Carcará: Sistema para Integração de Dados Meteorológicos e Hidrológicos no Semiárido Brasileiro

Mikael M F Sales[1], Jose Wellington Franco da Silva[2], Bruno Riccelli dos Santos Silva[3], Amanda Drielly Pires Venceslau[4]

[1]Universidade Federal do Ceará (UFC) – Campus Crateús

mikaelmfsales@gmail.com, wellington@crateus.ufc.br,
bruno@lesc.ufc.br, amanda.pires@ufc.br

Abstract. *The Brazilian Semiarid region has dispersed and poorly standardized meteorological and hydrological data, hindering integration and analysis. This work proposes Carcará, an automatic architecture for integrating INMET, ANA, and ERA5-Land data, based on an event-driven ETL pipeline, Data Lake, Data Warehouse, asynchronous messaging, and workflow orchestration. Validation was conducted in a laboratory environment over seven days of continuous operation, producing a dataset with 17,472 hourly records, 192 timestamps, 44 municipalities, 8 Brazilian states, and 91 reservoirs. The results validate the end-to-end operation and indicate the architecture's feasibility for continuous integration and hydrometeorological analyses in the Semiarid region.*

Resumo. *O Semiárido brasileiro apresenta dados meteorológicos e hidrológicos dispersos e pouco padronizados, dificultando sua integração e análise. Este trabalho propõe o Carcará, uma arquitetura de integração de dados do INMET, ANA e ERA5-Land, baseada em pipeline ETL orientado a eventos, Data Lake, Data Warehouse, mensageria assíncrona e orquestração de workflows. A validação ocorreu em ambiente laboratorial durante sete dias de operação contínua, gerando um dataset com 17.472 registros horários, 192 timestamps, 44 municípios, 8 UFs e 91 reservatórios. Os resultados validam a operação ponta a ponta e indicam a viabilidade da arquitetura para integração contínua e análises hidrometeorológicas no Semiárido.*

1. Introdução

As tecnologias digitais têm ampliado o uso de soluções orientadas por dados na agricultura, setor de elevada relevância socioeconômica e estratégica para a produção sustentável [Troisi et al. 2023]. No Brasil, o agronegócio representou 23,2% do PIB em 2024 [CEPEA 2025b], empregou 28,2 milhões de pessoas [CEPEA 2025a] e respondeu por 48,9% das exportações nacionais[1], enquanto a agricultura familiar é responsável por cerca de 70% dos alimentos consumidos internamente[2]. Nesse contexto, o Semiárido Brasileiro destaca-se pela elevada variabilidade climática, irregularidade das chuvas e risco recorrente de escassez hídrica, embora concentre cerca de 50% dos estabelecimentos de agricultura familiar do país e desempenhe papel relevante na segurança alimentar nacional [Fragoso et al. 2020]. A região abrange os estados do Nordeste e parte do norte

[1] www.gov.br/agricultura/pt-br [2] www.camara.leg.br

de Minas Gerais, correspondendo a aproximadamente 12% do território nacional e a uma população de cerca de 28 milhões de habitantes.

Diante desse cenário, abordagens orientadas por dados tornam-se estratégicas para a agricultura [Silva et al. 2025]. O uso de *Big Data* aprimora modelos de previsão, otimiza processos produtivos e aumenta a eficiência no uso de recursos naturais [Chen and Lv 2025]. A integração de dados climáticos, imagens de satélite e modelos de aprendizado profundo tem superado abordagens convencionais na predição de produtividade, oferecendo suporte mais preciso à tomada de decisão agrícola [Mena et al. 2024].

Em regiões vulneráveis, como o Semiárido brasileiro, a variabilidade climática intensifica incertezas produtivas e limita decisões baseadas apenas em práticas empíricas [Marengo et al. 2017]. Assim, a estruturação de fluxos automatizados de aquisição, integração e análise de dados constitui uma alternativa promissora para fortalecer a segurança hídrica e a resiliência agrícola regional.

Nesse contexto, este artigo apresenta o Carcará, uma arquitetura para integração contínua de dados hidrometeorológicos no Semiárido brasileiro, automatizando a aquisição, o pré-processamento, a padronização, a persistência e a disponibilização analítica de dados do INMET[3], ANA[4] e ERA5-Land[5]. A solução é estruturada como um *pipeline ETL* assíncrono e orientado a eventos, no qual a extração coleta os dados nas fontes, a transformação realiza limpeza, harmonização, normalização e integração espaço-temporal, e a carga materializa os dados em uma infraestrutura composta por *Data Lake* e *Data Warehouse*. Além da arquitetura, este trabalho apresenta sua implementação e validação operacional, demonstrando a capacidade de consolidar fontes heterogêneas em uma base integrada para análises climáticas, hidrológicas e agrícolas no Semiárido.

2. Trabalhos Relacionados

Os trabalhos relacionados abordam aspectos específicos da coleta, integração e processamento de dados ambientais e agrícolas. [Stephen et al. 2021] propõem uma arquitetura em nuvem para dados *Landsat*, *MODIS* e *Sentinel*, com padronização espacial e fusão de séries temporais, mas sem implementação prática ou integração com dados *in situ* e hidrológicos. [Garg and Alam 2023] desenvolvem um sistema agrícola baseado em *IoT*, com sensoriamento, *gateway*, nuvem e aplicação, utilizando *ThingSpeak*[6] para armazenar e analisar variáveis ambientais e biológicas; contudo, a solução restringe-se ao monitoramento local e depende de conectividade contínua. [Freitas et al. 2023] estruturam um *data lake* para consolidar observações meteorológicas governamentais no Rio de Janeiro, apoiando modelos preditivos de curto prazo, mas sem contemplar simultaneamente dados agrícolas, hidrológicos, sensoriamento remoto e medições *in situ*.

Esses estudos avançam em dimensões isoladas, como sensoriamento remoto, monitoramento por sensores e integração meteorológica urbana. Contudo, não apresentam uma arquitetura unificada que combine dados climáticos, hidrológicos e agrícolas, automação contínua, pré-processamento espaço-temporal, governança analítica e disponibilização pública no contexto do Semiárido brasileiro. Diferentemente deles, este trabalho propõe uma arquitetura baseada em *Data Lake* e *Data Warehouse*, integrando coleta *in situ*, sensoriamento remoto e dados hidrológicos, com limpeza,

[3] portal.inmet.gov.br [4] www.ana.gov.br/sar [5] cds.climate.copernicus.eu/datasets
[6] thingspeak.com

padronização, enriquecimento e integração espaço-temporal. Assim, busca-se preencher a lacuna referente à construção de uma infraestrutura analítica unificada, atualizável e orientada às demandas climáticas e agrícolas do Semiárido.

3. Metodologia

O Carcará foi implementado como um *pipeline* ETL em três camadas, conforme a Figura 1: *Extract*, *Transform* e *Load*. Na etapa *Extract*, os dados são coletados nas fontes e armazenados como artefatos brutos no *MinIO*. Eventos de mensageria sinalizam sua disponibilidade e acionam a etapa *Transform*, responsável pela limpeza, padronização, conversão e geração de dados processados. Por fim, a etapa *Load* carrega os dados no *PostgreSQL*, compondo um *data warehouse* em esquema estrela, com dimensões compartilhadas e tabelas fato por fonte. O *Apache Airflow* agenda e monitora a execução recorrente do fluxo.

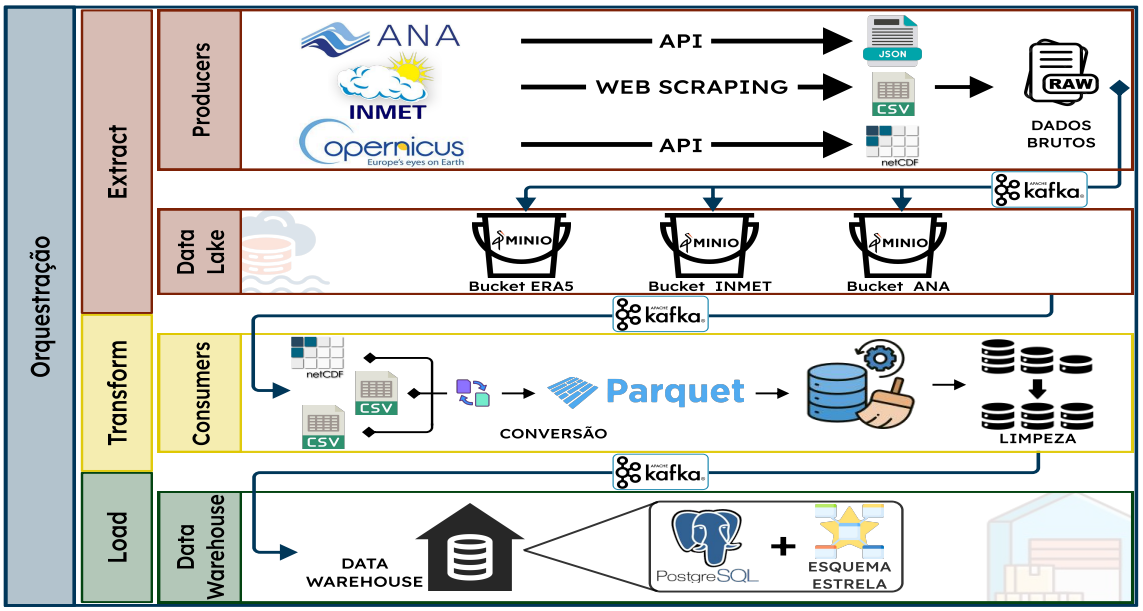

Figure 1. Fluxo metodológico de integração de dados do Carcará.

A coleta foi restrita ao semiárido, delimitado aproximadamente pelas coordenadas Norte $-2°$, Oeste $-46{,}5°$, Sul $-10°$ e Leste $-34{,}5°$. Foram considerados apenas municípios com cobertura simultânea das três fontes: estações ativas do INMET, sensores *in situ* da ANA e cobertura espacial do ERA. As séries do INMET e do ERA foram tratadas em resolução horária, enquanto os dados da ANA foram usados em resolução diária, sua máxima granularidade disponível. Devido à defasagem operacional do ERA, a coleta foi parametrizada para a data $D - 6$, em função do atraso na disponibilização dos dados pela API do *Climate Data Store* (CDS). Com isso, as bases do INMET e da ANA foram alinhadas à mesma referência temporal.

3.1. Extract

A etapa *Extract* realiza a aquisição inicial dos dados e sua persistência na camada bruta do *data lake*, preservando estrutura, granularidade e formato originais. Após o armazenamento, um evento é emitido para acionar o processamento assíncrono. Os dados da ANA

são coletados via API, por período e reservatórios, retornando medições diárias como capacidade, cota e volume. O ERA é obtido pela API do *Climate Data Store*, parametrizada por período, área e variáveis. O INMET é extraído por *Web Scraping*, a partir de arquivos compactados com séries meteorológicas das estações selecionadas.

3.2. Transform

A etapa *Transform* realiza o pré-processamento, a qualificação e a harmonização dos dados, conforme a Tabela 1. Após receber o evento de um artefato bruto, o sistema recupera o arquivo no *data lake*, aplica rotinas específicas por fonte, executa limpeza, padronização e organização estrutural, e materializa o resultado em `Parquet`. Essa etapa produz dados consistentes e adequados à integração espaço-temporal e à posterior carga.

Table 1. Resumo das etapas de transformação aplicadas nas fontes de dados, formatos de entrada e estruturas de saída geradas.

Fonte	Entrada	Operação	Saída
ANA	CSV	Normaliza município/UF e valida a estação; converte a data; aplica *pivot* (long→wide) e expande cada dia em 24 horas; quantiza.	Parquet com **Data**, **Hora**, **Município**, **UF**, **Reservatório**, **Capacidade**, **Cota**, **Volume**, **VolumeÚtil**.
ERA5	NetCDF (`.nc`)	Converte NetCDF para DataFrame e remove colunas de grade; agrega por `valid_time` (média); concatena múltiplos arquivos e consolida novamente; formata Data/Hora; quantiza.	Parquet com **Data**, **Hora** + variáveis climáticas (ex.: temperatura, pressão, etc.).
INMET	ZIP+CSV	Descompacta ZIP; detecta encoding/delimitador; remove linhas de metadado e valida estação; renomeia colunas; trata datas/horas e números (vírgula→ponto); adiciona município/UF; concatena verticalmente e limpa intermediários; quantiza.	Parquet unificado com **Data**, **Hora**, **Município**, **UF** + variáveis meteorológicas INMET.

Transform - ANA: Os dados da ANA foram convertidos para uma estrutura tabular padronizada, com validação espacial e normalização de município e unidade federativa. Os campos temporais foram uniformizados e os registros, originalmente em formato *long*, foram transformados para *wide* por meio de *pivot*, alocando capacidade, cota, volume e volume útil em colunas próprias. Como a fonte possui resolução diária, cada registro foi expandido para 24 observações horárias, replicando os valores ao longo do dia. Por fim, os dados foram quantizados e materializados em `Parquet`, contendo data, hora, município, UF, reservatório e métricas hidrológicas.

Transform - ERA5: Os dados do ERA, obtidos em `NetCDF`, foram convertidos para uma estrutura tabular compatível com as demais fontes, preservando o tempo como eixo principal de integração. As variáveis de interesse foram selecionadas, as coordenadas espaciais reorganizadas em colunas e atributos auxiliares de grade foram removidos. Variáveis multicamadas, como temperatura e umidade do solo em diferentes profundidades, foram desagregadas em colunas específicas. Em seguida, os dados foram agregados por instante temporal, quando necessário, concatenados e consolidados para evitar fragmentação entre recortes temporais. Os campos de data e hora foram padronizados, e os valores foram quantizados para reduzir custo computacional, uso de memória e armazenamento. O resultado final foi exportado em `Parquet`, contendo data, hora e variáveis climáticas de interesse.

Transform - INMET: Os dados do INMET foram recebidos em arquivos compactados contendo `CSV`s com medições horárias. A transformação iniciou-se pela descompactação, leitura dos arquivos, identificação de *encoding* e delimitador, remoção de metadados e validação das estações conforme o recorte espacial adotado. Em seguida, as colunas

foram padronizadas, com renomeação de cabeçalhos, tratamento de data e hora, conversão de valores com vírgula decimal e inclusão de município e UF associados a cada estação. Os arquivos de diferentes estações e períodos foram concatenados, removendo registros intermediários, redundantes ou inconsistentes. Por fim, os dados foram quantizados e materializados em `Parquet`, contendo data, hora, identificação espacial e variáveis meteorológicas observadas.

3.3. Load

A etapa *Load* incorpora os dados transformados ao *data warehouse*. Após receber o evento de disponibilidade, o sistema recupera o artefato processado, conecta-se ao *PostgreSQL* e executa inserções e atualizações. As dimensões são mantidas por operações de *upsert*, preservando identificadores e atributos descritivos, enquanto os registros são carregados nas tabelas fato específicas de cada fonte.

3.4. Orquestração e infraestrutura de execução

A automação e o monitoramento do *pipeline* foram realizados pelo *Apache Airflow* em ambiente conteinerizado. A ingestão foi programada para execução diária às 23h, enquanto transformação e carga operaram continuamente a partir dos eventos gerados no fluxo. A infraestrutura foi organizada com *Docker Compose*, garantindo portabilidade, empacotamento e integração entre serviços. A observabilidade foi mantida pelos estados de execução do *Airflow* e pelos logs dos contêineres.

4. Resultados Experimentais

O sistema foi validado em laboratório. Durante os testes, foram realizados três ajustes principais: verificação prévia de requisições pendentes no INMET, devido à limitação de uma requisição ativa por vez no BDMEP; verificações periódicas para lidar com respostas superiores a 24 horas; e submissão das requisições na última hora do dia de referência, reduzindo desalinhamentos entre ANA, ERA e INMET. Após os ajustes, o sistema operou continuamente de 27/12/2025 a 03/01/2026, executando com todas as funcionalidades.

A partir do *dataset* construído, as Figuras 2 e 3 exemplificam análises integradas entre variáveis do INMET, ANA e ERA5. A Figura 2 relaciona a umidade relativa média diária (UR) ao déficit de pressão de vapor (VPD), dado por $\mathrm{VPD} = e_s(T) - e_a(T_d)$, evidenciando relação inversa entre atmosfera mais seca e menor umidade: entre 27/12 e 30/12, o VPD aumenta enquanto a UR diminui; após esse período, o VPD reduz ou estabiliza e a UR aumenta. A Figura 3 relaciona a evaporação potencial diária (PEV) à temperatura média do ar a 2 m (*t2m*) e ao VPD, indicando que dias mais quentes e secos apresentam maior demanda evaporativa. Esses resultados devem ser interpretados como tendências espaço-temporais e condições potenciais de evaporação, não como causalidade direta ou evaporação real observada.

5. Conclusão e Trabalhos Futuros

No período experimental, o Carcará demonstrou a viabilidade de uma arquitetura assíncrona e orientada a eventos para integração automatizada de dados hidrometeorológicos no Semiárido brasileiro. A solução operou de ponta a ponta, da extração à materialização no *Data Warehouse*, gerando um *dataset* com 17.472 registros horários,

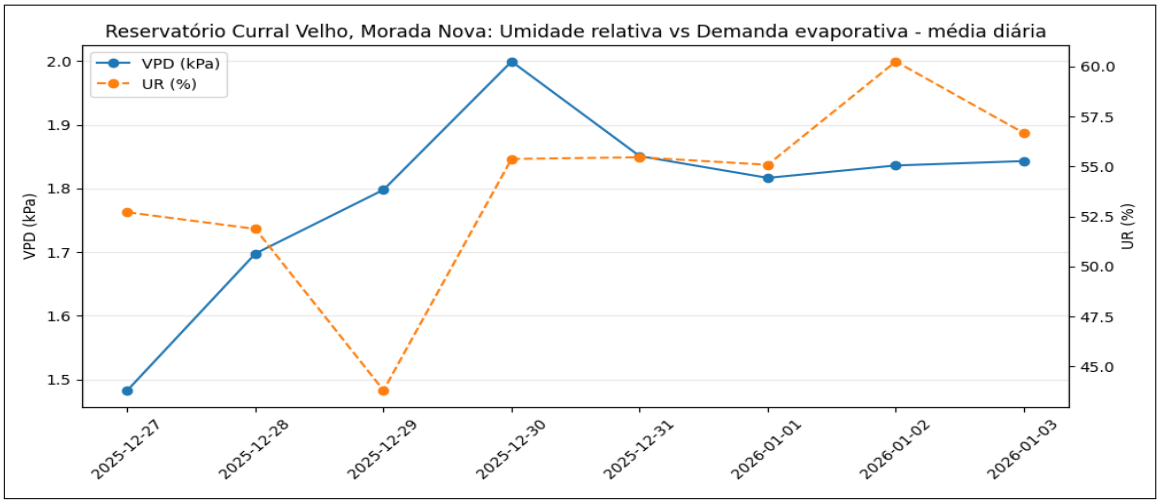

Figure 2. Relação entre umidade relativa e demanda evaporativa.

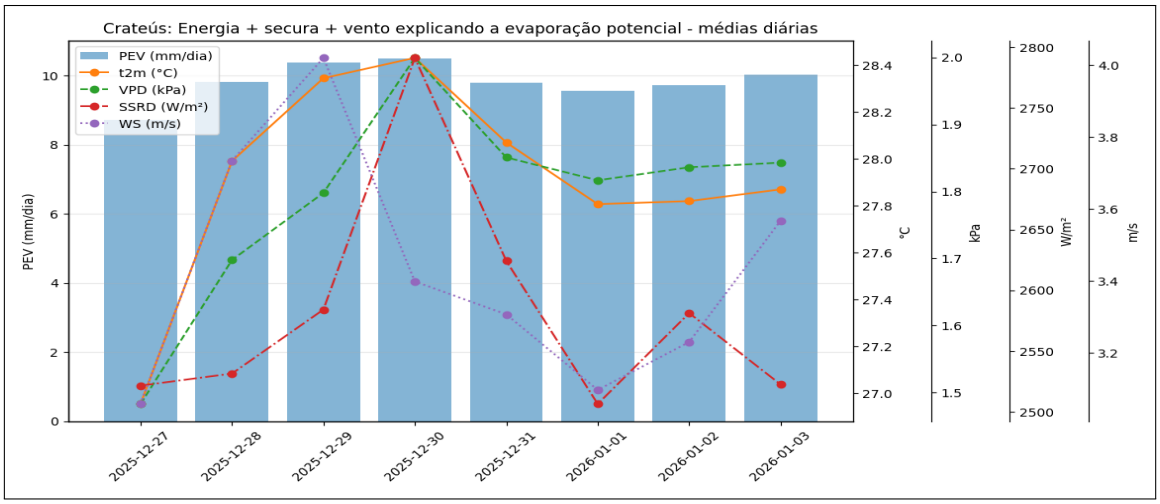

Figure 3. Influência da energia, VPD e vento na evaporação potencial.

192 *timestamps*, 44 municípios, 8 UFs e 91 reservatórios. A janela temporal D−6 garantiu o alinhamento entre as fontes, considerando a defasagem operacional do ERA5. Quanto à cobertura, o ERA5 apresentou completude integral, a ANA forneceu medições para 58 dos 91 reservatórios (63,7%) e o INMET foi a principal limitação, com dados em 20 dos 44 municípios e ao menos um campo meteorológico preenchido em 39,6% dos registros. Assim, os resultados indicam a viabilidade técnica da proposta, embora a cobertura final dependa da disponibilidade das bases externas. A arquitetura proposta e o *dataset* gerado nos experimentos estão disponíveis publicamente no GitHub[7].

Como trabalhos futuros, pretende-se ampliar a cobertura espacial e temporal, incorporar fontes com atualização próxima ao tempo real, fortalecer verificações de qualidade dos dados e expandir a observabilidade operacional, incluindo latência, falhas e *retries* por fonte. Também se prevê o desenvolvimento de uma camada de acesso para consulta, filtragem e visualização dos dados, além da generalização da arquitetura para outros recortes territoriais.

[7] https://github.com/mikaelmota13/carcara.git

6. Declaração de Responsabilidade e Uso de IA Generativa

Este trabalho é produto de um Trabalho de Conclusão de Curso (TCC) desenvolvido no curso de Sistemas de Informação. Não identificamos preocupações sociais ou éticas imediatas decorrentes de sua realização. Registramos, ainda, o uso do modelo de linguagem ChatGPT exclusivamente como apoio à revisão gramatical. Declaramos assumir responsabilidade integral por todo o conteúdo apresentado no manuscrito.

References

CEPEA (2025a). MERCADO DE TRABALHO/CEPEA: Agronegócio emprega 28,2 milhões de pessoas em 2024 e representa 26% das ocupações do País. Acesso em: 20 nov. 2025.

CEPEA (2025b). PIB-Agro/CEPEA: Desempenho do 4º trimestre reverte tendência de queda anual, e PIB do agronegócio avança 1,81% em 2024. Acesso em: 20 nov. 2025.

Chen, J. and Lv, N. (2025). Using big data and machine learning to optimize agricultural resource allocation and ecological and economic benefit evaluation. *International Journal of Agricultural and Environmental Information Systems*.

Fragoso, E. J. N., Coelho, P., Souza, P., Neto, A. F., dos Santos Santiago, A. M., Pacheco, C. S. G. R., and Melo, R. (2020). Agroecology and family farming: A perspective of sustainability in the brazilian semiarid region. *International Journal of Advanced Engineering Research and Science*.

Freitas, B. A., Fonseca, A., Bezerra, E., Bernardini, F., and Ferro, M. (2023). Diversidade de dados meteorológicos da cidade do rio de janeiro: Uma proposta de arquitetura de armazenamento e fluxo de dados para modelos de previsão. In *Anais Estendidos do XXXVIII Simpósio Brasileiro de Bancos de Dados*, pages 306–311, Porto Alegre, RS, Brasil. Sociedade Brasileira de Computação.

Garg, D. and Alam, M. (2023). A sensor data acquisition system for smart agriculture. *SN Computer Science*, 4(5):667.

Marengo, J. A., Torres, R. R., and Alves, L. M. (2017). Drought in northeast brazil—past, present, and future. *Theoretical and Applied Climatology*, 129(3):1189–1200.

Mena, F., Pathak, D., Najjar, H., Sanchez, C., Helber, P., Bischke, B., Habelitz, P., Miranda, M., Siddamsetty, J., Nuske, M., et al. (2024). Adaptive fusion of multi-view remote sensing data for optimal sub-field crop yield prediction. *arXiv preprint arXiv:2401.11844*.

Silva, J. L. P. d., Júnior, G. d. N. A., Silva Junior, F. B. d., Silva, T. G. F. d., Silva, J. B. A. d., Scheibel, C. H., Silva, M. V. d., Mingoti, R., Giongo, P. R., and Almeida, A. C. d. S. (2025). Performance of land use and land cover classification models in assessing agricultural behavior in the alagoas semi-arid region. *AgriEngineering*, 7(5):134.

Stephen, A., Punitha, A., and Chandrasekar, A. (2021). Using open remote sensing data to build an agriculture big data system. *Turkish Journal of Computer and Mathematics Education*, 12(2):429–436.

Troisi, O., Visvizi, A., and Grimaldi, M. (2023). Rethinking innovation through industry and society 5.0 paradigms: a multileveled approach for management and policy-making. *European Journal of Innovation Management*, 27(9):22–51.