto, para valores não extremos. Desse modo, tendo em vista o objetivo deste estudo em avaliar o modelo para que possa, futuramente ser aplicado à previsão de níveis de inundação para o rio Caí, entende-se a necessidade de serem realizados novos testes utilizando uma série temporal mais extensa espacial e temporalmente, contendo, assim, uma gama maior de níveis de alerta e de inundação. Entretanto, diante da dificuldade em obter mais dados de precipitação e nível a partir de fontes primárias, conforme relatado na etapa de consolidação da base de dados, uma possibilidade que se coloca é a de utilizar dados de precipitação e nível obtidos a partir de medições indiretas, como sensoriamento remoto. Além disso, é possível também consolidar uma série temporal utilizando dados de reanálise, que são aqueles gerados a partir da combinação de modelos de previsão em um sistema de assimilação de dados, técnica utilizada para suprir ausência de dados primários.

Contudo, apesar de ter subestimado a predição de níveis relacionados a eventos mais extremos, de forma geral o modelo apresentou boas medidas de avaliação, sobretudo o

coeficiente de Nash-Sutcliffe que apresentou valores satisfatórios indicando, assim, a adequação do uso da LSTM para prosseguimento de sua aplicação à predição de nível. Destaca-se, sobretudo, a facilidade e agilidade no processo de simulação de níveis quando se utiliza a rede neural em estudo, diferentemente do que ocorre com modelos hidrológicos, os quais necessitam uma série de condições de contorno que representem os fenômenos físicos que levam à inundação. A LSTM apresenta-se de forma mais facilitada na obtenção dos dados de níveis preditos, demonstrando grande potencial a futuros trabalhos na área de previsão de níveis de forma ágil.

REFERÊNCIAS

ABREU, M.C; CECÍLIO, R.A.; PRUSKI, F.F.; SANTOS, G.R.; ALMEIDA, L.T.; ZANETTI, S.S. **Critérios para escolha de distribuições de probabilidade em estudos de eventos extremos de precipitação.** Revista Brasileira de Meteorologia. https://doi.org/10.1590/0102-7786334004. Rio de Janeiro. 2018.

ALBERTON, G.; SEVERO, D.; MELO, M.N.V.; POTELICKI, H.; SARTORI, A. Aplicação de redes neurais artificiais para previsão de enchentes no Rio Itajaí-Açu em Blumenau, SC, Brasil. Revista Ibero Americana de Ciências Ambientais, v.12, n.4, p.686-696, 2021. DOI: http://doi.org/10.6008/CBPC2179-6858.2021.004.0053

BOUIX, Christian Pascal Silva. Modelagem de redes neurais artificiais MLP para previsão de vazões na bacia do rio Miranda afluente do Pantanal. 2024. Tese de Doutorado. Universidade de São Paulo.

BRASIL. Ministério da Integração Nacional. Instrução Normativa nº01, de 24 de agosto de 2012. Estabelece critérios para a decretação de situação de emergência ou estado de calamidade pública. Disponível em https://www.defesacivil.se.gov.br/wp-content/uploads/2020/07/instru%C3%A7%C3%A3o_normativa_n%C2%BA_01_de_24_de_agosto_de_2012-2.pdf. Consulta em 13 de abril de 2024.

BRASIL. Ministério da Integração Nacional. **Manual de planejamento em Defesa Civil. Volume 1.** Brasília: Ministério da Integração Nacional. 1999. Disponível em: https://www.defesacivil.rs.gov.br/upload/arquivos/201511/04145531-11-manual-de-planejamento-em-defesa-civil-volume-1.pdf. Consulta em 13 de abril de 2024

BRASIL. Serviço Geológico do Brasil. Programa de Gestão de Riscos e de Desastres. Projeto de regionalização de vazões nas bacias hidrográficas brasileiras. **Análise de frequência de cotas dos sistemas de alerta: Sistema de Alerta Bacia do Rio Caí.** 2022

BRASIL. Ministério da Integração e Desenvolvimento Regional. Secretaria Nacional de Defesa Civil. **Atlas Digital de Desastres Naturais no Brasil.** 2023. Disponível em http://atlasdigital.mdr.gov.br/. Consulta em: 14 de janeiro de 2024.

BRASIL. Ministério da Integração e Desenvolvimento Regional. Secretaria Nacional de Defesa Civil. **Sistema Integrado de Informações sobre Desastres (S2iD).** Disponível em https://s2id.mi.gov.br/. Consulta em: 14 de janeiro de 2024.

BROWNLEE, J. **Predict the future with MLPs, CNNs and LSTMs in python.** Machine Learning Mastery. Deep Learning fortimeseries. 2020. Disponível em:

https://books.google.com.br/books?id=o5qnDwAAQBAJ&pg=PR2&hl=pt-BR&source=gbs_selected_pages&cad=1#v=onepage&q&f=false. Consulta em: 02 de abril de 2024.

BRUCE, P.; BRUCE, A. Estatística prática para cientistas de dados: 50 conceitos essenciais. Traduzido por Luciana Ferraz. Rio de Janeiro. Alta Book. 2019

BRUNER, M., SLATER, L., TALLAKSEN, L.M. e CLARK, M. Challenges in modeling and predicting floods and droughts: A review**. Wires Wiley Interdisciplinary Reviews.** Volume 8, Issue 3. 2021. Disponível em: https://wires.onlinelibrary.wiley.com/doi/10.1002/wat2.1520 Consulta em: 13 de abril de 2024.

COLLISCHONN, W; DORNELLES, F. Hidrologia para engenharia e ciências ambientais. Porto Alegre. Associação Brasileira de Recursos Hídricos – ABRH. 2013

CRUZ, M.F.M; RODRIGUES, L.D.; VERSIANI, B.R. Previsão de vazões com metodologia DPFT e com redes neurais artificiais. Revista Brasileira de Recursos Hídricos. Volume 15. N.1. 2010.

FAGUNDES, M. **Previsão hidrológica como ferramenta para auxiliar no critério de fechamento da trilha do Rio do Boi (SC).** 2021. 148f Dissertação (Mestrado em Recursos Hídricos e Saneamento Ambiental) – Instituto de Pesquisas Hidráulicas, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2021

FARIAS, A.A; SOARES, J.F; CÉSAR, C.C. Introdução à estatística. 2ª edição. Rio de Janeiro. LTC. 1998

GAMA, H.A.G.; PEDROLLO, O.C. Previsão de níveis na bacia do Rio Mundaú a partir de redes neurais artificiais. XIV Simpósio de Recursos Hídricos do Nordeste. Associação Brasileira de Recursos Hídricos. 2018.

GOVINDARAJU, R.S. **Artificial Neural Networks in Hydrology. I: Preliminary Concepts.** Journal of Hydrologic Engineering. Volume 5. Issue 2. April 2000

HUANG, L.G.; ZHANG, N.; CROSBIE, R.S.; YE, L.; LIU, J.; GUO, ZHAOXIA.; MENG, Q.; FU, G.; BRYAN, B. A top-down deep learning model for predicting spatiotemporal dynamics of groundwater recharge. Environmental Modellung and Software. Elsevier. www.elsevier.com/locate/envsoft. 2023.

KIM, C.; KIM, C.-S. Comparsion of the performance of hydrologic model and a deep learning technique for rainfall-runoff analysis. **Tropical Cyclone and Review – Elsevier**, n.10 p.2015-2022. 2021. Doi: https://doi.org/10.1016/j.tcrr.2021.12.001

LIANG, C.; LI, H.; LEI, M.; DU, Q. Dongting lake water level forecast and its relationship with the three gorges dam based on a long short-term memory network. Water 2018, 10(10), 1389; https://doi.org/10.3390/w10101389

LUFI, SURYANINGTYAS; RISPININGTATI, ERY SUHARTANTO. **Hydrological Analysis of TRMM (Tropical Rainfall Measuring Mission) Data n Lesti Sub Watershed**. Civil Environmental Science Journal. Vol. III, N°.01, pp.018-030. 2020

MEDEIROS, Kevin Martins et al. Modelos de previsão de vazão afluente da UHE-Tucuruí: uma abordagem com redes neurais LSTM e CNN. 2023.

MIGLIATO, A.L.T. **Detecção de outliers em dados não vistos de séries temporais por meio de erros de predição com SARIMA e redes neurais recorrentes LSTM e GRU.** 2021. 107 f. Dissertação (Mestrado Profissional em Matemática, Estatística e Computação Aplicada à Indústria) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2021.

MONTE, B. E. O. **CRBi: Um índice de risco de inundações desenvolvido para municípios brasileiros.** 2022. 253 f. Tese (Doutorado em Recursos Hídricos e Saneamento Ambiental) – Instituto de Pesquisas Hidráulicas, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2022.

MONTE, B.E.O. MICHEL, G.P. GOLDENFUM, J.A. Reflexões sobre alguns conceitos relacionados a desastres naturais no brasil e no mundo. I ENCONTRO NACIONAL DE DESASTRES, 2018, Porto Alegre. Disponível em: https://lume.ufrgs.br/handle/10183/185945 Acesso em: 29 de março de 2024

NAGHETTINI, M; PINTO, E.J.A. Hidrologia estatística. Belo Horizonte. Serviço Geológico do Brasil - CPRM. 2013

NIELSEN, A. **Análise Prática de Séries Temporais. Predição com estatística e aprendizdo de máquina.** Rio de Janeiro. Alta Books. 2021

PAIVA, R.; COLLISCHONN, W.; MIRANDA, I.P.; DORNELLES, F.; GOLDENFUM, J.; FAN, F.; RUHOFF, A.; FAGUNDES, H. **Critérios hidrológicos para adaptação à mudança climática: Chuvas e cheias extremas na Região Sul do Brasil.** Nota técnica. Universidade Federal do Rio Grande do Sul. Instituto de Pesquisas Hidráulicas. Porto Alegre. 27 de maio de 2024.

PHEBO, L. **Usando Inteligência Artificial para alertar e prever inundações no Brasil.** Disponível em: https://blog.google/intl/pt-br/novidades/iniciativas/usando-inteligencia-artificial-para-alertar-e-prever-inundacoes-no-brasil/#:~:text=J%C3%A1%20presente%20em%20outros%20pa%C3%ADses,de%20inunda%C3%A7%C3%B5es%20em%20regi%C3%B5es%20ribeirinhas. Acesso em: 19 de dezembro de 2023.

RIO GRANDE DO SUL. Secretaria de Planejamento, Governança e Gestão. **Desastres naturais no Rio Grande do Sul: estudo sobre as ocorrências no período 2003-2021.** Departamento de Planejamento Governamental, 2022.

ROCHA, M.H.P.; MINE, M.R.M. e KAVISKY, E. Verificação do potencial das redes neurais artificiais em descrever o processo chuva-vazão mensal com cenários de modelos climáticos

regionais. XXI Simpósio Brasileiro de Recursos Hídricos, Associação Brasileira de Recursos Hídricos. 2015.

ROHN, M.C.; MINE, M.R.M. **Uma aplicação das redes neurais artificiais à previsão de chuvas de curtíssimo prazo.** XV Simpósio Brasileiro de Recursos Hídricos. Associação Brasileira de Recursos Hídricos. 2003

SAMBATI, S. MARTINS, R. G.; VILELA, R. B.; COTACALLAPA, M.; PESSOA, A. S. A.; DIAS, J.; BRESSIANI, D.; FERNANDES, G. **Previsão de riscos de alagamentos e inundações com uso de inteligência artificial.** Revista de Informática Aplicada, Volume 15, Número 1. 2019

SCHEN, C. A transdisciplinar review of deep learning research and its relevance for water resources scientists. **Water Resources Research**, 54, 8558–8593. 2018. https://doi.org/10.1029/2018WR022643

SCHMIDT, V.; LUCCIONI, A.; TENG, M.; ZHANG, T.; REYNAUD, A. S. R.; COSNE, G.; JURAVER, A.; VARDANYAN, V.; HERNANDEZ-GARCIA, A.; BENGIO, Y. **ClimateGAN: Raising Climate Change Awareness by Generating Images of Floods.** 2021. Disponível em: https://arxiv.org/pdf/2110.02871.pdf Consulta em 20 de janeiro de 2024

SPERB, R. M. et al. Prevent: Protótipo de um sistema de previsão de enchentes baseado em redes neurais. Simpósio Brasileiro de Recursos Hídricos, v. 13, 1999.

TOMINAGA, L.K.; SANTORO, J; AMARAL, R. **Desastres naturais: Conhecer para prevenir.** 3ª edição. São Paulo: Instituto Geológico, 2015.

TSCHIEDEL, A.F. **Abordagem de grande escala para simulações de cheias geradas por rompimentos de barragens de armazenamento de água.** 2022. 255 f. Tese (Doutorado em Recursos Hídricos e Saneamento Ambiental) – Instituto de Pesquisas Hidráulicas, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2022.

WORLD RISK REPORT. Bündnis Entwicklung Hilft / IFHV (2023): WeltRisikoBericht 2023. Berlin: Bündnis Entwicklung Hilft. 2023. Disponível em: https://weltrisikobericht.de/wp-content/uploads/2023/10/WRR_2023_english_online161023.pdf Acesso em 20 de março de 2024