Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias Físico Matemáticas

Estudio del espectro de energía de rayos cósmicos mediante técnicas de Machine Learning con datos del Observatorio Pierre Auger

T E S I S

como requisito para la obtención del grado de

Licenciado en Física

presenta

Victor Aldair Garmendia Fuentes

Asesores:
Dr. Enrique Varela Carlos
Dr. Humberto A. Salazar Ibargüen

Puebla, Puebla, México
Mayo, 2025

Estudio del espectro de energía de rayos cósmicos mediante técnicas de Machine Learning con datos del Observatorio Pierre Auger

Victor Aldair Garmendia Fuentes

Comité

Dr. Ivan Fuentecilla Carcamo
Presidente

Dr. José Juan Castro Alva
Secretario

Dr. Luis M. Villaseñor Cendejas
Vocal

Dr. Sebastián Rosado Navarro
Suplente

Dr. Enrique Varela Carlos
Asesor

Dr. Humberto A. Salazar Ibargüen
Asesor

A mis padres

Agradecimientos

Hoy 5 de mayo, mientras avanzo en la lectura final de esta tesis, consciente del esfuerzo, las horas dedicadas y los aprendizajes que este proceso me ha dejado, siento la necesidad de detenerme y agradecer.

En primer lugar, a mis padres *doña Rosario* y *don Victor*, cuyo apoyo incondicional y ejemplo de vida han sido fundamentales en esta etapa. A mis hermanos, Kari y David, gracias por estar siempre presentes. Sin ustedes, no sería la persona que soy hoy. Los amo.

A mis amigos, por su apoyo constante y por todos los momentos compartidos. Ustedes son prueba de que las etapas menos favorables pueden volverse memorables cuando se pasan en buena compañía. No importa el rumbo que tome la vida, llevaré siempre su amistad con gratitud y orgullo.

Por supuesto, al Dr. Enrique y al Dr. Humberto, mis asesores. Por su valiosa orientación y por haberme brindado las herramientas necesarias para desarrollar este trabajo.

A todos ustedes, mi más sincero y profundo agradecimiento.

Índice general

Resumen

En este trabajo se implementa un enfoque basado en modelos de series de tiempo utilizando técnicas de aprendizaje automático (Machine Learning, ML) para la predicción de la exposición acumulada en años futuros. A partir de estas proyecciones, se propone la estimación extendida del espectro de energía de los rayos cósmicos. Este estudio se centra en datos recopilados por el detector de superficie del Observatorio Pierre Auger durante el período 2004-2023.

Se espera que esta metodología proporcione un enfoque más sencillo en la resolución del espectro, así como una capacidad mejorada para la extrapolación. Además, este trabajo busca sentar una base sólida para la integración de modelos predictivos basados en ML dentro del análisis científico de rayos cósmicos, como complemento a los enfoques tradicionales sustentados en simulaciones.

Introducción

El estudio de los rayos cósmicos de alta energía constituye uno de los desafíos más relevantes en la astrofísica moderna, debido a las incertidumbres sobre su origen, mecanismos de aceleración y propagación a través del medio interestelar. Aunque los avances experimentales a través de enormes observatorios han permitido extender las mediciones de los rayos cósmicos a rangos sin precedentes, la reconstrucción del espectro de energía continúa limitada por factores sistemáticos; asociados a las incertidumbres estadísticas y sistemáticas en la reconstrucción de eventos, así como a la eficiencia y cobertura del detector.

Con el objetivo de aminorar estas limitaciones, las técnicas de aprendizaje automático ofrecen una vía prometedora para mejorar la estimación de variables clave, como la exposición acumulada, y facilitar la extrapolación del espectro energético en dominios donde los datos son escasos. Este trabajo propone el uso de modelos de series de tiempo, en particular LSTM y Prophet, para predecir el crecimiento de la exposición en años futuros, utilizando datos recopilados por el detector de superficie del Observatorio Pierre Auger.

A partir de estas predicciones, se busca estimar el comportamiento futuro del espectro de energía, evaluando la viabilidad de los modelos propuestos frente a métodos tradicionales basados en simulaciones. Este análisis tiene como propósito estudiar el espectro de energía a partir de los datos recolectados y compararlo con el comportamiento del espectro futuro a partir de la exposición predicha.

Esta tesis se estructura de la siguiente manera: el capítulo 1 presenta una revisión general sobre los rayos cósmicos y su espectro de energía; el capítulo 2 describe la instrumentación y métodos de detección del Observatorio Pierre Auger; el capítulo 3 introduce los fundamentos del aprendizaje automático y los modelos utilizados; el capítulo 4 expone el análisis del conjunto de datos y el cálculo de la exposición; el capítulo 5 aborda la reconstrucción y predicción del espectro energético; finalmente, el capítulo 6 presenta las conclusiones y oportunidades de mejora identificadas.

Capítulo 1

Rayos cósmicos

Al observar el cielo, difícilmente imaginamos la constante lluvia de partículas que atraviesan nuestro planeta de manera invisible para nuestros ojos. Estas partículas, conocidas como rayos cósmicos, llegan desde todas las direcciones y tienen diferentes orígenes. Se trata de núcleos atómicos y partículas subatómicas que viajan por el espacio a velocidades cercanas a la luz. Desde su descubrimiento, han desafiado nuestro entendimiento del universo y han dado lugar a un campo completamente nuevo de la física moderna.

1.1 Una breve historia

A inicios del siglo XX, los científicos notaron una radiación ionizante[1] inexplicable, presente en la atmósfera terrestre. Los primeros experimentos encaminados a entender esta radiación fueron realizados por Theodor Wulf y Domenico Pacini. En 1909, Wulf llevó a cabo mediciones utilizando un electroscopio[2] portátil colocado en la cima de la Torre Eiffel, intentando comprobar si la radiación disminuía con la altitud conforme se alejaba del suelo, que era considerado fuente de radiación terrestre. Sorprendentemente, observó que la radiación en la cima a más de 300 metros no era ni la mitad de su valor en el suelo, indicando que posiblemente existía otra fuente distinta a la terrestre, aunque esto no se consideró concluyente en su momento.

Posteriormente, Pacini en 1911 realizó experimentos bajo el agua, utilizando recipientes sumergidos en el Mar Mediterráneo, para evaluar si la radiación observada en la superficie disminuía considerablemente al protegerse bajo una gran masa de agua. Pacini notó una reducción significativa en la radiación al aumentar la profundidad, concluyendo que existía una componente significativa de la radiación proveniente del espacio exterior.

Los estudios de Victor Hess, llevados a mayores alturas mediante vuelos en globos aerostáticos equipados con electroscopios, profundizaron en la comprensión de la radiación. En 1911, ascendió aproximadamente 1100 metros, sin detectar variaciones significativas

[1] Energía emitida en forma de partículas u ondas electromagnéticas capaz de ionizar átomos, es decir, de arrancar electrones de ellos.

[2] Dispositivo utilizado para detectar la presencia y polaridad de cargas eléctricas mediante la observación de la separación de láminas metálicas.

en la cantidad de radiación en comparación con la superficie terrestre. Posteriormente, el 17 de abril de 1912, durante un eclipse solar casi total, Hess alcanzó los 5300 metros de altura, la fotografía 1.1 fue tomada en el aterrizaje de dicho vuelo. Dado que la ionización atmosférica no disminuyó durante el eclipse, concluyó que la fuente de la radiación no podía ser el Sol, sino que debía originarse en regiones más lejanas en el espacio [1]. Estos hallazgos fueron reafirmados después por Werner Kolhörster en 1913 en vuelos realizados hasta más de 9000 metros de altitud.

Figura 1.1: Fotografía célebre de Victor Hess, dentro de la canasta de un globo aerostático, en 1912. *Fotografía de dominio público.*

Robert Milikan fue quien introdujo el concepto de rayos cósmicos en 1925, haciendo referencia a su naturaleza proveniente del espacio exterior.

Entre los años 1927 y 1928, durante un viaje desde Indonesia a los Países Bajos, Jacob Clay observó que la intensidad aumentaba a medida que se alejaba del ecuador. Este hallazgo indicaba que la radiación cósmica era influenciada por el campo geomagnético, lo que implicaba que estaba compuesta principalmente por partículas cargadas.

Poco después, en 1929, Skobeltsyn empleó una cámara de niebla con el propósito de investigar las propiedades de los electrones resultantes de las desintegraciones radiactivas. Durante las observaciones, identificó ciertas trazas que presentaban desviaciones, semejantes a las de los electrones, pero que se curvaban en dirección opuesta cuando se les aplicaba un campo magnético [2]. Skobeltsyn contribuyó al descubrimiento del positrón mediante la implementación de un campo magnético en su cámara de niebla y la identificación de rayos cósmicos de partículas cargadas. Para entonces, se inventó el detector Geiger-Müller, que permitió detectar rayos cósmicos individuales y determinar sus tiempos de llegada con gran precisión [3].

El 2 de agosto de 1932, Anderson descubrió el positrón, antipartícula del electrón (véase figura 1.2 izquierda). En ese mismo año, Blackett y Occhialini facilitaron la aplicación en cámaras de niebla con contadores Geiger-Müller para la observación de partículas cargadas (ver figura 1.2 derecha) [4].

Estos experimentos demostraron que los rayos cósmicos iniciaban chubascos de partículas cargadas.

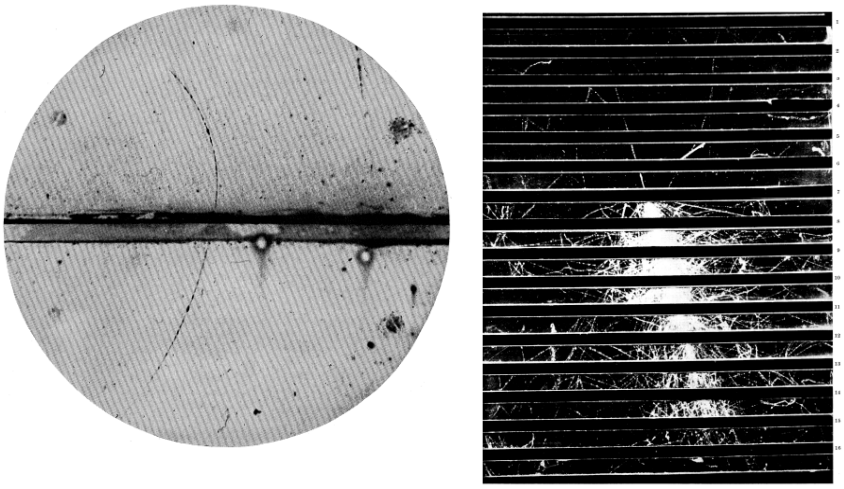

Figura 1.2: *Izquierda:* Fotografía de la trayectoria de un positrón. Se dispuso una lámina de plomo de 6 mm, como separador entre las secciones superior e inferior de la cámara. La trayectoria del positrón inicia desde la parte inferior, dado que la curvatura de la traza en la sección superior, bajo la influencia del campo magnético, tiene mayor intensidad. Que indica una perdida de energía al atravesar la lamina de plomo. *Imagen y referencia:* [5]. *Derecha:* Registro en cámara de niebla a 3027 m de altitud de una cascada generada por un protón de $\sim$ 10 GeV. La interacción inicial ocurre en una placa de plomo de 13 mm, seguida de la producción de piones neutros y cargados que amplifican la cascada e interactúan o decaen en muones visibles. *Fuente:* [4].

Fue el año 1933, gracias a los trabajos independientes de Rossi y Johnson, lograron confirmar que estas radiaciones estaban compuestas por partículas cargadas, y que, en efecto, las partículas se desviaban al interactuar con el campo magnético terrestre.

Un avance significativo ocurrió en 1937, de la mano del físico francés **Pierre Auger** y sus colegas, que realizaron experimentos pioneros al estudiar las coincidencias entre detectores colocados a varios metros de distancia.

A través de estas observaciones, descubrieron los llamados chubascos atmosféricos extensos[3] **(EAS, por sus siglas en inglés)**, que sentaron las bases para el estudio experimental de los rayos cósmicos a gran escala [6].

Aunque la comunidad científica tardó varios años en aceptar la existencia de los rayos cósmicos, Victor Hess finalmente fue galardonado con el Premio Nobel en 1936 por su

[3]Fenómeno en el que los rayos cósmicos, al chocar con la atmósfera terrestre, generan una cascada de partículas secundarias.

descubrimiento. Para entonces, aunque muchas preguntas seguían sin respuesta, los rayos cósmicos se convirtieron en herramientas para explorar el mundo subatómico, sentando las bases de la física de partículas moderna. Hasta el inicio de la década de 1950, esta metodología constituyó el principal recurso para la identificación de nuevas partículas.

Para 1953, la física de partículas y la de los rayos cósmicos tomaron caminos separados. La tecnología de aceleradores de partículas permitió avances significativos. A su vez, los rayos cósmicos no perdieron relevancia. En las últimas décadas, han experimentado un resurgimiento, especialmente al detectarse partículas con energías extremas, como el famoso rayo cósmico Oh-My-God de 3.2×10^{20} eV detectado el 15 de octubre de 1991 en Utah, Estados Unidos, por el detector Fly's Eye [7]. Tres décadas después se ha detectado Amaterasu[4], que registró una energía de alrededor de 2.4×10^{20} eV, observado el 27 de mayo de 2021 por el detector Telescope Array en Utah [8]. Este evento demostró que el universo naturalmente produce energías inimaginables, superando con creces lo alcanzable en aceleradores terrestres como el Gran Colisionador de Hadrones (LHC).

Hoy, los rayos cósmicos ya no son solo herramientas para explorar el mundo subatómico, sino mensajeros que nos permiten estudiar fenómenos cósmicos distantes. Han revelado que hay más en el cosmos de lo que podemos ver; aceleradores naturales de partículas, campos magnéticos intensos, y posiblemente, materia oscura, que podría constituir la mayor parte de la masa del universo. En la actualidad, observatorios puestos en tierra como el Observatorio Pierre Auger, continúan explorando los misterios de los rayos cósmicos. Aunque sabemos que están compuestos principalmente de protones y otras partículas cargadas, muchas preguntas permanecen sin respuesta: ¿Cuál es su origen? ¿Qué mecanismos los aceleran? ¿Qué nos dice sobre el medio interestelar? Estas interrogantes siguen impulsando la investigación en astropartículas, manteniendo vivo el interés por este campo de estudio.

1.2 Estudio de los rayos cósmicos

Investigar el origen de los rayos cósmicos implica explorar algunos de los fenómenos más extremos y sorprendentes del universo. Estas partículas altamente energéticas nos ofrecen información sobre eventos astrofísicos como las supernovas, colisiones galácticas y regiones del espacio con campos magnéticos intensos.

El análisis de los rayos cósmicos comienza con su detección. Dado que estas partículas cargadas interactúan con la atmósfera terrestre antes de llegar a la superficie, su observación directa requiere superar esta barrera natural. Esta interacción atmosférica protege la vida en la Tierra de radiaciones dañinas. Hoy, los métodos de detección se dividen en dos enfoques principales: experimentos directos, que buscan capturar las partículas antes de que interactúen con la atmósfera, y experimentos de superficie, que estudian las partículas secundarias generadas por estas interacciones [9].

[4]En honor a la diosa del Sol y del universo en la mitología japonesa.

1.2.1 Detección directa: más allá de la atmósfera

Los experimentos directos, como los realizados con globos estratosféricos o satélites en órbita terrestre baja, permiten observar los rayos cósmicos antes de que colisionen con la atmósfera. Estos métodos son ideales para estudiar partículas de menor energía (energías muy por debajo de 10^{15} eV), dado que su tasa de llegada es lo suficientemente alta como para ser detectadas con instrumentos de tamaño moderado. Sin embargo, a energías más altas, el flujo[5] de rayos cósmicos disminuye drásticamente, como se ilustra en la figura izquierda de 1.3: mientras que a 10^8 eV se detectan alrededor de 100 partículas por metro cuadrado por segundo, a 10^{15} eV (3 PeV) la tasa se reduce a una partícula por metro cuadrado por año, y a energías extremas por encima de 5×10^{18} eV (5 EeV), se espera solo una partícula por kilómetro cuadrado por siglo [10]. Esta escasez de eventos hace que los experimentos directos sean inviables para estudiar las partículas más energéticas, ya que requerirían detectores de tamaño imposible para implementar en el espacio.

1.2.2 Detección indirecta: la atmósfera como laboratorio

Los experimentos de superficie utilizan la atmósfera como un detector natural. Cuando un rayo cósmico choca con un núcleo atmosférico, genera una cascada de partículas secundarias, conocida como EAS, por sus siglas en inglés. Estas cascadas pueden extenderse por varios kilómetros cuadrados en la superficie, lo que permite su detección mediante arreglos de detectores, como los utilizados en el Observatorio Pierre Auger [9], que se discutirá en detalle en el siguiente capítulo. Aunque este enfoque no permite observar directamente el rayo cósmico primario, proporciona información muy importante sobre su energía, dirección y composición, aspectos fundamentales para comprender los mecanismos de aceleración y los objetos astrofísicos que los originan.

1.3 Espectro de rayos cósmicos

El espectro energético constituye uno de los observables fundamentales en la física de astropartículas. Este fenómeno presenta una amplia variedad de rayos cósmicos primarios; entre ellos destaca la abundancia de electrones, protones, núcleos de oxígeno, hierro y otros elementos químicos. Además, incluye rayos cósmicos secundarios, que son generados a partir de la interacción de los rayos cósmicos primarios con el medio interestelar o la atmósfera terrestre. Aunque menos abundantes, encontramos litio, berilio y boro, entre otros. El espectro abarca un rango energético excepcionalmente amplio, que se extiende desde el orden de 10^9 eV hasta valores superiores a 10^{20} eV. En primera aproximación, el espectro puede describirse mediante una ley de potencia de la forma $N(E) \propto E^{-\gamma}$, donde γ corresponde al índice espectral ($\gamma \approx -2.7$). La función $N(E)$ denota el número de partículas por unidad de área, tiempo, ángulo sólido y energía, es decir, el flujo diferencial de rayos cósmicos. El parámetro γ experimenta variaciones significativas en función del intervalo de energía [11].

[5] Magnitud física que define el número de partículas que impactan una superficie (o ángulo sólido) por unidad de tiempo y energía.

A lo largo del espectro se identifican tres características significativas que permiten inferir detalles fundamentales sobre los mecanismos de aceleración y propagación de los rayos cósmicos (figura 1.3, derecha). La primera estructura es la denominada rodilla (o knee en inglés), observada aproximadamente entre 3×10^{15} y 5×10^{15} eV. En esta región, el índice espectral cambia notablemente de $\gamma \approx -2.7$ a $\gamma \approx -3.1$. Se cree que la mayor parte de los rayos cósmicos, al menos hasta la rodilla, se originan dentro de nuestra galaxia, siendo probablemente acelerados en su mayoría por los remanentes de supernovas [12], [13].

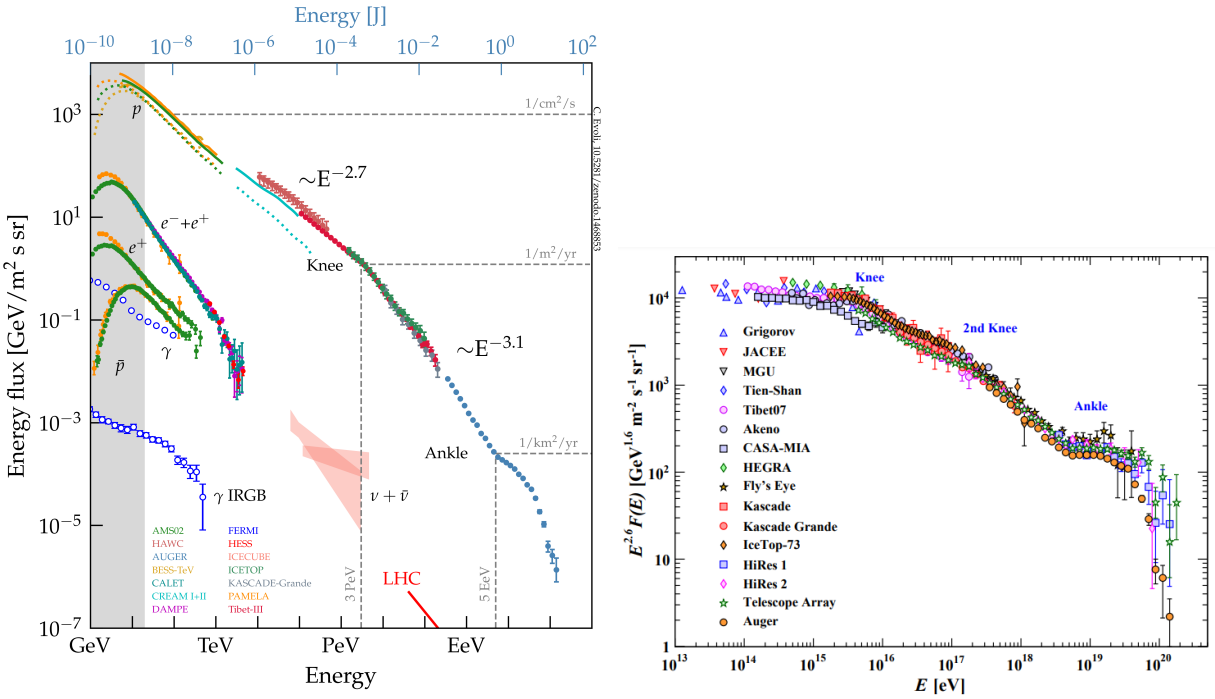

Figura 1.3: *Izquierda:* Espectro de energía de diferentes tipos de rayos cósmicos (flujo en función de la energía). *Fuente imagen:* [14]. *Derecha:* El espectro combinado de energía multiplicado por una potencia de $E^{2.6}$ obtenido por diversos experimentos, para resaltar las estructuras resultantes de los cambios en el índice espectral. *Imagen tomada de:* [15].

A energías superiores, alrededor de 1×10^{17} eV, se observa otra estructura conocida como la segunda rodilla (second knee), aquí el índice espectral varía de aproximadamente $\gamma \approx -3.0$ a hasta una nueva disminución de la pendiente cercana a $\gamma \approx -3.3$, dependiendo del estudio particular considerado [16]. Este cambio en la pendiente aún no está completamente claro, pero se relaciona estrechamente con una posible transición en la composición de masa de los rayos cósmicos, presentando una fracción dominante de núcleos más pesados, probablemente originada de diferencias en la eficiencia de aceleración y en la propagación de astropartículas[6] según la rigidez de las partículas [17].

La última estructura prominente del espectro es el tobillo (ankle), identificado aproximadamente a 5×10^{18} eV, en donde el índice espectral se reduce nuevamente, con valores

[6]Depende casi exclusivamente de la relación entre la energía y la carga de la partícula E/Z.

reportados alrededor de $\gamma = -2.7$ [18], [19]. Esta característica se interpreta generalmente como la región de transición entre rayos cósmicos de origen galáctico y aquellos de origen extragaláctico. A estas energías, el radio de Larmor[7] de las partículas es comparable al tamaño de la Vía Láctea, dificultando considerablemente su confinamiento dentro de la galaxia y favoreciendo una dominancia de partículas extragalácticas, especialmente protones precedentes de núcleos activos de galaxias (AGNs) y eventos astrofísicos altamente energéticos [19].

Finalmente, a energías extremas alrededor de 1×10^{20} eV, se observa un corte; el origen sigue siendo objeto de debate. Un intento de explicar este fenómeno es por el conocido límite Greisen-Zatsepin-Kuzmin (GZK). Este límite surge debido a interacciones inevitables entre los rayos cósmicos ultraenergéticos y el fondo cósmico de microondas[8] (CMB) que está presente en todo el universo. En este proceso, los rayos cósmicos pierden energía, restringiendo la distancia desde la cual las partículas pueden llegar a la Tierra. Dicho efecto lleva a una conclusión interesante, de que las partículas deben tener su origen en fuentes situadas a una distancia menor de 100 Mpc[9] [20], [21]. O en su defecto, la posibilidad de que el corte energético se origine de la energía máxima que los aceleradores cósmicos pueden proporcionar. Lo que plantea nuevas interrogantes sobre los mecanismos físicos que rigen tales procesos [10].

Mediciones recientes proporcionadas por observatorios como el Pierre Auger y Telescope Array han permitido estudiar detalladamente la composición química del espectro de alta energía, revelando una evolución desde partículas predominantemente ligeras en energías cercanas al tobillo hacia núcleos progresivamente más pesados en el régimen ultra energético. Esto sugiere que las características de las fuentes extragalácticas y su entorno inmediato influyen sustancialmente en la composición química y en la pendiente del espectro observado [22]. Actualmente, investigaciones adicionales se orientan hacia la identificación de fuentes extragalácticas específicas, mediante el análisis de anisotropías, que es la variación direccional de llegada en la distribución del flujo de rayos cósmicos más energéticos. Aunque aún no se han identificado fuentes individuales, estos estudios proporcionan información crucial sobre los entornos físicos extremos capaces de acelerar partículas hasta energías excepcionalmente altas [18].

1.3.1 Fuentes

La identificación y caracterización de las fuentes de rayos cósmicos constituye uno de los desafíos fundamentales en la física de astropartículas. Aunque en esta subsección se presenta una breve descripción sobre las fuentes principales, es importante enfatizar, querido lector, que estos son tópicos de investigación activa y continua. Aquí resumimos muy brevemente las principales categorías de fuentes de los rayos cósmicos.

[7] Radio de curvatura de la trayectoria de una partícula cargada que se mueve en un campo magnético.

[8] Radiación electromagnética residual que se cree se originó durante el Big Bang. Está compuesta por fotones de baja energía, influyendo en la propagación de los rayos cósmicos.

[9] Un megaparsec equivale a un millón de parsecs, donde un parsec es la distancia a la que una unidad astronómica (UA, la distancia media entre la Tierra y el Sol) subtiende un ángulo de un segundo de arco. Equivale aproximadamente 3.26 años luz, casi 3.086×10^{13} Km.

Rayos cósmicos solares

Los rayos cósmicos solares están constituidos principalmente por protones, electrones y núcleos ligeros acelerados durante fenómenos solares altamente energéticos como las erupciones solares y las eyecciones de masa coronal. El proceso dominante para la aceleración de estas partículas es la reconexión magnética rápida en regiones activas de la corona solar. Estos eventos solares pueden acelerar partículas hasta energías típicamente en el rango de unos pocos MeV hasta varios GeV [23], [24].

Rayos cósmicos galácticos

La mayoría de los rayos cósmicos detectados en la Tierra son de origen galáctico. Las fuentes más prominentes son los remanentes de supernovas (SNR), capaces de acelerar partículas hasta energías en el orden de 10^{15} a 10^{17} eV mediante el mecanismo de aceleración por choque difusivo[10]. Además de los SNR, objetos compactos tales como pulsares, estrellas de neutrones, microcuásares y nebulosas asociadas a pulsares, son también considerados fuentes importantes debido a sus intensos campos magnéticos y procesos dinámicos altamente energéticos [13], [11]. La composición química en esta región energética muestra predominancia de protones, núcleos de helio y elementos más pesados hasta el hierro [17].

Rayos cósmicos extragalácticos

En las energías más altas del espectro energético, específicamente por encima de 10^{18} eV, los rayos cósmicos son predominantemente de origen extragaláctico. Como mencionamos en la sección anterior, se han propuesto varias fuentes potenciales, incluyendo núcleos activos de galaxias, estallidos de rayos gamma (GRB) y chorros relativistas provenientes de agujeros negros supermasivos en AGNs.

1.3.2 Mecanismos de aceleración

Los mecanismos de aceleración más estudiados incluyen la aceleración de Fermi (en sus variantes de primer y segundo orden) y la aceleración difusiva por choque, los cuales explican cómo las partículas pueden alcanzar energías extremas en diversos entornos astrofísicos. Aquí hablamos brevemente sobre ello.

En 1949, Enrico Fermi propuso un mecanismo en el que las partículas cargadas ganan energía al interactuar con nubes interestelares y campos magnéticos turbulentos. En este modelo, conocido como aceleración de Fermi de segundo orden, las partículas son reflejadas por espejos magnéticos asociados con irregularidades en el campo magnético galáctico. Estas reflexiones ocurren de manera aleatoria, y las partículas ganan energía de forma estocástica en cada colisión [25].

La energía promedio ganada por colisión está dada por:

$$\left\langle \frac{\Delta E}{E} \right\rangle = \frac{8}{3}\left(\frac{V}{c}\right)^2, \tag{1.3.1}$$

[10]Implica la aceleración repetida de partículas mediante múltiples cruces de ondas de choque astrofísicas, incrementando así su energía.

donde V es la velocidad de las nubes interestelares y c es la velocidad de la luz. Este término cuadrático en V/c hace que la ganancia de energía sea relativamente pequeña, especialmente dado que $V/c \approx 10^{-4}$ en el medio interestelar. Aunque este modelo predice un espectro de energía de tipo ley de potencia $N(E) \propto E^{-\gamma}$, no explica por qué el índice espectral observado ($\gamma \approx -2.7$) es consistente en diversas mediciones [26].

Una versión más eficiente del mecanismo de Fermi, conocida como aceleración de Fermi de primer orden, ocurre cuando las partículas interactúan con ondas de choque supersónicas, como las producidas por explosiones de supernovas o AGNs. En este caso, la ganancia de energía es lineal en V/c, lo que hace que el proceso sea mucho más efectivo [26]:

$$\left\langle \frac{\Delta E}{E} \right\rangle = \frac{4}{3}\left(\frac{V}{c}\right). \tag{1.3.2}$$

Este mecanismo predice un espectro de energía $N(E) \propto E^{-2}$, que se acerca más al valor observado de $\gamma \approx -2.7$. La diferencia puede atribuirse a efectos de propagación en el medio interestelar, como la difusión y las pérdidas de energía [26].

El modelo más aceptado para la aceleración de rayos cósmicos es la aceleración difusiva por choque, una variante del mecanismo de Fermi de primer orden.

En este mecanismo, las partículas cargadas son aceleradas al interactuar repetidamente con una onda de choque. Siguiendo a [26], supongamos que una partícula con energía inicial E_0 experimenta una colisión con una onda de choque, ganando una fracción de energía β en cada cruce. Si P es la probabilidad de que la partícula permanezca en la región de aceleración después de una colisión, después de n colisiones, el número de partículas restantes será $N = N_0 P^n$, y su energía será $E = E_0 \beta^n$. Esto implica que el número de partículas con energía mayor o igual a E sigue una ley de potencias:

$$N(E) \propto E^{-1+\ln(P)/\ln(\beta)}, \text{ donde } \frac{\ln(P)}{\ln(\beta)} \approx -1 \tag{1.3.3}$$

En el caso de colisiones frontales, la ganancia de energías es de primer orden en V/c, lo que simplifica el modelo y permite predecir un espectro de energía $N(E) \propto E^{-2}$ [26].

Las fluctuaciones en el campo magnético dispersan las trayectorias de las partículas, permitiéndoles permanecer en la región de choque y ganar energía de manera significativa. Este mecanismo predice un espectro de energía $\gamma = -2$, que coincide con las observaciones en muchas fuentes astrofísicas. Esto podría explicar cómo se aceleran partículas cerca de la rodilla a energías extremas observadas hasta 10^{20} eV.

1.4 Chubascos atmosféricos extensos

Cuando los rayos cósmicos, típicamente compuestos por protones o núcleos más pesados, ingresan en la atmósfera superior terrestre, interactúan con núcleos atmosféricos como el nitrógeno u oxígeno. Esta interacción desencadena procesos hadrónicos que producen múltiples partículas secundarias, generando así una cascada o chubasco de partículas que se propaga hacia la superficie terrestre (ver figura 1.4 arriba). La cascada se expande rápidamente a medida que se producen sucesivas interacciones nucleares y electromagnéticas,

alcanzando un máximo número de partículas a una profundidad atmosférica específica, conocida como X_{max}, que varía según la energía del rayo cósmico primario [27]. La relevancia del estudio de estos eventos radica en que permite reconstruir información sobre la composición, energía y dirección de llegada de los rayos cósmicos primarios de altas energías (más de 10^{14} eV).

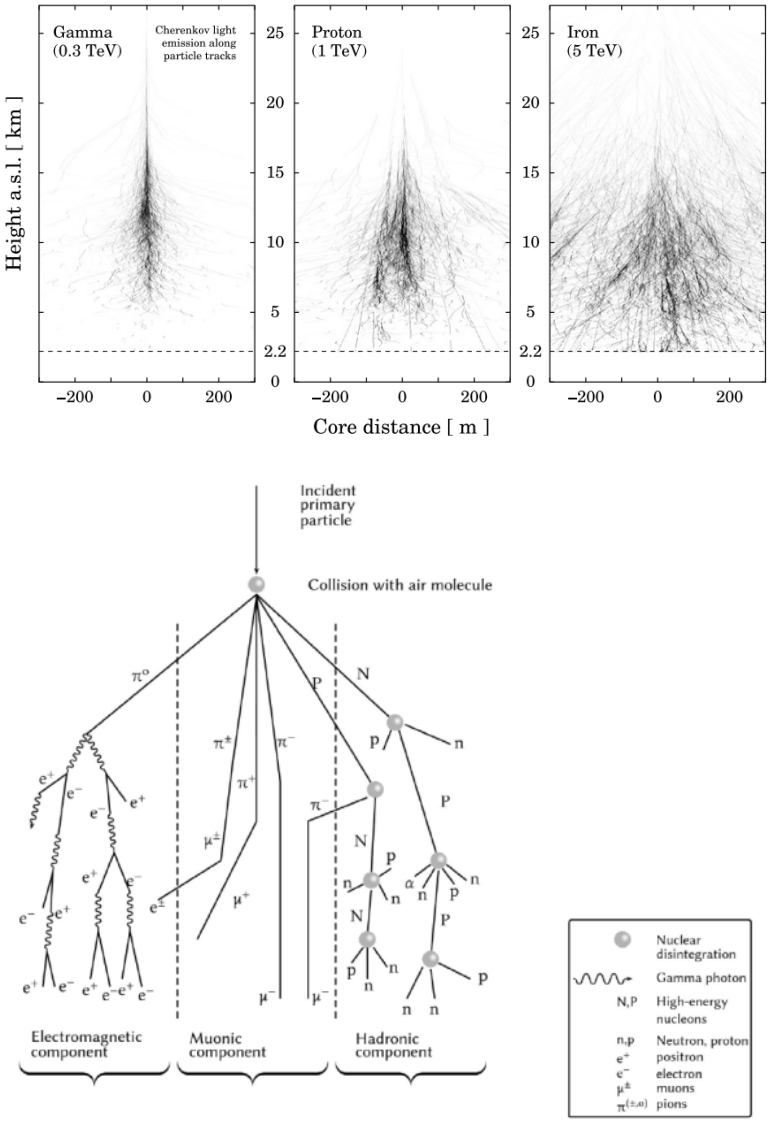

Figura 1.4: *Arriba:* Simulaciones realizadas con CORSIKA de EAS generadas por partículas primarias: rayo gamma, protón y núcleo de hierro. Mostrando la producción de luz Cherenkov. Esta representación permite visualizar la extensión espacial de la cascada atmosférica y su desarrollo en la atmósfera. *Fuente imagen:* [28]. *Abajo:* Esquema de un EAS. La partícula incidente colisiona con una molécula de aire, generando tres componentes principales: electromagnético (fotones gamma, electrones y positrones), muónico (muones) y hadrónico (nucleones, piones y otros hadrones). El proceso incluye desintegraciones nucleares y la producción de subcascadas. *Fuente imagen:* [29].

La estructura de un chubasco atmosférico extenso se puede describir en términos generales mediante tres componentes principales: electromagnética, muónica y hadrónica (véase figura 1.4 abajo).

1.4.1 Componente electromagnética

La componente electromagnética es la más abundante, representando aproximadamente el 90 % del número total de partículas en la cascada [17]. Esta componente se origina principalmente por el decaimiento rápido de piones neutros π^0, generados en las interacciones hadrónicas iniciales. Dichos piones decaen casi inmediatamente en pares de fotones según la reacción: $\pi^0 \rightarrow \gamma + \gamma$, generando así una cascada de electrones, positrones y fotones por medio de producción de pares[11] y radiación bremsstrahlung[12] o radiación de frenado. Estas interacciones generan sucesivamente más partículas, hasta alcanzar una energía crítica E_c, debajo de la cual predominan procesos de ionización y absorción de $\gamma \rightarrow e^- + e^+$, $e^-(e^+) \rightarrow e^-(e^+) + \gamma$.

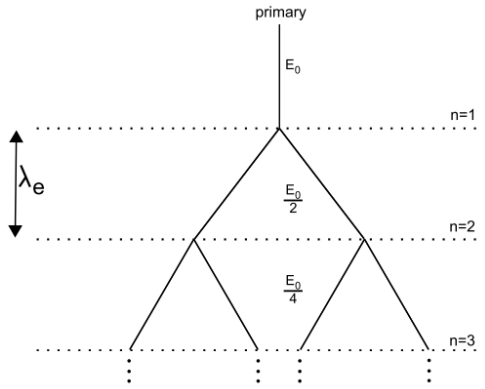

Figura 1.5: Representación esquemática del modelo simplificado de Heitler para el desarrollo de la componente electromagnética. Se ilustra la cascada de partículas iniciada por un primario de energía E_0, mostrando la división sucesiva de energía y la longitud de interacción λ_e. *Imagen tomada de:* [30]

.

El modelo simplificado propuesto por Heitler (véase figura 1.5) describe cualitativamente el crecimiento exponencial del número de partículas en función de la profundidad atmosférica X:

$$N(X) = 2^{X/\lambda}, \tag{1.4.1}$$

donde λ es la longitud de interacción o camino libre medio. La energía de las partículas a

[11] Transformación de un fotón en un electrón y un positrón, que ocurre en la proximidad de un núcleo atómico para conservar el momento. Este proceso requiere que el fotón tenga una energía mínima equivalente a la masa combinada del electrón y el positrón.

[12] Radiación electromagnética producida cuando una partícula cargada, al ser frenada o desviada por un campo eléctrico fuerte (generalmente el de un núcleo atómico), pierde energía cinética, la cual se emite en forma de fotones.

una profundidad X está dada por:

$$E(X) = \frac{E_0}{2^{X/2}} \tag{1.4.2}$$

donde E_0 es la energía inicial del rayo cósmico primario. El número máximo de partículas N_{max} y la profundidad a la que ocurre X_{max} están relacionados con la energía crítica mediante:

$$N_{max} = \frac{E_0}{E_c}, \quad X_{max} = \lambda \ln\left(\frac{E_0}{E_c}\right), \tag{1.4.3}$$

[31].

1.4.2 Componente muónica

La componente muónica suele representar aproximadamente entre el 5 % y el 10 % del número total de partículas secundarias en un EAS [32]. Esta componente surge principalmente del decaimiento de piones y kaones cargados producidos en interacciones hadrónicas iniciales. Estas partículas decaen según los procesos: $\pi^+ \to \mu^+ + \nu_\mu$, $\pi^- \to \mu^- + \overline{\nu}_\mu$, generando muones altamente penetrantes capaces de alcanzar la superficie terrestre debido a su masa y baja interacción con la materia atmosférica.

Una de sus características más notables es que su sección eficaz de interacción, que describe la probabilidad de interactuar con la materia, es inversamente proporcional a su masa. Que les confiere una alta capacidad de penetración, permitiéndoles atravesar grandes espesores de material con una pérdida de energía comparativamente baja a la de los electrones. En contraste, los neutrinos, caracterizados por una sección eficaz extremadamente baja, rara vez son detectados directamente debido a su mínima interacción con la materia. No obstante, su contribución energética se incluye en los cálculos globales del chubasco, particularmente cuando se emplean técnicas como la fluorescencia [31], [6], de las que hablamos en el próximo capítulo.

El número de muones N_μ está relacionado con la energía del primario según la expresión general:

$$N_\mu = \left(\frac{E_0}{E_{dec}}\right)^\alpha, \tag{1.4.4}$$

donde E_{dec} es la energía de decaimiento típica de los piones y α es un exponente que depende de la multiplicidad de partículas secundarias en la interacción [27]. La componente muónica transporta información crucial sobre la composición química del primario incidente [17].

1.4.3 Componente hadrónica

La componente hadrónica se genera a partir de interacciones hadrónicas entre el rayo cósmico primario y los núcleos atmosféricos. Estas interacciones producen un chubasco complejo con partículas tales como piones cargados y neutros, kaones y nucleones secundarios. La evolución del número y energía de estas partículas en el chubasco depende fuertemente de la energía inicial del primario [26]. El modelo de Heitler puede extenderse

para describir la componente hadrónica (ver figura 1.6). En este caso, se asume que una partícula hadrónica primaria produce n_{tot} partículas secundarias en cada interacción, de las cuales dos tercios son piones cargados $\pi^{\pm}$ y un tercio son piones neutros π^0. Los piones neutros decaen rápidamente en fotones gamma: $\pi^0 \rightarrow \gamma + \gamma$, mientras que los piones cargados decaen en muones y neutrinos: $\pi^+ \rightarrow \mu^+ + \nu_\mu$ y $\pi^- \rightarrow \mu^- + \overline{\nu}_\mu$.

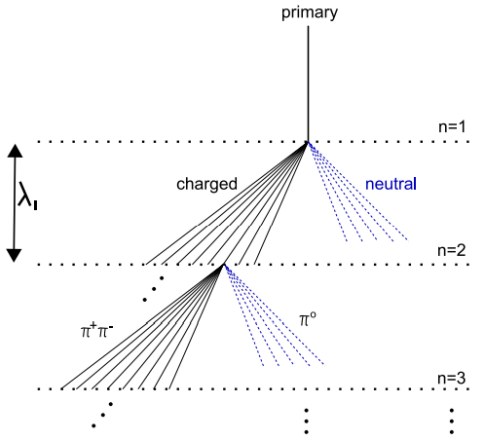

Figura 1.6: Extensión del modelo de Heitler para describir la componente hadrónica. Se muestra la cascada de hadrones, incluyendo la producción de piones $\pi^{\pm}, \pi^0$, con la longitud de interacción λ_I y los niveles de división n. *Imagen tomada de:* [30]

.

La componente hadrónica es fundamental para el desarrollo de la componente electromagnética de la cascada. Los piones neutros, producidos en las interacciones hadrónicas, decaen en fotones gamma, que a su vez generan cascadas electromagnéticas secundarias. La energía de las componentes hadrónica y electromagnética después de n interacciones está dada por:

$$E_h = \left(\frac{2}{3}\right)^n E_0, \quad E_{EM} = \left[1 - \left(\frac{2}{3}\right)^n\right] E_0, \tag{1.4.5}$$

Después de varias interacciones $n = 6$, la componente electromagnética porta la mayor parte de la energía $\sim 90\,\%$, incluso si la cascada fue iniciada por un hadrón [33].

Capítulo 2

El Observatorio Pierre Auger

El Observatorio Pierre Auger es actualmente el experimento más extenso en el mundo dedicado al estudio de rayos cósmicos de alta energía y de ultra alta energía (UHECR, por sus siglas en inglés). Ubicado en la Pampa Amarilla, en la provincia de Mendoza, cerca de la ciudad de Malargüe, en Argentina, cubre un área aproximada de 3000 km^2 (ver figura 2.1 izquierda) y está situado a una altitud media de 1400 metros sobre el nivel del mar, que corresponde a una profundidad atmosférica de aproximadamente 875 g/cm^2. Esta ubicación fue elegida por sus condiciones atmosféricas estables y baja contaminación lumínica (véase figura 2.1 derecha). Su principal objetivo es permitir estudios detallados del espectro energético, la composición química y la anisotropía de los UHECR [9].

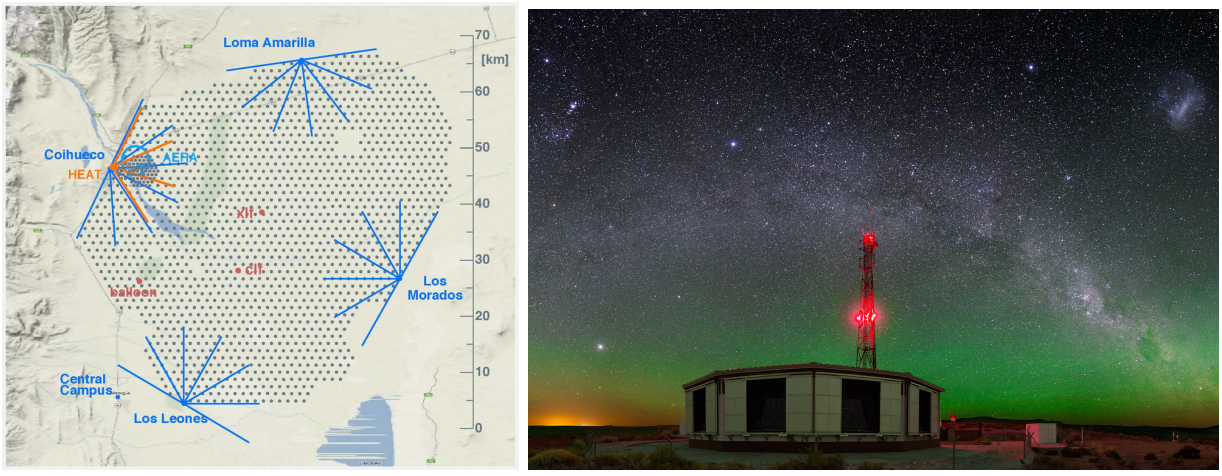

Figura 2.1: *Izquierda:* Distribución geográfica del Observatorio Pierre Auger, mostrando el arreglo de estaciones de SD representadas con puntos negros. Las líneas radiales azules que apuntan al centro del arreglo indican el campo de visión de los telescopios en las cuatro estaciones del detector de fluorescencia: Los Leones, Los Morados, Loma Amarilla y Coihueco, además del sistema de telescopios HEAT. *Imagen tomada de:* [34]. *Derecha:* Fotografía de la estación Los Morados. *Fotografía por:* Steven Saffi y Colaboración Pierre Auger.

Desde su inauguración el 1 de enero de 2004, el Observatorio Pierre Auger es un esfuerzo colaborativo internacional que involucra a más de 400 científicos de 18 países.

2.1 Descripción general

El observatorio opera bajo un diseño híbrido, combinando dos técnicas de detección complementarias: el detector de superficie (SD) y el detector de fluorescencia (FD). El SD, compuesto por una red de detectores Cherenkov, opera las 24 horas del día con una eficiencia cercana al 100 % para energías superiores a 3×10^{17} eV, mientras que el FD, que utiliza telescopios de fluorescencia, funciona durante noches despejadas y sin luna, proporcionando una medida calorimétrica[1] directa de la energía de los rayos cósmicos [35]. Los telescopios de FD miran hacia el interior del arreglo de SD, de modo que su campo de visión abarca la región cubierta por la matriz de SD (véase figura 2.2).

Aunque cada sistema de detección puede operar de manera independiente, su funcionamiento conjunto permite mejorar significativamente la precisión y confiabilidad de las reconstrucciones de rayos cósmicos.

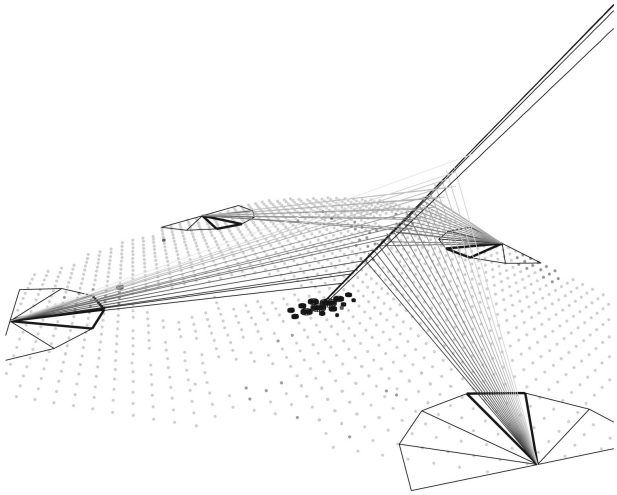

Figura 2.2: Esquema de la detección híbrida de un evento de rayo cósmico. Muestra la trayectoria de un EAS, registrada simultáneamente por el FD y el SD. *Fuente:* [9]. Los tanques del SD resaltados en color negro representan las estaciones que registraron una señal significativa del evento. Asimismo, se observa cómo un telescopio de cada edificio del FD se activa al detectar la luz de fluorescencia emitida por el chubasco atmosférico.

La capacidad del SD para funcionar de manera continua proporciona una gran cantidad de eventos observados, mientras que el FD, aunque limitado a condiciones, brinda información sobre la evolución longitudinal de los chubascos. Al integrar ambas técnicas, como se ilustra en la figura 2.2, se obtiene un conjunto de datos híbridos de alta calidad, lo que permite corregir incertidumbres sistemáticas y validar los métodos empleados en

[1] Esto implica que la cantidad de luz de fluorescencia emitida por las moléculas del aire es proporcional a la energía depositada por la cascada de partículas en la atmósfera.

el SD, mejorando la reconstrucción de la energía y dirección de los rayos cósmicos. Además, minimiza la dependencia de simulaciones en la estimación de la energía primaria, reduciendo la influencia de simulaciones de modelos hadrónicos [9].

2.2 Detector de superficie

El SD es una red de 1660 detectores Cherenkov distribuidos en forma triangular con una separación de 1.5 km entre estaciones (SD-1500); cada detector consiste en un tanque cilíndrico de polietileno de 3.6 m de diámetro y 1.2 m de altura, lleno con 12,000 litros de agua ultrapura. Adicionalmente, el sistema SD incorpora una matriz complementaria más pequeña de 71 detectores, donde las estaciones están separadas a 750 m (SD-750) cubriendo aproximadamente un área de 27 km^2. Esta configuración de 750 m se encuentra integrada dentro de la matriz de 1.5 km como se ilustra en la figura 2.3, con la finalidad de ampliar el rango de observación hasta energías del orden de 10^{17} eV [36].

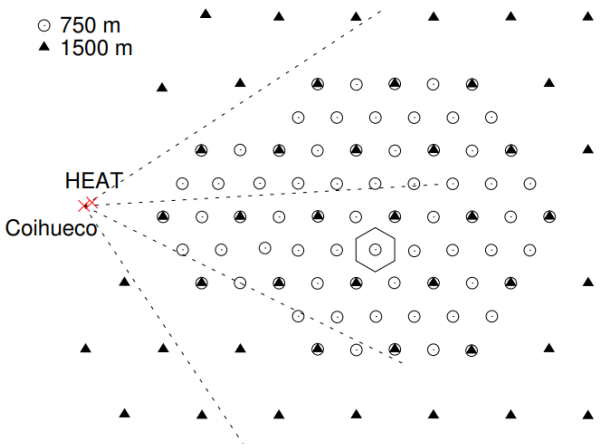

Figura 2.3: Esquema de la matriz SD-750, ubicada frente a las estaciones de FD: Coihueco y HEAT. Esta matriz alcanza una eficiencia del 100 % para energías superiores a 3×10^{17} eV, en EAS con ángulos cenitales menores a 55°, que implica una reducción del umbral energético respecto al SD-1500, que alcanza el 100 % de eficiencia para energías superiores a 3×10^{18} eV. *Fuente:* [36].

El agua dentro de los tanques actúa como medio para la producción de luz Cherenkov cuando es atravesada por partículas cargadas generadas en los EAS. Para optimizar la detección, el interior del tanque está revestido con un material altamente reflectivo que maximiza la captación de luz [9].

Cada tanque está equipado con tres fotomultiplicadores[2] (PMTs) de 9 pulgadas, ubicados simétricamente en la parte superior del tanque mirando hacia abajo. Estos PMTs detectan la luz Cherenkov emitida por las partículas secundarias (principalmente electrones, positrones y muones) que atraviesan el agua dentro del tanque. Las señales eléctricas

[2]Es un dispositivo detector extremadamente sensible que convierte fotones en señales eléctricas. Su funcionamiento se basa en el efecto fotoeléctrico, seguido de una serie de etapas de multiplicación electrónica.

generadas por los PMTs se digitalizan mediante convertidores analógicos-digitales flash [9], que registran tanto la amplitud como el tiempo de llegada de las partículas.

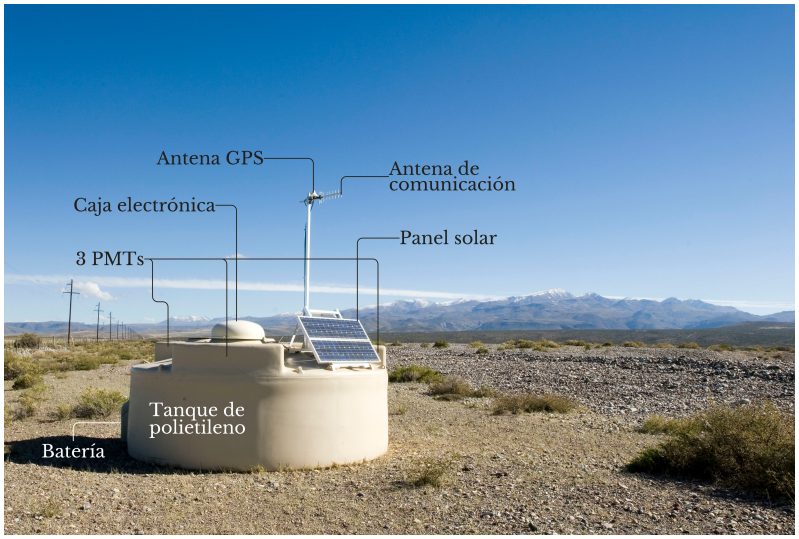

Figura 2.4: Fotografía modificada de uno de los tanques de SD-1500. *Fotografía por:* Colaboración Pierre Auger.

La sincronización temporal de las estaciones del SD se realiza mediante receptores GPS, montados en la parte superior del tanque junto con las antenas de comunicación. Además, cada estación del SD opera de manera independiente, alimentada por paneles solares y baterías recargables, lo que permite su operación continua en condiciones climáticas adversas (ver figura 2.4). La información registrada por los detectores se transmite mediante un sistema de comunicación inalámbrica a través de cuatro antenas instaladas junto a los edificios del FD. Estas antenas reciben la información desde las estaciones del SD y la retransmiten, junto con la información del FD, al Sistema Central de Adquisición de Datos (CDAS, por sus siglas en inglés), ubicado en la sede del observatorio en Malargüe. CDAS es responsable del procesamiento, almacenamiento y análisis de los datos [9].

Se emplea un sistema jerárquico de disparador para seleccionar eventos de interés científico y descartar señales de fondo no relevantes. Este esquema consta de cinco niveles de disparo, diseñados para filtrar y procesar los eventos antes de enviarlos al CDAS. Los disparadores locales (T1 y T2) operan en cada estación del SD, aplicando umbrales de energía y patrones de señal para descartar eventos irrelevantes. Posteriormente, los disparadores globales (T3, T4 y T5) refinan la selección en el CDAS, evaluando correlaciones espacio-temporales entre las estaciones y clasificando los eventos según su calidad y relevancia física. Un análisis más detallado del sistema de disparo del SD puede encontrarse en [37]. Este sistema jerárquico permite identificar con precisión chubascos atmosféricos extensos de alta energía, optimizando el ancho de banda de la red de comunicación y evitando la sobrecarga.

La precisión de las mediciones del SD depende de su calibración, este proceso se basa en el concepto del muón vertical equivalente (VEM) [9], definido como la señal promedio producida por un muón que atraviesa el tanque de forma vertical. Este procedimiento

permite establecer una escala común y uniforme para todas las estaciones. El proceso de calibración consiste en convertir las señales detectadas en unidades VEM. Para ello, se ajustan las ganancias de los PTMs en cada estación y se establece un sistema de referencia que permite la comparación con simulaciones del detector. La calibración se realiza de manera continua cada 60 segundos; los parámetros de calibración se actualizan y se almacenan junto con los datos de los eventos registrados [9]. Este procedimiento garantiza la fiabilidad de las mediciones y su compatibilidad con los datos obtenidos mediante el FD, facilitando así la reconstrucción híbrida de los rayos cósmicos.

2.2.1 Radiación de Cherenkov

La radiación Cherenkov es un fenómeno electromagnético que ocurre cuando una partícula cargada se mueve a través de un medio dieléctrico con una velocidad superior a la velocidad de la luz en dicho medio, generando un cono de luz característico. Este fenómeno fue descubierto por Igor Tamm e Ilya Frank y experimentalmente por Pavel Cherenkov en 1934, lo que les valió el Premio Nobel de Física en 1958 [38].

La condición para que se emita radiación de Cherenkov es que la velocidad de la partícula v sea mayor que la velocidad de la luz en el medio c/n, donde n es el índice de refracción del medio y c es la velocidad de la luz en el vacío:

$$v > \frac{c}{n}. \tag{2.2.1}$$

Para partículas relativistas, esta condición se traduce en un umbral de energía mínima E_{umbral} que la partícula debe superar para emitir radiación de Cherenkov. La energía umbral se calcula como [39]:

$$E_{\text{umbral}} = \frac{m_0 c^2}{\sqrt{1 - \frac{1}{n^2}}}, \tag{2.2.2}$$

donde m_0 es la masa en reposo de la partícula.

El ángulo de emisión θ de la radiación de Cherenkov está dado por la relación:

$$\cos\theta = \frac{1}{\beta n}, \tag{2.2.3}$$

donde $\beta = v/c$. Este ángulo define la geometría del cono de luz de Cherenkov.

La distribución espectral de la radiación de Cherenkov, es decir, el número de fotones emitidos por unidad de longitud de onda λ, está dado por la fórmula de Frank-Tamm [40]:

$$\frac{d^2N}{dxd\lambda} = \frac{2\pi\alpha Z^2}{\lambda^2}\left(1 - \frac{1}{\beta^2 n^2}\right), \tag{2.2.4}$$

donde, α es la constante de estructura fina ($\alpha \approx 1/137$), Z es la carga de la partícula en unidades de la carga del electrón, y λ es la longitud de onda de la luz emitida. Esta ecuación muestra que la radiación de Cherenkov es más intensa en longitudes de onda cortas (región del azul y ultravioleta), lo que explica el característico resplandor azulado observado en los detectores de Cherenkov.

Los tanques de SD aprovechan la radiación de Cherenkov para detectar partículas secundarias, con una eficiencia que depende de su energía y carga.

2.3 Detector de fluorescencia

El FD consta de 24 telescopios Schmidt[3] distribuidos en cuatro estaciones: Los Leones, Los Morados, Loma Amarilla y Coihueco, ubicadas en los bordes del arreglo SD. Cada estación consta de 6 telescopios que, en conjunto, cubren un campo de visión de 180° en azimut, como se ilustra en la figura 2.5. Cada telescopio cubre un campo de visión de 30° × 30° en elevación y azimut, lo que permite una cobertura tridimensional del 90 % de eventos por encima de 10^{19} eV, ofreciendo una resolución angular por píxel cercana a 1.5° [9]. Estos telescopios detectan la luz ultravioleta emitida por moléculas de nitrógeno cuando son excitadas por las partículas secundarias del chubasco. Dado que esta emisión es isotrópica, la intensidad de la luz de fluorescencia es proporcional a la energía del rayo cósmico primario.

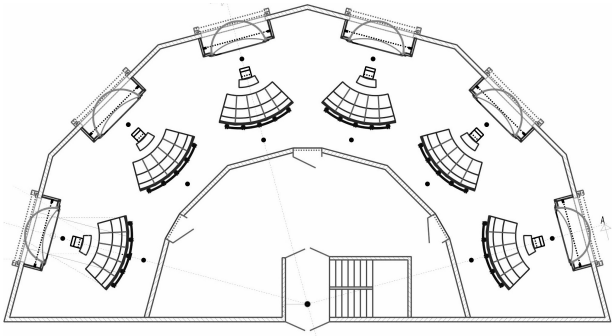

Figura 2.5: Esquema de las seis bahías que componen la estructura del edificio de los FD. Cada bahía alberga un telescopio. *Imagen por:* Colaboración Pierre Auger.

Como extensión del FD, el sistema HEAT incluye tres telescopios más, ubicados a 180 m frente a la estación Coihueco y diseñados para observar chubascos atmosféricos con ángulos elevados. Estos telescopios son estructuralmente similares a los del FD, pero incorporan un sistema hidráulico accionado eléctricamente que les permite cubrir un rango de elevación de 30° a 58°, complementando así el campo de visión de los del FD [9].

Aunque el FD opera aproximadamente el 15 % del tiempo total [9], su capacidad para medir la profundidad del máximo desarrollo (X_{max}) del EAS lo convierte en una herramienta esencial para estudiar la composición química de los rayos cósmicos. X_{max} es la profundidad atmosférica donde el chubasco alcanza la mayor producción de partículas secundarias, y su valor depende de la masa del núcleo primario: los núcleos ligeros, como protones, tienden a penetrar más en la atmósfera, lo que resulta en un X_{max} más profundo, mientras que los núcleos pesados, como hierro, interactúan antes en la atmósfera, lo que resulta en un X_{max} a menor profundidad atmosférica [41].

Cada telescopio está equipado con un filtro de entrada, el cual bloquea la luz visible, también con un espejo esférico segmentado de 3.3 m de diámetro que enfoca la luz ultravioleta emitida por las moléculas de nitrógeno excitadas en la atmósfera hacia una cámara de 440 fotomultiplicadores hexagonales. Para minimizar aberraciones ópticas y mejorar la

[3]Se basa en un tipo de telescopio de gran campo que emplea una combinación de un espejo esférico y una lente correctora para minimizar aberraciones ópticas.

eficiencia de detección, el FD incorpora un anillo de corrector, una adaptación del diseño óptico, que amplía el área efectiva de detección y duplica la sensibilidad del telescopio (véase figura 2.6 derecha) [35].

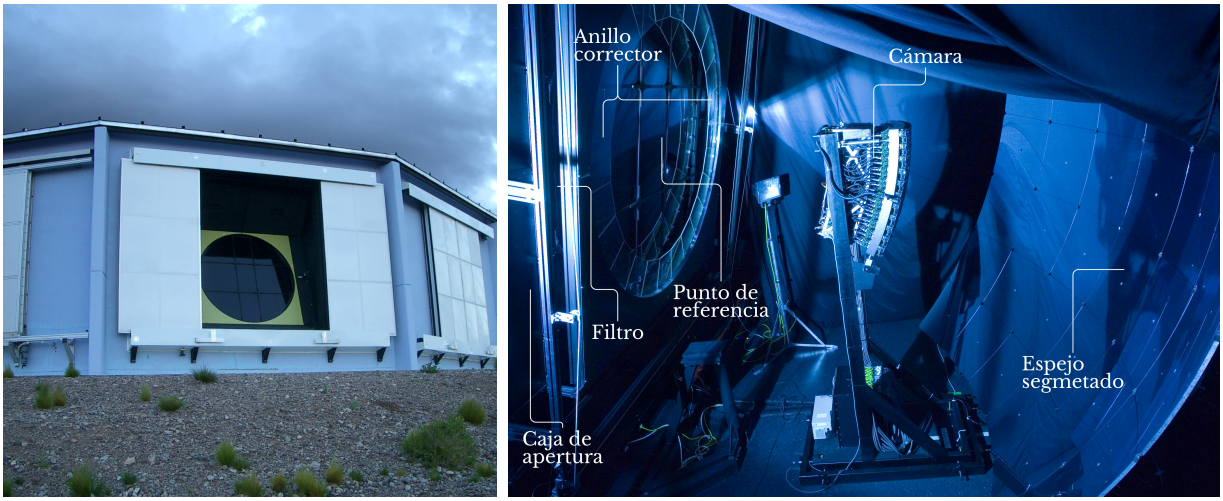

Figura 2.6: *Izquierda:* Vista de la estación Los Leones con una de sus bahías en posición abierta. *Derecha:* Fotografía modificada del interior de una bahía, mostrando la configuración del telescopio *Fotografías por:* Colaboración Pierre Auger.

A partir de las señales registradas por el FD, cada evento es transmitido al CDAS. Posteriormente, los eventos se combinan con la información del SD para realizar una reconstrucción híbrida de la energía y la dirección del rayo cósmico. Para garantizar precisión en las mediciones, el FD se calibra mediante dos métodos complementarios: una calibración absoluta basada en fuentes de luz controladas y una calibración relativa mediante un sistema de monitoreo continuo. Adicionalmente, se implementa un monitoreo atmosférico a través de la Instalación Central de Láser (CLF) y la Instalación eXtremo de Láser (XLF), que emiten pulsos láser con parámetros conocidos que permiten caracterizar las condiciones atmosféricas y corregir la atenuación de la luz de fluorescencia; una descripción más detallada de este proceso se encuentra en [35].

2.4 Métodos de reconstrucción

La física de los EAS varía de manera significativa con el ángulo cenital de llegada. A medida que este ángulo aumenta, la cantidad de atmósfera que atraviesa el chubasco también se incrementa, lo que provoca una atenuación considerable del componente electromagnético. En consecuencia, en eventos inclinados (mayores a 60°) predominan los muones al nivel del suelo, mientras que en eventos verticales (menores a 60°) el componente electromagnético es relevante y debe considerarse en la reconstrucción del evento.

Dada esta diferencia en la composición de las señales, se han desarrollado algoritmos de reconstrucción específicos para eventos verticales e inclinados definidos por la colaboración Pierre Auger. En este trabajo, se emplea exclusivamente la reconstrucción estándar aplicada a eventos con ángulos cenitales menores a 60°.

La reconstrucción híbrida de eventos en el Observatorio Pierre Auger combina las mediciones obtenidas por el FD y el SD. Si bien el proceso completo incluye la reconstrucción realizada con el FD, en lo que sigue nos centraremos en las principales etapas aplicadas al SD. Una descripción detallada del procedimiento completo puede consultarse en [9] y [42].

2.4.1 Selección de eventos

Para asegurar la calidad de los datos, se aplican dos disparadores adicionales fuera de línea: el disparador físico T4 y el corte fiducial 6T5. Estos disparadores aseguran que solo se seleccionen eventos bien contenidos y con una geometría definida.

- Disparador físico T4: Este nivel selecciona eventos físicos del conjunto de datos activado en línea por T3, eliminando señales de fondo o ruido. El criterio T4 se basa en coincidencias temporales entre estaciones adyacentes, verificando que los tiempos de llegada de las señales sean consistentes con la propagación del frente del chubasco en la dirección reconstruida [9].

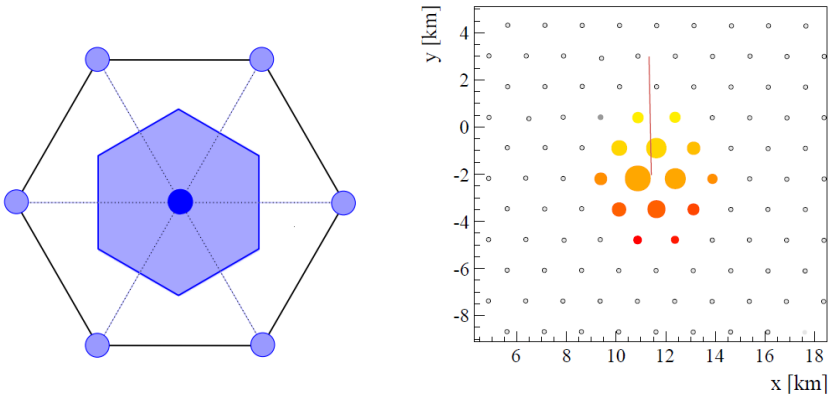

Figura 2.7: *Izquierda:* Esquema del disparador fiducial 6T5. Se define como tal a cualquier estación activa (el punto del centro) que cuente con seis estaciones vecinas también activas. *Imagen tomada de:* [43] *Derecha:* Esquema de la distribución espacial de las señales registradas por el SD-1500 en un evento. Cada círculo representa una estación activa, donde la intensidad del color indica el tiempo de llegada del frente del chubasco (de menos intenso para las más tempranas a intenso rojo para las más tardías). La línea trazada representa la dirección reconstruida de llegada del evento. *Fuente e imagen de:* [9].

- Corte fiducial 6T5: Entre los eventos que cumplen con el disparador físico T4, es necesario identificar aquellos que pueden ser reconstruidos con precisión. Para ello, se impone un criterio fiducial conocido como 6T5: se requiere que la estación con la mayor señal esté rodeada por seis vecinos operativos, formando un hexágono funcional como se ilustra en las figuras 2.7. Esto garantiza una cobertura geométrica adecuada alrededor del núcleo del chubasco, lo cual optimiza la determinación del punto de impacto [9]. Este criterio aprovecha la geometría hexagonal para estimar

la apertura del arreglo como múltiplo de las estaciones activas que permite calcular de forma sencilla la exposición del detector.

2.4.2 LDF

La función de distribución lateral (LDF, por sus siglas en inglés) describe cómo varía la amplitud de las señales en función de la distancia al núcleo del chubasco. Esta función permite estimar el tamaño del chubasco y, por tanto, la energía del rayo cósmico primario. La LDF se ajusta utilizando la función modificada de Nishamura-Kamata-Greisen (NKG), que tiene en cuenta las características específicas de los EAS detectados por el SD:

$$S(r) = S(r_{opt}) f_{\mathrm{LDF}}(r) = S(r_{opt}) \left(\frac{r}{r_{opt}}\right)^{\beta} \left(\frac{r + r_1}{r_{opt} + r_1}\right)^{\beta+\gamma}, \tag{2.4.1}$$

donde r es la distancia al núcleo de la lluvia, $r_{opt} = 1000$ m es la distancia óptima para el SD-1500, $r_1 = 700$ m es un parámetro de escala, y β y γ son parámetros que dependen del ángulo cenital y del tamaño del chubasco [42].

2.4.3 Reconstrucción geométrica

La reconstrucción geométrica comienza con la selección de eventos que hayan superado los criterios de disparo establecidos a nivel de estación. Esta configuración inicial permite realizar una primera estimación de la dirección geométrica del evento, la cual se reconstruye a partir de los tiempos de llegada de las señales registradas en estaciones activas.

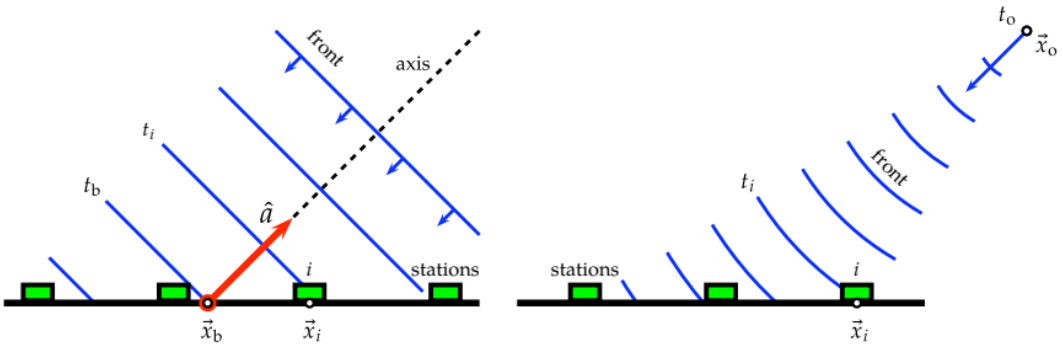

Figura 2.8: *Izquierda:* Representación esquemática del modelo de frente plano para describir el desarrollo de la cascada atmosférica. *Derecha:* Representación del modelo esférico del frente de cascada. *Imagen tomada de:* [42]

En una primera aproximación, se emplea un modelo de frente plano (ver figura 2.8 izquierda), que asume que el frente del chubasco se propaga como una superficie plana que avanza a la velocidad de la luz c, a lo largo del eje del chubasco $\hat{a}$. El tiempo $t(\vec{x})$ en el que el frente plano alcanza una posición cualquiera $\vec{x}$ se describe mediante la siguiente ecuación [42]:

$$c\, t(\vec{x}) = c\, \vec{t_b} - (\vec{x} - \vec{x}_b) \cdot \hat{a}, \tag{2.4.2}$$

donde $\vec{x}_b$ es el centro espacial de las estaciones activadas, calculado como el baricentro ponderado por la intensidad de las señales, al instante t_b, obtenido como el promedio ponderado de los tiempos de inicio de las señales.

Para eventos con al menos cuatro estaciones activas y LDF bien definida, se emplea un modelo de frente de onda esférico (véase figura 2.8 derecha). Este modelo describe la expansión del chubasco como una superficie esférica que se propaga a la velocidad de la luz, con un origen virtual en el espacio. La ecuación que describe este modelo es [42]:

$$c\, t(\vec{x}) = c\, t_0 + |\vec{x} - \vec{x}_0|\,, \tag{2.4.3}$$

donde $\vec{x}_0$ y t_0 son las coordenadas espaciales y temporales, respectivamente, del origen virtual del frente de onda esférico.

La dirección de llegada del chubasco se obtiene a partir del origen virtual $\vec{x}_0$ y el punto de impacto al nivel del suelo $\vec{x}_s$, que se determina a partir del ajuste de la LDF. La dirección del eje del chubasco $\hat{a}$ se calcula como [42]:

$$\hat{a} = \frac{\vec{x}_0 - \vec{x}_s}{|\vec{x}_0 - \vec{x}_s|}. \tag{2.4.4}$$

La resolución angular en la reconstrucción de EAS mejora conforme aumenta el número de estaciones activadas en un evento. En el caso de eventos de baja energía, donde típicamente se activan solo tres o cuatro estaciones, los datos temporales disponibles no permiten un ajuste confiable de la curvatura. Para estos casos, se recurre al modelo esférico manteniendo fijo el radio de la curvatura, el cual se determina a partir de una parametrización obtenida con eventos de mayor multiplicidad. Esta solución permite preservar la precisión de la reconstrucción, incluso cuando la información recolectada es limitada.

Capítulo 3

Aprendizaje automático

La revolución digital de las últimas décadas ha transformado a las ciencias empíricas en disciplinas centradas en el manejo de grandes volúmenes de datos. Este nuevo paradigma, impulsado por avances en computación y almacenamiento, dio origen a la ciencia de datos como un campo interdisciplinario. En física se generan flujos continuos de datos complejos y multidimensionales que, en algunos casos, superan las capacidades de los métodos analíticos tradicionales. Aunque el enfoque clásico parte de algoritmos basados en modelos teóricos, desarrollarlos no es tarea fácil, especialmente cuando el volumen de datos excede la capacidad de análisis manual o las variables presentan relaciones no lineales.

En este contexto, el aprendizaje automático (Machine Learning, ML) ha demostrado ser una alternativa eficaz para abordar estos desafíos. El ML introduce un cambio de perspectiva: en lugar de depender exclusivamente de algoritmos diseñados manualmente, **los modelos aprenden directamente a partir de los datos**, identificando patrones y relaciones complejas mediante procesos iterativos de optimización, lo que les permite modelar funciones desconocidas, sin requerir una formulación explícita del fenómeno subyacente.

3.1 Fundamentos

Desde un punto de vista formal, el aprendizaje automático puede entenderse como un conjunto de métodos que buscan aproximar funciones $f : X \to Y$, donde X representa variables de entrada y Y variables de salida. A través de la exposición a ejemplos etiquetados o no etiquetados, los algoritmos se ajustan a su estructura interna para generalizar el comportamiento observado y aplicarlo a nuevos datos no vistos.

3.1.1 Tipos de aprendizaje

Existen diversas categorías de aprendizaje automático, entre las que destacan tres: el aprendizaje supervisado, no supervisado y por refuerzo. Cada una de estas responde a distintos tipos de problemas y niveles de información disponible en los datos.

- **Aprendizaje supervisado**, este enfoque se utiliza cuando se dispone de un conjunto de datos etiquetados, es decir, pares de entradas y salidas deseadas. El objetivo es

encontrar una función f que, dada una nueva entrada x, prediga su correspondiente salida y. Este enfoque es común en problemas de clasificación, donde Y es una variable discreta de etiquetas, por ejemplo identificación de dígitos manuscritos y en problemas de regresión, donde Y es una variable continua, como predicciones. Para abordar estos problemas, se emplean diferentes algoritmos en los que destacan la regresión lineal y no lineal como los modelos de series temporales, las redes neuronales artificiales, las máquinas de soporte vectorial y los modelos basados en árboles de decisión.

- **Aprendizaje no supervisado**, se aplica cuando solo se cuenta con datos de entrada sin etiquetas asociadas. El objetivo es descubrir estructuras o patrones subyacentes en los datos. Una técnica común es el agrupamiento, que busca dividir el conjunto de datos $\{x_i\}_{i=1}^n$ en k grupos o clústers, de forma que las muestras dentro de cada clúster compartan una mayor similitud entre sí que con las de otros clústers. El algoritmo k-means, por ejemplo, minimiza la suma de las distancias cuadradas entre cada punto y el centroide de su clúster asignado. Este enfoque es útil para análisis exploratorio de datos y en reducción de dimensionalidad como el método de PCA.

- **Aprendizaje por refuerzo**, a diferencia de los enfoques supervisado y no supervisado, el aprendizaje por refuerzo se fundamenta en la interacción dinámica entre un agente y un entorno. En este marco, el algoritmo no aprende a partir de un conjunto de datos estáticos, sino que adquiere conocimiento a través de la experiencia acumulada al ejecutar acciones y observar sus consecuencias. Su objetivo es aprender una estrategia de comportamiento denominada política que le permita maximizar una medida acumulada de recompensa en el tiempo. Con base en esta experiencia, ajusta su comportamiento futuro para obtener mejores resultados. Este tipo de aprendizaje es especialmente útil en problemas de toma de decisiones secuenciales, como en tareas de navegación, control robótico, optimización de procesos o juegos de estrategia, donde no siempre es evidente cuál es la mejor acción en cada situación.

3.1.2 Componentes

La estructura de algoritmos de aprendizaje automático puede integrarse en cuatro elementos esenciales. Estos componentes permiten construir algoritmos capaces de aprender relaciones funcionales complejas a partir de datos [44]:

- **Conjunto de datos**, se puede definir como $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ el conjunto total de datos, donde $x_i \in \mathbb{R}^n$ es un vector de características o variables observables para la i-ésima muestra y y_i es la variable objetivo. En el caso de problemas de regresión, y_i puede ser un valor escalar, mientras que para clasificación, y_i es una etiqueta categórica. En un modelo generalmente se dividen en tres subconjuntos de datos, el conjunto de entrenamiento $\mathcal{D}_{\text{train}}$, el conjunto de validación $\mathcal{D}_{\text{val}}$, empleado para seleccionar hiperparámetros y monitorear el proceso de entrenamiento, y el conjunto de prueba $\mathcal{D}_{\text{test}}$ usado únicamente para evaluar el desempeño final del modelo.

- **Modelo**, es una función parametrizada $f(x;\theta)$ donde θ representa los parámetros del modelo. El aprendizaje consiste en encontrar los valores óptimos θ que mejor

explique la relación entre entradas y salidas según los datos observados. Este modelo puede ser lineal como una regresión lineal, no lineal como árboles de decisión o modelos complejos como en redes neuronales profundas.

- **Función de costo o pérdida** denotada como $\mathcal{L}(f(x;\theta), y)$ que cuantifica el error entre la predicción del modelo y el valor real.
- **Algoritmo de optimización**, el objetivo del algoritmo de optimización es minimizar la función de perdida $\mathcal{L}$ sobre los datos de entrenamiento, ajustando iterativamente los parámetros θ. Para modelos no lineales, este proceso no puede resolverse de forma analítica, por lo que se emplean métodos iterativos como el descenso por gradiente estocástico[1]. En la práctica, se utilizan variantes avanzadas como *Adam*, *RMSprop* o *momentum*, que son métodos de optimización de primer orden que utilizan el gradiente de la función de perdida.

3.1.3 Generalización

La capacidad de un modelo para generalizar es un aspecto fundamental: debe desempeñarse bien no solo sobre los datos de entrenamiento, sino sobre nuevos datos. Dos fenómenos clásicos pueden limitar este objetivo:

- **Subajuste (underfitting)** ocurre cuando el modelo tiene una capacidad insuficiente para captar la complejidad de los datos (ver el modelo lineal en la figura 3.1 de la izquierda).
- **Sobreajuste (overfitting)** ocurre cuando el modelo memoriza el ruido del conjunto de entrenamiento y falla al generalizar, ajustándose en exceso a los datos de entrenamiento (ver el modelo complejo en la figura 3.1 de la derecha).

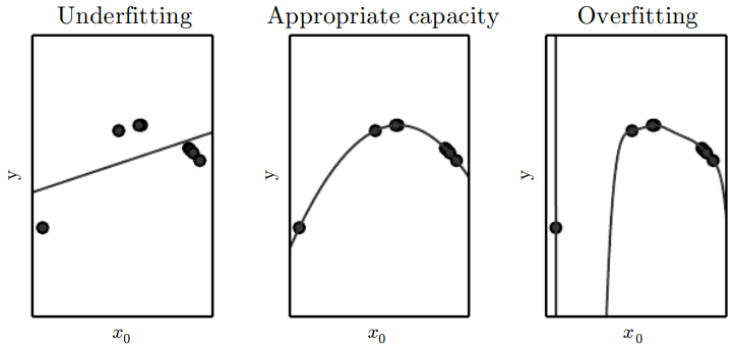

Figura 3.1: Ilustración del concepto de subajuste y sobreajuste. El gráfico central muestra un modelo con la capacidad adecuada para ajustarse a los datos. *Imagen tomada de:* [44]

El subajuste se asocia a alto sesgo y el sobreajuste a alta varianza. Alcanzar un equilibrio óptimo entre sesgo (error por simplificaciones) y varianza (error debido a la sensibilidad en los datos de entrenamiento) se logra mediante técnicas de regularización.

[1] Actualiza los parámetros del modelo usando el gradiente calculado en una muestra aleatoria (o lote pequeño) de los datos de entrenamiento.

Para evitar el sobreajuste, se incorporan técnicas de regularización que penalizan la complejidad del modelo. Una forma común es agregar un término de penalización a la función de pérdida, como la norma $L2$ (o Ridge) y $L1$ (o Lasso). Además, métodos como la interrupción temprana del entrenamiento (early stopping), la introducción de ruido aleatorio en la entrada, la eliminación de neuronas durante el entrenamiento (dropout) y la normalización de activaciones (batch normalization) son técnicas que permiten mejorar la capacidad de generalización de los modelos.

3.1.4 Evaluación

La evaluación de modelos se realiza para medir la capacidad de generalización, es decir, cómo se desempeñan con datos nuevos no vistos durante el entrenamiento. Este proceso consta de dos componentes principales: las métricas de evaluación y las estrategias de validación.

Métricas de evaluación

La selección de la métrica adecuada depende del tipo de problema abordado. Entre las más utilizadas, encontramos en problemas de regresión:

- **Error cuadrático medio (MSE)** mide el promedio de los errores al cuadrado.
- **Error absoluto medio (MAE)**, promedio de diferencias absolutas.
- **Coeficiente $\mathbf{R}^2$**, eficiencia predictiva relativa del modelo.

En problemas de clasificación:

- **Exactitud (Accuracy)**, porcentaje de predicciones correctas.
- **Precisión-sensibilidad (Recall)** trade-off entre falsos positivos y detección de casos relevantes.
- **F1-score** es la media armónica entre precisión y sensibilidad.
- **Log-loss** mide la calidad de las probabilidades predichas.

Estrategias de validación

Para obtener estimaciones confiables del rendimiento, existen diferentes enfoques:

- **Hold-out** es la división básica de datos de entrenamiento (70-80 %) y prueba (20-30 %). Simple pero puede ser inestable con pocos datos.
- **Validación cruzada k-fold**, divide los datos en k grupos (típicamente 5 o 10). Entrena k veces, usando cada grupo como prueba una vez. Proporciona estimaciones más estables que hold-out.

- **Leave-One-Out (LOOCV)** es la versión extrema de k-fold, donde k es igual al número de muestras n. Usa todas las muestras menos una para el entrenamiento. Precisa pero costosa computacionalmente.
- **Validación cruzada estratificada**, que mantiene la proporción original de clases en cada división. Esencial para datos desbalanceados.

La elección de estos depende del conjunto de datos, del balance entre precisión en la estimación y costo computacional.

3.2 Aprendizaje profundo

El auge del aprendizaje profundo (Deep Learning, DL) responde a una necesidad concreta: superar las limitaciones de los modelos lineales clásicos y problemas donde los datos son complejos y con estructuras no lineales. El DL es un subcampo del ML que se enfoca en el uso de arquitecturas de redes neuronales artificiales con múltiples capas ocultas (ver figura 3.2). La profundidad del modelo, es decir, el número de capas, le permite representar funciones altamente no lineales y de gran capacidad expresiva. Estas arquitecturas son parametrizaciones jerárquicas: las primeras capas suelen aprender representaciones de bajo nivel, por ejemplo, correlaciones locales, mientras que las siguientes capas capturan las relaciones de orden superior relevantes para la tarea objetivo, lo que resulta especialmente útil en el análisis de alta dimensionalidad o gran variedad de datos, como el procesamiento de imágenes, lenguaje natural o series temporales [44].

Un modelo profundo se constituye como una composición de funciones no lineales, cada una representando una capa de la red. Si denotamos por $f_1, f_2, \ldots, f_L$ las funciones que definen cada capa, el modelo total se expresa como: $f(x;\theta) = f_L(f_{L-1}(\ldots f_1(x)))$, donde cada función f_l está parametrizada por un conjunto de θ_l (pesos y sesgos) que son aprendidos durante el proceso de entrenamiento.

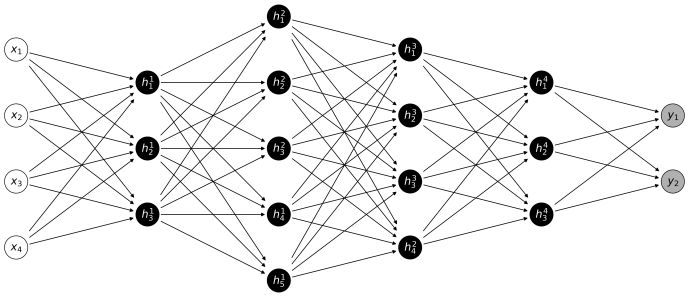

Figura 3.2: Componentes de una red neuronal feedforward con cuatro capas ocultas representadas por los nodos de color negro. La capa de entrada de color blanco y la capa de salida de color gris. Cada flecha representa un peso.

En ML una red neuronal feedforward es una arquitectura donde la información fluye en una sola dirección, desde la capa de entrada hacia la capa de salida, sin bucles o retroalimentación. Una red neuronal de este tipo consiste en:

- Una capa de entrada que recibe el vector de características $(x_1, x_2, \ldots, x_i)$.
- Varias capas ocultas, cada una definida como: $h^l = \sigma(W^l h^{l-1} + b^l)$, donde σ es una función de activación no lineal, por ejemplo ReLU[2], W^l son los pesos y b^l los sesgos [44].
- Una capa de salida $(y_1, y_2, \ldots, y_i)$ adaptada al tipo de problema (clasificación, regresión, series temporales, etc.).

Las redes de este tipo, también conocidas como *multiplayer perceptrons* (MLP), constituyen la base conceptual sobre la cual se construyen modelos más avanzados. Entre las arquitecturas más representativas del aprendizaje profundo se encuentran:

- **Redes neuronales convolucionales (CNN)**, que están diseñadas para explorarla estructura espacial local en datos como imágenes o señales bidimensionales. Estas redes aplican filtros convolucionales, que realizan productos punto locales entre una pequeña matriz de pesos y regiones de la entrada, permitiendo extraer automáticamente patrones como bordes, texturas o formas. Son ampliamente usadas en tareas de clasificación visual, reconocimiento facial o análisis médico por imagen.
- **Redes neuronales recurrentes (RNN)**, introducen conexiones temporales entre unidades, permitiendo que la red mantenga una memoria de entradas anteriores. Esta propiedad las hace especialmente eficaces para modelar secuencias, como texto, audio o datos temporales. Las RNN y una de sus variantes *Long Short-Term Memory* (LSTM), serán abordadas con mayor detalle en la siguiente subsección.
- **Redes generativas profundas**, como las **redes antagónicas (GAN)** o los **autoencoders**, que aprenden distribuciones de probabilidad sobre los datos y son capaces de generar nuevas muestras realistas. Estas redes tienen aplicaciones en síntesis de imágenes, reducción de dimensionalidad y modelado no supervisado.
- **Redes de creencia profunda (DBN)** y modelos basados en **restricted boltzamann machines (RBM)**, que fueron pioneros en el preentrenamiento no supervisado de redes profundas y jugaron un papel importante en el desarrollo temprano del DL.

El aprendizaje profundo destaca por su capacidad para realizar ingeniería de características automática, extrayendo representaciones cada vez más abstractas conforme se avanza en la profundidad. No obstante, estas arquitecturas también presentan desafíos, como la necesidad de grandes cantidades de datos, un alto poder computacional y mayor riesgo de sobreajuste en comparación con modelos más simples.

En este trabajo, donde analizaremos datos con estructura temporal, específicamente en problemas como la estimación de la exposición de los detectores de superficie, los modelos feedforward convencionales no tienen memoria ni capacidad para modelar dependencias temporales. Para resolver esta limitación, se emplean modelos de series de tiempo.

[2]ReLU (Rectified Linear Unit) es una función de activación. Su forma funcional es $g(z) = \max(0, z)$, lo que implica que las neuronas solo se activan si tienen una salida positiva para entradas positivas, mientras que para entradas negativas la salida es cero.

3.2.1 Redes neuronales recurrentes

Los modelos tradicionales suponen independencia entre las entradas, lo cual es una limitación cuando las observaciones actuales dependen de eventos anteriores. Para capturar estas dependencias temporales, se introducen las redes neuronales recurrentes.

Las RNNs poseen conexiones cíclicas (ver figura 3.3) que permiten que la información se retroalimente dentro de la red, proporcionando una memoria interna dinámica [44]. En términos generales, una RNN procesa una secuencia de entradas $(x_1, x_2, \ldots, x_T)$, generando una secuencia de estados ocultos $(h^1, h^2, \ldots, h^T)$ definidos de manera recursiva:

$$h^t = \sigma(U^h x_t + W^h h^{t-1} + b^h), \quad y_t = \sigma(W^y h^t + b^y) \tag{3.2.1}$$

donde σ es una función de activación (por ejemplo, tanh[3] o ReLU), U^h es la matriz de peso que conecta la entrada x_t, W^h y W^y son matrices de pesos aprendibles (recurrentes) que conectan con el estado oculto h_t, b^h y b^y son sesgos y y_t representa la salida en el paso del tiempo t [44]. Este estado oculto actúa como una forma de memoria, permitiendo que la red incorpore contexto histórico al procesar nuevas entradas.

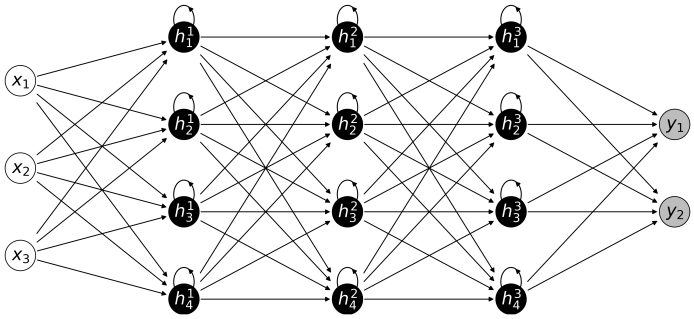

Figura 3.3: Esquema de una red neuronal recurrente con una capa de entrada representados por los nodos color blanco, las tres capas ocultas de color negro y la capa de salida de color gris, además se añade un bucle en cada neurona oculta, ilustrando la idea de estado recurrente dentro de cada capa.

El entrenamiento de una RNN se basa en el principio de retropropagación a través del tiempo [45]. Dado que cada salida depende no solo de la entrada actual sino también de las anteriores, los gradientes deben propagarse hacia atrás en el tiempo, de forma general:

$$\frac{\partial \mathcal{L}}{\partial W} = \sum_{t=1}^{T} \frac{\partial \mathcal{L}}{\partial h^t} \cdot \frac{\partial h^t}{\partial W}, \tag{3.2.2}$$

donde $\mathcal{L}$ es la función de pérdida total a lo largo de la secuencia. Esto introduce un problema práctico: cuando la secuencia es larga, los gradientes pueden tender a cero (desvanecimiento) o crecer exponencialmente (explosión), dificultando el entrenamiento del modelo. Las RNNs básicas son capaces de modelar dependencias temporales cortas, pero su desempeño se degrada notablemente cuando se requiere capturar relaciones a largo

[3]La tangente hiperbólica es una función de activación no lineal que mapea cualquier valor de entrada a un rango entre -1 y 1. Su forma es similar a la función sigmoide, pero centrada en cero.

plazo, debido al problema del desvanecimiento del gradiente [46]. Este fenómeno impide que las actualizaciones de los parámetros reflejen correctamente la influencia de eventos anteriores lejanos en el tiempo, lo que limita la capacidad de la red para aprender patrones de largo alcance.

Una arquitectura particularmente robusta de este marco es LSTM, permitiendo el aprendizaje de dependencias de largo alcance y el modelo Prophet en series temporales.

3.3 Modelos de series de tiempo

El análisis de series temporales se enfoca en modelar y predecir datos ordenados cronológicamente, considerando su estructura secuencial y la posible presencia de tendencias, estacionalidades y autocorrelaciones. En este contexto, se han desarrollado diversos enfoques que permiten capturar dinámicas temporales a diferentes niveles de complejidad. En lo que sigue, hablamos de dos modelos relevantes dentro de este ámbito.

3.3.1 Modelo LSTM

Las redes neuronales de memoria a largo y corto plazo (LSTM) son una variante de las redes neuronales recurrentes diseñadas, como su nombre lo indica, para superar las limitaciones en el aprendizaje de dependencias a largo plazo. Propuestas por Hochreiter y Schmidhuber [47], las LSTM introducen una estructura interna que regula el flujo de información mediante puertas (gates), permitiendo preservar y actualizar información relevante durante largas secuencias.

El estado oculto en una LSTM se acompaña de un **estado de celda** c_t, que funciona como una memoria explícita. La dinámica se controla mediante tres puertas:

Puerta de olvido: decide qué información eliminar del estado anterior,

$$f_t = \sigma(U_f x_t + W_f h_{t-1} + b_f). \tag{3.3.1}$$

Puerta de entrada: controla cuánta información nueva se incorpora,

$$i_t = \sigma(U_i x_t + W_i h_{t-1} + b_i), \tag{3.3.2}$$

$$\tilde{c}_t = \tanh\left(U_c x_t + W_c h_{t-1} + b_c\right). \tag{3.3.3}$$

Puerta de salida: determina qué parte de la memoria se transmite como salida,

$$o_t = \sigma(U_o x_t + W_o h_{t-1} + b_o). \tag{3.3.4}$$

Donde U representa las matrices de pesos que conectan **la entrada actual** x_t, W representa las matrices de pesos que conectan el **estado oculto anterior** h_{t-1} y b son los **vectores de sesgo**. Todas estas variables están asociadas a cada una de las puertas del modelo.

La figura 3.4 muestra el esquema funcional de una celda LSTM. Este modelo incorpora un conjunto de operaciones que regulan el flujo de información mediante las tres puertas (f_t, i_t, o_t). Cada una de estas puertas aplica una activación sigmoide σ, cuyo resultado

controla qué fracción de la información se conserva, actualiza o expulsa del estado de la celda.

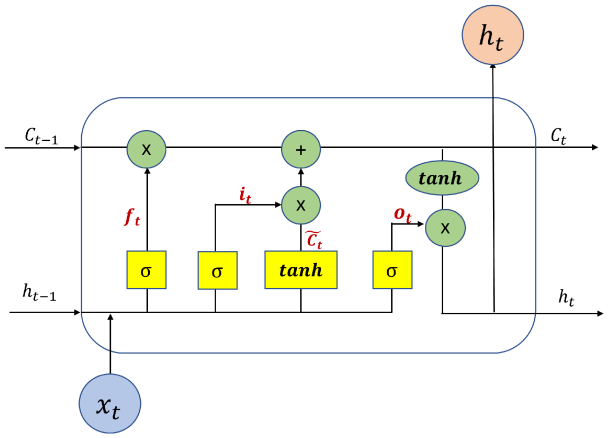

Figura 3.4: Esquema de una celda LSTM. *Imagen tomada de:* [48]

El proceso comienza con la concatenación del estado anterior oculto h_{t-1} y la entrada actual x_t, la cual se introduce a cada una de las puertas. La puerta de olvido determina qué parte del estado de la celda anterior c_{t-1} debe conservarse. Simultáneamente, la puerta de entrada modula la información nueva, combinando i_t y la candidata a nuevo contenido $\tilde{c}_t$, que se genera mediante una activación tangente hiperbólica. La suma ponderada entre estos dos elementos define el **nuevo estado de la celda** c_t:

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t. \tag{3.3.5}$$

Finalmente, la puerta de salida regula qué parte del estado actualizado se utiliza para calcular el **nuevo estado oculto** h_t, aplicando:

$$h_t = o_t \odot \tanh c_t, \tag{3.3.6}$$

donde $\odot$ denota el producto elemento a elemento. Esta arquitectura permite que la celda LSTM mantenga y manipule información relevante a lo largo de largas secuencias temporales. Sin embargo, a pesar de su capacidad para modelar dependencias temporales complejas, el desempeño de las LSTM depende críticamente de la cantidad y calidad de los datos disponibles. Sin una muestra suficientemente representativa y extensa, el modelo no puede aprender patrones generalizables, incurriendo en sobreajuste o convergencia prematura a soluciones poco óptimas.

Además, las LSTM implican un alto coste computacional tanto en memoria como en tiempo de entrenamiento. Aunque el uso de unidades de procesamiento gráfico (GPU) ha sido esencial para acelerar el entrenamiento, el proceso sigue siendo intensivo, llegando a requerir horas de cómputo en equipos convencionales.

3.3.2 Modelo Prophet

Un enfoque alternativo es el uso del modelo Profeta (Prophet), desarrollado por Taylor y Letham en el equipo de investigación de Meta (anteriormente Facebook). Prophet es un

modelo de series temporales desarrollado para realizar pronósticos precisos, interpretables y escalables en contextos empresariales [49]. Su diseño responde a la necesidad de modelar series con tendencias no lineales, múltiples estacionalidades, eventos atípicos y efectos recurrentes o irregulares, sin recurrir a métodos estadísticos complejos.

Desde un punto de vista matemático, Prophet se basa en un modelo aditivo de descomposición[4] de la serie temporal $y(t)$, el cual se expresa como [49]:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t, \tag{3.3.7}$$

donde, $g(t)$ representa la **tendencia a largo plazo**, $s(t)$ captura las **estacionalidades** (diarios, semanales, anuales), modeladas mediante series de Fourier, $h(t)$ incorpora los **eventos especiales** o interrupciones definidos manualmente y ε_t es un **término de error** aleatorio (ruido no estructurado).

Este enfoque lo vincula estrechamente con los modelos aditivos generalizados[5] (GAMs), permitiendo ajustar componentes no lineales.

En el componente de tendencia, Prophet ofrece dos formulaciones: una forma **logística** que permite crecimiento saturado, y una forma **lineal por tramos** adecuada para tasas de crecimiento constante. Ambas se complementan con los puntos de cambio (changepoints) que permiten capturar transiciones en la dinámica temporal.

Tendencia logística

Cuando se modela un crecimiento que tiende a una capacidad límite, Prophet puede adoptar una forma logística generalizada. Descrita por [49]:

$$g(t) = \frac{C}{1 + \exp\left(-k(t - m)\right)}, \tag{3.3.8}$$

donde C es la capacidad máxima (especificada por el usuario), k representa la tasa de crecimiento (pendiente inicial) y m el desplazamiento temporal.

Para incorporar cambios en la tasa de crecimiento a lo largo del tiempo, Prophet extiende este modelo mediante una suma de indicadores de cambio activados en puntos específicos, lo que da lugar a la siguiente expresión [49]:

$$g(t) = \frac{C(t)}{1 + \exp\left(-(k + a(t)^\top \delta)(t - (m + a(t)^\top \gamma))\right)}, \tag{3.3.9}$$

donde, δ representa los cambios en la tasa de crecimiento en cada uno de los changepoints, $a(t)$ es un vector indicativo que activa los cambios en la pendiente una vez que el tiempo t ha sobrepasado el punto de cambio correspondiente y γ asegura la continuidad de la función $g(t)$ al momento del cambio de pendiente. Esta forma permite al modelo adaptarse a comportamientos logísticos por tramos.

[4] Asumen que la serie temporal puede representarse como la suma de sus componentes individuales, como tendencia, estacionalidad y ruido.

[5] Son modelos de regresión flexibles donde la variable respuesta se modela como una suma aditiva de funciones suaves de los predictores, sin asumir formas funcionales específicas.

Tendencia lineal por tramos

En escenarios donde el crecimiento no muestra saturación, una alternativa más simple es el modelo de crecimiento lineal por segmentos. En este caso, la tendencia se modela como [49]:

$$g(t) = (k + a(t)^\top \delta)t + (m + a(t)^\top \gamma), \tag{3.3.10}$$

los términos conservan el mismo significado que en el caso logístico. Aquí la función $g(t)$ representa una línea continua con pendientes ajustables en los puntos de cambio e igual que en la tendencia logística, la función $a(t)^\top \delta$ modifica la pendiente, y $a(t)^\top \gamma$ ajusta el intercepto para mantener la continuidad.

En ambas formulaciones, los parámetros k, m, δ y γ son estimados automáticamente a partir de los datos mediante procesos de optimización durante el entrenamiento. El usuario únicamente especifica la forma general de la tendencia `growth="linear"` o `growth="logistic"` y, en el caso logístico, la capacidad máxima mediante el parámetro `cap`.

Prophet no requiere que los datos estén equiespaciados ni imputar valores faltantes, lo que lo hace particularmente flexible ante datos reales. Esta automatización contribuye a la facilidad de uso del modelo.

Capítulo 4

Análisis del conjunto de datos

La caracterización del espectro de energía de los rayos cósmicos requiere datos experimentales con un alto nivel de fidelidad reconstructiva. Para ello, es necesaria una gestión técnica del conjunto de datos, tanto en su adquisición como en su procesamiento. Este capítulo presenta una descripción del análisis de datos utilizados en esta investigación, así como los procedimientos seguidos para el diseño de los modelos de predicción. Los datos provienen del detector de superficie del Observatorio Pierre Auger; estos se almacenan originalmente en un sistema jerárquico por fecha y son exportados y consolidados en archivos binarios `.root`[1] dentro del CDAS.

Los eventos analizados abarcan el período comprendido entre el 1 de enero de 2004 y el 14 de abril de 2023 para los eventos detectados por el SD-1500, con más de 8 millones de eventos registrados, mientras que los eventos del SD-750 abarcan del 14 de septiembre de 2007 al 14 de abril de 2023, con más de 5 millones de eventos. Originalmente, la información se encuentra distribuida en miles de archivos individuales organizados jerárquicamente por fecha de adquisición. Para optimizar el procesamiento y análisis, se implementó un pipeline que une todos los archivos individuales en un solo archivo `.root`.

4.1 Cortes de calidad

La depuración inicial del conjunto de datos se llevó a cabo mediante la ejecución de un script desarrollado en el entorno Linux que permitió reducir drásticamente los tiempos de procesamiento, pasando de un estimado de varias semanas en equipos convencionales a tan solo 21 horas de cómputo continuo. Como resultado, se obtuvo un conjunto de datos compuesto por más de 5 millones de eventos registrados por el SD-1500 y más de 1 millón de eventos por el SD-750.

Los cortes de calidad aplicados en esta primera etapa tienen como objetivo garantizar la confiabilidad de los eventos considerados para el análisis posterior de la exposición y del espectro energético de los RC. Estos criterios se resumen en el cuadro 4.1 y se describen a continuación:

[1] Formato desarrollado por el CERN para almacenamiento eficiente de grandes volúmenes de datos, que permite estructura jerárquica.

Cuadro 4.1: Resumen de los cortes al conjunto de datos del SD-1500 y SD-750.

Cortes
lightning
min RecLevel
max Zenith $\theta < 60°$
T4 Trigger
T5 Trigger
T5 TriggerUB
min $\log_{10}(E/\text{eV}) > 10^{17.5}$
bad PeriodsRejection
min $\log_{10}(E/\text{eV}) > 10^{17}$

- **lightning**, se descartan eventos cuya señal registrada en los detectores presenta características eléctricas anómalas asociadas a descargas atmosféricas.

- **min RecLevel**, se impuso un umbral mínimo al nivel de reconstrucción del evento para asegurar que solo se incluyeran eventos con reconstrucción geométrica y energética confiable.

- **max Zenith**, se seleccionaron únicamente eventos con un ángulo cenital $\theta < 60°$, correspondiente al rango vertical definido por la Colaboración Pierre Auger para garantizar la eficiencia plena del arreglo SD.

- **T4 y T5 Triggers**, se exigió que todos los eventos cumplieran con los criterios de disparo T4 y T5.

- **T5 TriggerUB**, variante extendida del T5 que considera una geometría más permisiva.

- **min** $\log_{10}(E/\text{eV}) > 10^{17.5}$, se aplico un umbral inferior de energía con el fin de restringir el análisis a eventos por encima del umbral de eficiencia total del SD-1500.

- **bad PeriodsRejection**, se eliminaron eventos registrados durante intervalos de tiempo con anomalías operativas detectadas en el monitoreo del estado del detector, por ejemplo, fallas de estaciones o mantenimiento del sistema.

- **min** $\log_{10}(E/\text{eV}) > 10^{17}$, adicionalmente, un corte más bajo de energía aplicado exclusivamente al conjunto de datos de la matriz SD-750 por encima del umbral de eficiencia donde el arreglo alcanza su máximo rendimiento.

Una vez consolidados los dos conjuntos de datos correspondientes al SD-1500 y SD-750, y tras la aplicación de los cortes iniciales, la lectura de los archivos se realizó por separado mediante la biblioteca `uproot` (ver código 4.1), que permite acceder a los datos contenidos en el archivo `.root` desde Python, sin necesidad de un entorno C++. Los archivos `.root` son estructuras de almacenamiento desarrolladas en el marco del software ROOT,

ampliamente utilizado en física de altas energías para manejar grandes volúmenes de datos jerárquicos. Estos archivos contienen objetos organizados en forma de árboles, donde cada `árbol` (`TTree`) puede contener múltiples `ramas` (`branches`). Cada rama almacena una variable o conjunto de variables asociadas a los eventos registrados, lo que permite acceder de forma eficiente a subconjuntos específicos de la información.

Dado que el procedimiento aplicado a ambos conjuntos es idéntico, a continuación se describe el flujo de trabajo correspondiente a uno de ellos.

Código 4.1: Lectura del archivo cosolidado.

```
import uproot
import pandas as pd

archivo = uproot.open("archivo.root")
archivo.keys()  # Visualizacion de los arboles disponibles
archivos = archivo["recData"]  # Acceso al arbol principal

# Obtencion y listado de todas las ramas del arbol recData
ramas = archivo["recData"].keys()
print("Todas las ramas:")
for rama in ramas:
    print(rama)
```

Tras explorar las ramas disponibles del árbol `recData`, que contiene los datos reconstruidos por el detector, se aplicó un filtro para conservar únicamente aquellas asociadas al SD (`event.fSDEvent`). Posteriormente, se seleccionaron de forma manual las variables relevantes para el análisis y se exportó a un archivo en formato `.csv`, con el objetivo de evitar un conjunto de datos innecesariamente pesado y más amigable de tratar. Este procedimiento se muestra en el código 4.2.

Código 4.2: Extracción de ramas y organización en un DataFrame.

```
EventosSD = {
    "EventId":pd.DataFrame(archivos["event./event.fSDEvent/event.
       fSDEvent.fSdEventId"].array(library="pd")),
    # ... otras ramas seleccionadas
    "NoOfCandidateStations":pd.DataFrame(archivos["event./event.fSDEvent
        event.fSDEvent.fNoOfCandidateStations"].array(library="pd")),
    "AxisCoreCS":pd.DataFrame(archivos["event./event.fSDEvent/event.
       fSDEvent.fSdRecShower.RecShower/event.fSDEvent.fSdRecShower.
       Shower/event.fSDEvent.fSdRecShower.fAxisCoreCS"].array(library="
       pd")),
    }

dfSD = pd.concat(EventosSD, axis=1)
dfSD.to_csv("datosSD.csv", index=False)
```

Por último, se realizó un filtrado final basado en la variable `NoOfCandidateStations` que representa las estaciones activadas por evento, con el propósito de verificar el cumplimiento estricto del criterio fiducial. Asimismo, se eliminaron eventos con valores atípicos o inconsistentes en los parámetros operativos. Como resultado, se obtuvo un conjunto depurado compuesto por 187,843 eventos correspondientes al arreglo SD-1500 y 279,835 eventos

del arreglo SD-750, ambos considerados físicamente confiables. Estos eventos constituyen la base sobre la cual se desarrolla esta tesis.

4.2 Descripción de los datos

Los datos corresponden a casi dos décadas de observación continua por el SD-1500 y más de 15 años por el SD-750. El proceso de exploración y selección de variables está enfocado en el análisis de la exposición del detector, parámetro fundamental para la determinación del flujo de rayos cósmicos y el posterior análisis del espectro energético. El conjunto final incluye variables asociadas a los identificadores del evento, las características de las estaciones activas, así como las propiedades geométricas y energéticas reconstruidas. Entre las que se encuentran:

- Identificadores de evento: `EventId`.
- Identificadores de tiempo: `YYMMDD`, `HHMMSS`, `GPSSecond`, `GPSNanoSecond`.
- Reconstrucción de energía: `Energy`, `EnergyError`.
- Condiciones de calidad: `NoOfCandidateStation`, `NumberOfAccidentalStation`.
- Dirección de llegada reconstruida: `AxisCoreCS`.

Los identificadores de evento y tiempo están sincronizados mediante el sistema GPS, alcanzando precisiones del orden de nanosegundos. En particular, la variable `YYMMDD` codifica la fecha en formato año-mes-día, mientras que `HHMMSS` representa la hora en formato hora-minutos-segundos. Las variables `GPSSecond` y `GPSNanoSecond` permiten una referencia temporal absoluta.

La variable `Energy` mide con mucha precisión la energía reconstruida de los eventos detectados. Por su parte, `NoOfCandidateStation`, `NumberOfAccidentalStation` cuantifican respectivamente el número de estaciones que participaron activamente en la reconstrucción del evento y aquellas que se encontraban activadas de forma accidental o sin cumplir criterios de calidad.

Un aspecto relevante fue la reconstrucción de los ángulos cenital (θ) y azimutal (ϕ), necesarios para caracterizar la dirección de llegada de los eventos. Dado que estos ángulos no se encontraban explícitamente en los datos brutos, se dedujeron a partir de la rama `AxisCoreCS`, que representa el vector unitario de dirección del eje del chubasco en coordenadas cartesianas (fX, fY, fZ) en el sistema de referencia del detector. A partir de estas componentes, los ángulos se obtuvieron de la siguiente manera:

$$\theta = \arccos(fZ), \tag{4.2.1}$$

$$\phi = \arctan 2(fY, fX). \tag{4.2.2}$$

Ambos se expresan en radianes y posteriormente se convierten en grados para su interpretación física (ver código 4.2).

Código 4.3: Cálculo de los ángulos de llegada.

```python
import numpy as np

# Extraer la rama de interes
axis_core_data = filtrado["fAxisCoreCS"]

# Convertir en DataFrame estructurado
df = pd.DataFrame(axis_core_data.tolist(),
                  columns=["fX", "fY", "fZ"])

# Calcular cenit y azimut
df["theta"] = np.degrees(np.arccos(df["fZ"]))
df["phi"] = np.degrees(np.arctan2(df["fY"], df["fX"]))

# Ajustar el rango de phi
df["phi"] = df["phi"] % 360
```

La validación de este procedimiento se realizó mediante comparación directa con los valores mostrados por el software *Event Browser*[2], obteniendo una concordancia exacta para todos los eventos visualizados. Esta verificación confirma la validez del método de cálculo empleado, así como la consistencia interna del sistema de coordenadas del detector.

Verificada esta etapa, se analizó la distribución temporal de los eventos depurados en función de la energía reconstruida para ambos arreglos (véase figura 4.1). Cada punto representa un evento individual, con la energía expresada en electronvoltios y graficada en escala lineal.

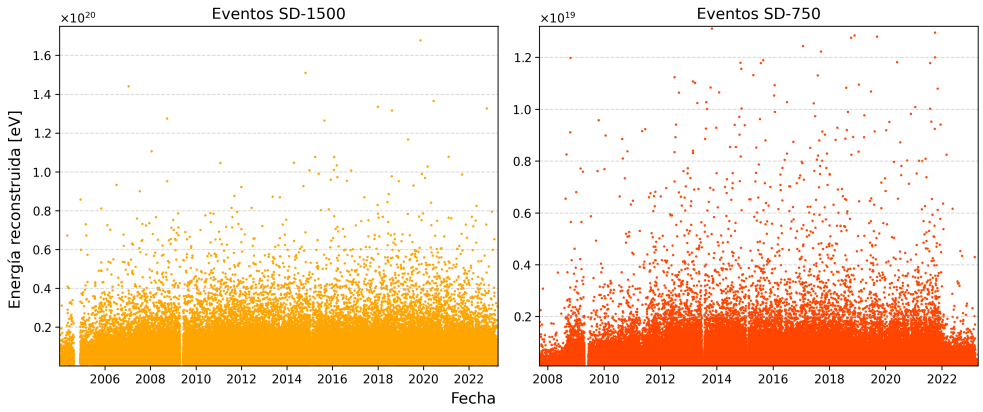

Figura 4.1: Distribución temporal en función de la energía de los eventos reconstruidos del arreglo SD-1500 y el SD-750.

En el gráfico izquierdo se observa una alta densidad de eventos con energías superiores a 10^{18} eV, así como la presencia de eventos extremos por encima de 10^{20} eV. Además, se identifican dos lagunas temporales de datos, la primera en el intervalo del año 2004 y la segunda en el año 2009, atribuibles a problemas técnicos en el sistema de comunicaciones del arreglo, según ha sido documentado en el artículo [9] de la Colaboración Pierre Auger.

[2]Herramienta de visualización y análisis de eventos de RC desarrollado por la colaboración Pierre Auger basada en ROOT.

Discontinuidades similares, aunque en menor magnitud, son visibles alrededor de los años 2015 y 2020.

La gráfica de la derecha, como era esperable, la mayoría de los eventos se concentran en el intervalo energético de 10^{17}-10^{18} eV, para el cual este arreglo fue optimizado. También en este conjunto se observa una ligera interrupción en el año 2009, así como una pequeña reducción de eventos alrededor de 2013, 2015, 2016 y 2020.

Estas observaciones permiten confirmar que, salvo por las interrupciones documentadas técnicas, el conjunto de datos conserva tanto la continuidad como la representatividad estadística en los rangos energéticos de interés. Esta estabilidad es fundamental para el análisis de la exposición acumulada.

4.3 Cálculo de la exposición

Para estimar el flujo diferencial de rayos cósmicos a partir del número de eventos detectados por el SD, es indispensable conocer la **exposición acumulada** del detector, expresada en unidades de $\mathrm{km}^2 \cdot \mathrm{sr} \cdot \text{año}$, que determina la capacidad del arreglo para detectar partículas.

La exposición se calcula mediante la integración en el tiempo del número de **hexágonos elementales activos** donde cada hexágono consta de seis estaciones activas, multiplicado por la apertura proyectada de cada celda, definida como el área efectiva en el plano perpendicular a la dirección de llegada del chubasco. El área de un hexágono elemental se calcula como:

$$A_{\mathrm{cell}} = \frac{\sqrt{3}}{2} d^2. \tag{4.3.1}$$

Las áreas correspondientes a cada arreglo (SD-750 y SD-1500) son:

$$A_{\mathrm{cell}}^{750} = \frac{\sqrt{3}}{2}(0.75\ \mathrm{km})^2 = 0.487\ \mathrm{km}^2,$$

$$A_{\mathrm{cell}}^{1500} = \frac{\sqrt{3}}{2}(1.5\ \mathrm{km})^2 = 1.949\ \mathrm{km}^2.$$

Cada evento detectado se caracteriza por su dirección de llegada, determinada mediante los ángulos cenital θ y azimutal ϕ. La **apertura o área efectiva instantánea** asociada a los hexágonos activos en un instante t se estima utilizando la siguiente expresión:

$$\mathrm{d}A_{\mathrm{6T5}} = A_{\mathrm{cell}} \cos\theta\ \mathrm{d}\Omega = -A_{\mathrm{cell}} \cos\theta\ \mathrm{d}(\cos\theta)\ \mathrm{d}\phi, \tag{4.3.2}$$

donde $\cos\theta$ es un factor de proyección que ajusta el área efectiva según la inclinación del evento, $\mathrm{d}\Omega$ es el elemento de ángulo sólido en coordenadas esféricas, y la relación $\mathrm{d}\Omega = -\mathrm{d}(\cos\theta)\ \mathrm{d}\phi$ permite reescribir la expresión en términos de la variable azimutal.

La integración sobre el ángulo sólido para eventos verticales $\theta \in [0°, 60°]$ se realiza de la siguiente manera:

$$A_{\mathrm{6T5}} = A_{\mathrm{cell}} \int_0^{2\pi} \mathrm{d}\phi \int_{\cos\theta_{\mathrm{max}}}^{\cos\theta_{\mathrm{min}}} \mathrm{d}\cos\theta \cos\theta \tag{4.3.3}$$

$$A_{\mathrm{6T5}} = A_{\mathrm{cell}} \pi (\sin\theta_{\mathrm{max}} - \sin^2\theta_{\mathrm{min}}),$$

de esta forma, para SD-750 y SD-1500 se obtiene:

$$A_{6\mathrm{T5}}^{750} = 1.0269 \text{ km}^2 \cdot \text{sr},$$

$$A_{6\mathrm{T5}}^{1500} = 4.5912 \text{ km}^2 \cdot \text{sr}.$$

Con esta apertura por celda, la exposición total ε se calcula integrando en el tiempo la apertura instantánea del sistema:

$$\varepsilon = \int_{t_0}^{t_f} N_{\mathrm{act}}(t) \cdot A_{6\mathrm{T5}} \, \mathrm{d}t, \tag{4.3.4}$$

donde $N_{\mathrm{act}}(t)$ representa el número de hexágonos activos en el instante t. En la práctica, esta integral se aproxima mediante una suma discreta sobre los instantes registrados:

$$\varepsilon = \sum_i N_{\mathrm{act}}(t_i) \cdot A_{6\mathrm{T5}}. \tag{4.3.5}$$

4.3.1 La exposición acumulada a partir de los datos

La estimación de la exposición acumulada se realizó de la forma discreta y diferenciada para los arreglos SD-1500 y SD-750, mediante un enfoque evento por evento. Esta metodología se adoptó ante la ausencia de información sobre el número de hexágonos activos en cada instante de observación.

Se obtuvo el instante temporal en formato de calendario a partir de la conversión del tiempo GPS almacenado en la variable `GPSSecond` tal y como se muestra en el código 4.4.

Código 4.4: Función para conversión de GPS a datetime.

```
def gps_to_datetime(gps_time):
    gps_epoch = dt.datetime(1980, 1, 6)
    return gps_epoch + dt.timedelta(seconds=gps_time)
```

Además, se calculó la apertura efectiva considerando el número de estaciones activadas por evento. En estudios más rigurosos, la exposición suele corregirse mediante factores que consideran condiciones atmosféricas, proyecciones geométricas, eficiencia instrumental y estabilidad operativa. En este trabajo, sin embargo, se adopta una forma general de cálculo, aplicando únicamente una función de eficiencia angular y energética a los datos del SD-1500 (ver código 4.5), dado que no opera con eficiencia plena ($E < 10^{18.5}$ eV).

Código 4.5: Función de eficiencia.

```
from scipy.special import erf

def eficiencia(E, theta, p1=0.373):
    cos_theta = np.cos(theta)
    p0_theta = 18.63 - 3.18 * cos_theta**2 + 4.38 * cos_theta**4 - 1.87
        * cos_theta**6
    return 0.5 * (1 + erf((np.log10(E) - p0_theta) / p1))
```

La función de eficiencia $\epsilon(E, \theta)$ empleada reproduce la parametrización propuesta por la colaboración Pierre Auger [50], basada en estudios con eventos híbridos. Esta función

depende de la energía reconstruida del evento y de su ángulo cenital, y está dada por:

$$\epsilon(E,\theta) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\log_{10}(E/\mathrm{eV}) - p_0(\theta)}{p_1}\right)\right], \tag{4.3.6}$$

donde $p_1 = 0.373$ controla la pendiente de transición, erf es la función de error y el término $p_0(\theta)$ incorpora la dependencia angular, definido como:

$$p_0(\theta) = 18.63 - 3.18\cos^2\theta + 4.38\cos^4\theta - 1.87\cos^6\theta.$$

Esta formulación permite modelar una transición sigmoidal desde eficiencia baja a plena.

Para el SD-750, se asumió eficiencia unitaria debido a que los eventos analizados se encuentran en el régimen de eficiencia plena ($E > 10^{17}$ eV). Sin embargo, esta aproximación sobreestimaba significativamente la exposición. Esto se debe a que no se cuenta con información detallada sobre la actividad temporal de los hexágonos, ni la fracción del arreglo efectivamente operativo en cada instante.

Ante esta carencia, se adoptó una aproximación intermedia: el cálculo se mantuvo a nivel de evento por evento, pero se incorporó un factor de normalización de 0.1 sobre la apertura efectiva estimada, con el objetivo de aproximar el valor total al rango de la exposición acumulada reportado para este arreglo [51]. Esta corrección refleja de manera conservadora la fracción promedio de celdas operativas no observables en los datos, y permite estimar la exposición sin introducir supuestos.

El cálculo de la exposición acumulada fue implementado mediante el código 4.6, que estima la apertura efectiva por evento e integra la exposición acumulada diferenciando entre los arreglos SD-1500 y SD-750:

Código 4.6: Cálculo de la exposición acumulada para SD-1500 y SD-750.

```
# Calculo de exposicion por evento
def calcular_exposicion(df, arreglo="1500"):
    df = df.copy()
    df["datetime"] = df["GPSSecond"].apply(gps_to_datetime)
    df["year"] = df["datetime"].dt.year + (df["datetime"].dt.dayofyear /
        365.25)
    theta_rad = np.radians(df["theta"])

    # Area por celda segun arreglo
    if arreglo == "1500":
        A_hex = 1.95
        eff = eficiencia(df["Energy"], theta_rad)
    elif arreglo == "750":
        A_hex = 0.49
        eff = 1.0  # Eficiencia plena
    else:
        raise ValueError("Arreglo no reconocido")

    # Apertura efectiva base
    A_t = df["NoOfCandidateStations"] * A_hex * eff

    if arreglo == "750":
        A_t *= 0.1  # Aplicar factor de normalizacion
```

```
23
24      df["A_t"] = A_t
25      df["Exposure_cumulative"] = A_t.cumsum()
26      return df[["datetime", "year", "A_t", "Exposure_cumulative"]]
27
28  # Calculo de exposicion acumulada
29  def exposicion_en_anios(df_expo):
30      df_expo = df_expo.sort_values("datetime")
31      t0 = df_expo["datetime"].min()
32      tf = df_expo["datetime"].max()
33      delta_t_years = (tf - t0).total_seconds() / (365.25 * 24 * 3600)
34      exposicion_total_km2sr = df_expo["A_t"].sum()
35      exposicion_km2sryr = exposicion_total_km2sr / delta_t_years
36      return exposicion_km2sryr, delta_t_years
37
38  # Calcular exposicion
39  expo_1500 = calcular_exposicion(df_sd1500, arreglo="1500")
40  expo_750 = calcular_exposicion(df_sd750, arreglo="750")
41
42  # Calcular exposicion acumulada
43  expo_total_1500 = exposicion_en_anios(expo_1500)
44  expo_total_750 = exposicion_en_anios(expo_750)
```

Con base en lo anterior, el cuadro 4.2 resume los principales resultados generales obtenidos para ambos arreglos.

Cuadro 4.2: Resumen general de exposición acumulada por arreglo.

Arreglo	**Tiempo de operación**	**Eventos**	**Exposición [km^2 · sr · año]**
SD-1500	01/01/2004 - 14/04/2023	187,843	$\varepsilon_{1500} = 89,777$
SD-750	14/09/2007 - 14/04/2023	279,835	$\varepsilon_{750} = 772.7$

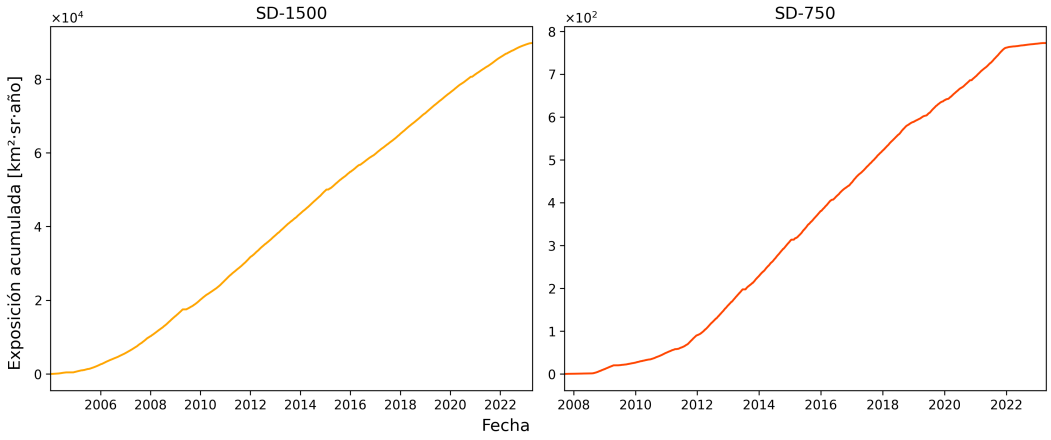

Figura 4.2: Evolución de la exposición de los conjuntos de datos del SD-1500 y SD-750.

De este modo, la figura 4.2 muestra la evolución temporal de la exposición acumulada para ambos arreglos. El conjunto SD-1500 presenta un crecimiento continuo y sostenido

desde el inicio del período de observación hasta la fecha de corte del conjunto de datos, aproximándose a la estimación de [50]. Por su parte, en el arreglo SD-750 se observa una tendencia ascendente similar, aunque con un ritmo más moderado en los primeros años, debido al cambio de la electrónica de lectura, tal y como se reporta en [52]. La diferencia en las pendientes acumuladas entre ambos sistemas está asociada a su tamaño relativo.

El comportamiento regular y sostenido en la evolución de la exposición acumulada sugiere que es posible construir modelos predictivos que permitan extrapolar dicha magnitud hacia intervalos de tiempo no observados.

4.4 Implementación de los modelos

Una de las motivaciones principales de este trabajo es explorar modelos de aprendizaje automático para predecir la exposición, en lugar de simulaciones tradicionales como CONEX o CORSIKA. Aunque estas simulaciones basadas en Monte Carlo son fundamentales en la física de rayos cósmicos, su uso prolongado se ve limitado por el alto costo computacional, las complejas parametrizaciones y la acumulación de incertidumbres. En contraste, los modelos de series de tiempo permiten aprovechar directamente datos reales del experimento, facilitando proyecciones rápidas, ajustables y con menor carga computacional.

En este trabajo se implementan modelos LSTM y Prophet para estimar el crecimiento de la exposición acumulada de los arreglos SD-1500 y SD-750, utilizando los recursos computacionales del Laboratorio Nacional de Supercómputo del Sureste de México (LNS). A continuación, se describen sus características y el proceso de implementación.

4.4.1 Implementación del modelo LSTM

La elección del modelo se fundamenta en su capacidad teórica para capturar relaciones temporales, especialmente en series de tiempo no estacionarias.

Figura 4.3: Incremento mensual de la exposición acumulada.

No obstante, diversos estudios, incluido el de [53] advierten sobre las limitaciones del modelo LSTM. Específicamente, se señala que su rendimiento se degrada cuando se en-

frenta a series con crecimiento suave, irregularidades no periódicas o con escasa cantidad de datos.

Con el objetivo de superar estas limitaciones, se propuso entrenar el modelo LSTM sobre sus incrementos mensuales. Como se ilustra en la figura 4.3, esta transformación resalta la variabilidad mensual de la exposición, revelando comportamientos más dinámicos durante el período de observación de los detectores y fases de estabilidad operacional. La hipótesis planteada fue que el LSTM podría beneficiarse de esta estructura temporal al enfocarse en las tasas de cambio, en lugar de la acumulación.

Construcción de la serie temporal y preparación del conjunto de datos

Para construir la serie temporal utilizada como variable objetivo del modelo LSTM, se calcularon los incrementos mensuales de exposición de los arreglos. Este procedimiento es el mismo para SD-1500 y SD-750, por lo que seguimos la lógica del código implementado al SD-750. Este método consistió en ordenar cronológicamente los eventos registrados y aplicar una derivada discreta a la columna `Exposure_cumulative` del código 4.6, generando así la variable `delta_exposure`, como se muestra en el código 4.7. Posteriormente, los datos se agruparon por mes calendario utilizando el tipo de periodo mensual de `pandas`, y se sumaron los incrementos individuales para cada mes contenidos en la variable `exposure_monthly`.

Código 4.7: Cálculo de los incrementos mensuales.

```
# Calcular los incrementos mensuales de exposicion
expo_750_sorted = expo_750.sort_values("datetime").reset_index(drop=True
    )
expo_750_sorted["delta_exposure"] = expo_750_sorted["Exposure_cumulative
    "].diff()
expo_750_sorted = expo_750_sorted.dropna(subset=["delta_exposure"]).
    reset_index(drop=True)

# Agrupar por mes y sumar los incrementos de exposicion mensual
expo_750_sorted["month"] = expo_750_sorted["datetime"].dt.to_period("M")
monthly_exposure = expo_750_sorted.groupby("month")["delta_exposure"].
    sum().reset_index()
monthly_exposure["month"] = monthly_exposure["month"].dt.to_timestamp()
monthly_exposure.rename(columns={"delta_exposure": "exposure_monthly"},
    inplace=True)
```

Para el desarrollo del modelo se utilizaron las bibliotecas que se muestran en el código 4.8. `PyTorch` fue elegida por su compatibilidad con el entorno de GPU del LNS. Las métricas de evaluación consideradas fueron MAE, MSE, RMSE y R^2.

Código 4.8: Librerías.

```
import torch
import torch.nn as nn
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import mean_absolute_error, mean_squared_error,
    r2_score
from torch.utils.data import DataLoader, TensorDataset
```

La serie temporal de exposición mensual fue normalizada mediante `MinMaxScaler` para mejorar la estabilidad numérica durante el entrenamiento. Posteriormente, se construyeron secuencias de longitud fija `look_back=12` a partir de la serie normalizada, permitiendo al modelo LSTM aprender dependencias temporales mensuales. Finalmente, el conjunto de datos fue dividido en entrenamiento (80 %) y validación (20 %), utilizando cargadores de datos con mini-lotes para una optimización eficiente.

Código 4.9: Preparación y división de datos.

```
# Preparar la serie temporal
serie = monthly_exposure["exposure_monthly"].values.reshape(-1, 1)
scaler = MinMaxScaler()
serie_scaled = scaler.fit_transform(serie)

def create_sequences(data, look_back=12):
    X, y = [], []
    for i in range(len(data) - look_back):
        X.append(data[i:i+look_back])
        y.append(data[i+look_back])
    return np.array(X), np.array(y)

look_back = 12
X, y = create_sequences(serie_scaled, look_back)
X_tensor = torch.tensor(X, dtype=torch.float32)
y_tensor = torch.tensor(y, dtype=torch.float32)

# Division en entrenamiento y validacion
split_idx = int(len(X_tensor) * 0.8)
X_train, X_val = X_tensor[:split_idx], X_tensor[split_idx:]
y_train, y_val = y_tensor[:split_idx], y_tensor[split_idx:]

train_loader = DataLoader(TensorDataset(X_train, y_train), batch_size=8,
    shuffle=True)
val_loader = DataLoader(TensorDataset(X_val, y_val), batch_size=8,
   shuffle=False)
```

Implementación y optimización del modelo

El modelo LSTM implementado es simple para evitar sobreajuste, con dos capas de LSTM `num_layers=2`, `hiden_size=32` unidades ocultas por capa, además de una capa densa final con activación `Softplus` que garantiza que las predicciones sean positivas y el modelo fue optimizado mediante el algoritmo `Adam`:

Código 4.10: Definición del modelo, función de perdida y optimización.

```
class LSTMModel(nn.Module):
    def __init__(self, input_size=1, hidden_size=32, num_layers=2):
        super().__init__()
        self.lstm = nn.LSTM(input_size, hidden_size, num_layers=
           num_layers, batch_first=True)
        self.fc = nn.Sequential(
            nn.Linear(hidden_size, 1),
            nn.Softplus()
```

```
        )

    def forward(self, x):
        out, _ = self.lstm(x)
        return self.fc(out[:, -1, :])

model = LSTMModel()
loss_fn = nn.MSELoss()
optimizer = torch.optim.Adam(model.parameters(), lr=0.001)
```

Este diseño cumple con varios criterios recomendados en la literatura. No se incorporan técnicas adicionales de regularización como penalización `L2` o `dropout` debido a la naturaleza relativamente pequeña del conjunto de datos y a la baja complejidad del modelo; como se discutirá, su efectividad se ve limitada por la estructura de los datos. En este mismo sentido, se utilizó la función `Softplus` en la capa de salida con el objetivo de restringir las predicciones del modelo a valores positivos. Esta elección se fundamenta en el hecho de que, desde el punto de vista físico, la exposición no puede ser negativa, incluso en casos donde existan caídas operativas del arreglo.

El modelo fue entrenado durante 200 épocas utilizando mini-lotes con optimización por descenso del gradiente estocástico. En cada época, se actualizaron los pesos a partir del MSE calculado sobre el conjunto de entrenamiento. Al finalizar cada iteración, se evaluó el desempeño del modelo sobre el conjunto de validación, desnormalizando previamente las predicciones para obtener el error en unidades físicas:

Código 4.11: Proceso de entrenamiento y validación del modelo.

```
# Entrenamiento
epochs = 200
for epoch in range(epochs):
    model.train()
    for X_batch, y_batch in train_loader:
        output = model(X_batch)
        loss = loss_fn(output, y_batch)
        optimizer.zero_grad()
        loss.backward()
        optimizer.step()

    # Validacion por epoca
    model.eval()
    with torch.no_grad():
        y_val_pred = model(X_val).squeeze().numpy()
        y_val_true = y_val.squeeze().numpy()
        y_val_pred_inv = scaler.inverse_transform(y_val_pred.reshape(-1,
            1))
        y_val_true_inv = scaler.inverse_transform(y_val_true.reshape(-1,
            1))
        val_loss = np.mean((y_val_pred_inv - y_val_true_inv) ** 2)

    if epoch % 20 == 0:
        print(f"Epoch {epoch} - Val MSE: {val_loss:.6f}")
```

Evaluación

Al finalizar el entrenamiento, se evaluó el desempeño del modelo sobre el conjunto de validación mediante métricas estándar de regresión (código 4.12). Estas métricas permiten cuantificar la capacidad explicativa del modelo respecto a los datos reales. Las predicciones fueron previamente desnormalizadas, como se mencionó anteriormente.

Código 4.12: Cálculo de las métricas de desempeño sobre el conjunto de validación.

```
# Metricas de desempenio
mae = mean_absolute_error(y_val_true_inv, y_val_pred_inv)
mse = mean_squared_error(y_val_true_inv, y_val_pred_inv)
rmse = np.sqrt(mse)
r2 = r2_score(y_val_true_inv, y_val_pred_inv)

print("\n=== Metricas de validacion ===")
print(f"MAE  = {mae:.4f} [km^{2} sr year]")
print(f"MSE  = {mse:.4f}")
print(f"RMSE = {rmse:.4f}")
print(f"R^{2}   = {r2:.4f}")
```

Resultados

Los resultados obtenidos se resumen en la tabla 4.3. Se observa que el modelo logra capturar la tendencia general de crecimiento en ambos arreglos, aunque con mejor ajuste en el caso del SD-750, donde el valor de R^2 indica una mayor capacidad explicativa. El arreglo SD-1500, si bien la estructura general de la curva es recuperada, el modelo tiende a suavizar los descensos abruptos, lo cual se refleja en errores promedio más elevados.

Cuadro 4.3: Métricas de validación del modelo LSTM para SD-750 y SD-1500.

Arreglo	**MAE**	**MSE**	**RMSE**	R^2
SD-750	17.02	434.77	20.85	0.698
SD-1500	1019.17	1,714,833.13	1309.52	0.579

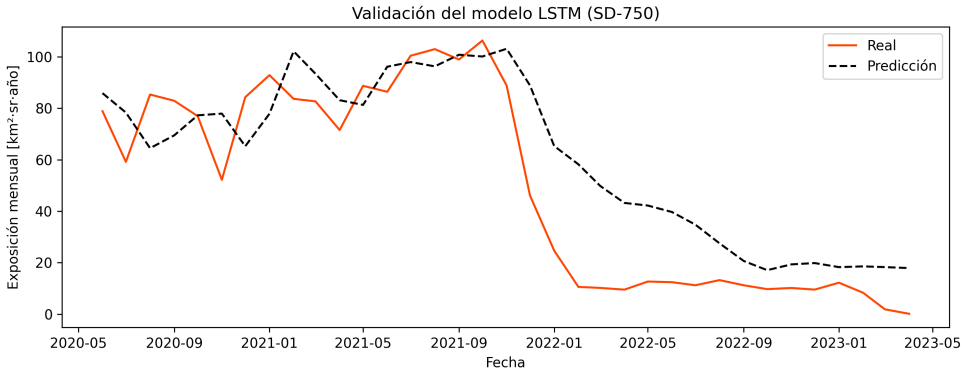

Figura 4.4: Comparación entre la exposición mensual real y la predicción generada por el modelo LSTM durante el periodo de validación para el arreglo SD-750.

La figura 4.4 muestra la comparación entre los valores reales y las predicciones generadas por el modelo en el conjunto de validación para el caso del SD-750. Se aprecia un ajuste razonable, especialmente durante los períodos de operación estable. En contraste, en la figura 4.5, correspondiente al SD-1500, se observa una mayor discrepancia en los meses con descensos pronunciados, relacionados con caídas operativas, las cuales el modelo no logra capturar completamente.

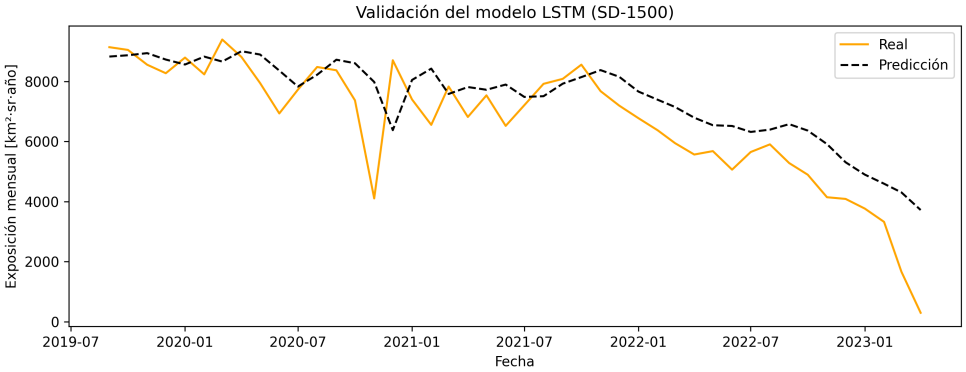

Figura 4.5: Comparación entre la exposición mensual real y la predicción generada por el modelo LSTM durante el periodo de validación para el arreglo SD-1500.

Conclusiones

Los resultados obtenidos muestran que el modelo LSTM es capaz de capturar parcialmente la tendencia general en el crecimiento mensual de la exposición de los arreglos. Sin embargo, su desempeño presenta limitaciones importantes que deben considerarse.

En primer lugar, es importante señalar que el cálculo de la exposición utilizado en este trabajo se basa en una reconstrucción aproximada a partir de la suma de incrementos. Esta aproximación no incorpora información detallada sobre la operación de los detectores individuales; como paros parciales, fallos en estaciones específicas o condiciones meteorológicas adversas, los cuales pueden tener un impacto significativo en la variabilidad real de la exposición. Por ende, ciertas fluctuaciones abruptas observadas en los datos no pueden ser explicadas ni anticipadas por el modelo, lo que introduce un sesgo sistemático significativo.

En segundo lugar, si bien el modelo mostró un desempeño aceptable para el SD-750 ($R^2 = 0.698$), no alcanzó el umbral de $R^2 > 0.7$ que comúnmente se considera como indicador mínimo de validez predictiva en investigaciones científicas. En el caso del SD-1500, el valor de $R = 0.579$ refleja una menor capacidad explicativa del modelo, lo que puede atribuirse tanto a la mayor complejidad de la serie temporal como a una mayor variabilidad estructural en los datos. Por lo tanto, este modelo no debe considerarse como la herramienta para extrapolar la exposición.

En consecuencia, la arquitectura LSTM, si bien potente para series con patrones complejos o no lineales, no resultó adecuada para este tipo de datos. Se considera que enfoques más simples, basados en modelos aditivos o lineales adaptativos, como Prophet, podrían

ser más apropiados, al menos como línea base, para predecir tendencias suaves en el enfoque de la exposición acumulada.

4.4.2 Implementación del modelo Prophet

A diferencia de LSTM, Prophet no requiere grandes volúmenes de datos ni entrenamiento profundo, y puede generar predicciones confiables con bajo costo computacional.

Preparación del conjunto de datos

Antes de aplicar Prophet, es necesario ajustar el formato del conjunto de datos según los requerimientos del modelo: la columna temporal debe llamarse `ds` (de datestamp), y la variable a predecir `y`. A partir de la curva de exposición acumulada obtenida de 4.6 y reestructura el DataFrame para que Prophet pueda interpretarlo correctamente. El bloque de código 4.13 sigue el análisis aplicado al conjunto de datos del SD-1500.

Código 4.13: Preparación de los datos.

```
# Tomar datetime y exposicion acumulada (en km^{2} sr)
df_prophet = expo_1500[["datetime", "Exposure_cumulative"]].copy()

# Renombrar columnas como exige Prophet
df_prophet.columns = ["ds", "y"]
```

Para evaluar el desempeño del modelo, se llevó a cabo una división temporal del conjunto de datos. El modelo se ajusta únicamente con la información anterior al 31 de diciembre de 2019, fecha de corte definida, y posteriormente se valida la capacidad de extrapolación sobre datos posteriores no vistos durante el ajuste:

Código 4.14: División de datos en entrenamiento y prueba.

```
cutoff_date = "2019-12-31"
df_train = df_prophet[df_prophet["ds"] <= cutoff_date]
df_test = df_prophet[df_prophet["ds"] > cutoff_date]
```

Implementación y optimización del modelo

Se ajustaron las dos versiones del modelo Prophet, permitiendo comparar tanto una tendencia constante como un posible límite superior.

En ambos casos, se activó la estacionalidad anual (`yearly_seasonaluty=True`) y se ajustaron los hiperparámetros que regulan la flexibilidad del modelo frente a cambios (`changepoint_prio_scale`) y a variaciones periódicas (`seasonality_prior_scale`).

EL modelo con crecimiento lineal se entrena directamente con el conjunto de entrenamiento. Para el caso logístico, se especifica además un parámetro de capacidad (`cap`) que representa el valor máximo alcanzable por la serie, definido aquí como un 20 % por encima del valor máximo observado, que fue el que mejor se ajustó en la validación.

Código 4.15: Ajuste de modelos: lineal y logistico.

```
from prophet import Prophet

```

```
model_lin = Prophet(
    growth="linear",
    yearly_seasonality=True,
    changepoint_prior_scale=0.3,
    seasonality_prior_scale=10.0
)
model_lin.fit(df_train)

capacidad_maxima = 1.2 * df_prophet["y"].max()

df_train_log = df_train.copy()
df_train_log["cap"] = capacidad_maxima

model_log = Prophet(
    growth="logistic",
    yearly_seasonality=True,
    changepoint_prior_scale=0.3,
    seasonality_prior_scale=10.0
)
model_log.fit(df_train_log)
```

Una vez entrenados los modelos, se generaron las predicciones para el periodo de prueba. Prophet requiere la creación explícita de un DataFrame que contenga las fechas futuras sobre las que se desea proyectar la serie.

Se calculó la duración del intervalo de prueba en días, y se utilizaron los métodos `make_future_dataframe` y `predict` para generar las predicciones tanto para el modelo lineal como para el logístico.

Código 4.16: Generación de predicciones.

```
# Duracion del test
periods = (df_test["ds"].max() - df_test["ds"].min()).days + 1

# Lineal
future_lin = model_lin.make_future_dataframe(periods=periods)
forecast_lin = model_lin.predict(future_lin)

# Logistico
future_log = model_log.make_future_dataframe(periods=periods)
future_log["cap"] = capacidad_maxima
forecast_log = model_log.predict(future_log)
```

Evaluación

Para cuantificar el rendimiento predictivo de los modelos Prophet, se compararon las predicciones generadas con los valores reales de exposición acumulada en el conjunto de validación. Posteriormente, se calcularon tres métricas estándar:

Código 4.17: Evaluación del modelo mediante métricas de error.

```
# Reindexar sobre las fechas del test set
fechas_eval = df_test["ds"]
y_true = df_test["y"].values
```

```
4
5  yhat_lin = forecast_lin.set_index("ds").reindex(fechas_eval, method="
       nearest")["yhat"].values
6  yhat_log = forecast_log.set_index("ds").reindex(fechas_eval, method="
       nearest")["yhat"].values
7
8  # Metricas
9  def evaluar(y_true, y_pred):
10     rmse = np.sqrt(mean_squared_error(y_true, y_pred))
11     mae = mean_absolute_error(y_true, y_pred)
12     r2 = r2_score(y_true, y_pred)
13     return rmse, mae, r2
14
15 rmse_lin, mae_lin, r2_lin = evaluar(y_true, yhat_lin)
16 rmse_log, mae_log, r2_log = evaluar(y_true, yhat_log)
17
18 print("=== Evaluacion sobre el conjunto de validacion ===")
19 print("\nModelo lineal:")
20 print(f"  RMSE = {rmse_lin:.2f}")
21 print(f"  MAE  = {mae_lin:.2f}")
22 print(f"  R^{2}   = {r2_lin:.4f}")
23
24 print("\nModelo logistico:")
25 print(f"  RMSE = {rmse_log:.2f}")
26 print(f"  MAE  = {mae_log:.2f}")
27 print(f"  R^{2}   = {r2_log:.4f}")
```

Resultados

El cuadro 4.4 resume las métricas de validación obtenidas en ambos conjuntos de datos. Por otro lado, la figura (4.6 arriba) muestra la evolución de la exposición acumulada del arreglo SD-1500 junto con las validaciones generadas por Prophet bajo las dos configuraciones. En ambos casos se incluyen las bandas de incertidumbre del 95 %. Como se observa, ambos modelos capturan adecuadamente la tendencia general de los datos, aunque el modelo logístico muestra un mejor ajuste en el tramo final del período, cuando se comienza a evidenciar una ligera desaceleración en el crecimiento. Este comportamiento se refleja también en las métricas de evaluación.

Cuadro 4.4: Métricas de validación de modelos: lineal y logístico para SD-750 y SD-1500.

Arreglo	**Modelo**	**RMSE**	**MAE**	R^2
SD-750	Lineal	161.82	101.64	0.928
	Logístico	191.34	142.65	0.898
SD-1500	Lineal	34,974.87	27,033.49	0.781
	Logístico	8,631.40	8,108.37	0.976

En el caso del SD-750 (figura 4.6 abajo) ambos modelos también reproducen correctamente la tendencia observada. Sin embargo, a diferencia del SD-1500, el modelo lineal presenta una ligera ventaja en las métricas de evaluación; esto se debe a que a inicios de

2022 hay una caída en la tendencia de exposición, por lo que el que mejor suaviza esta caída es el modelo logístico.

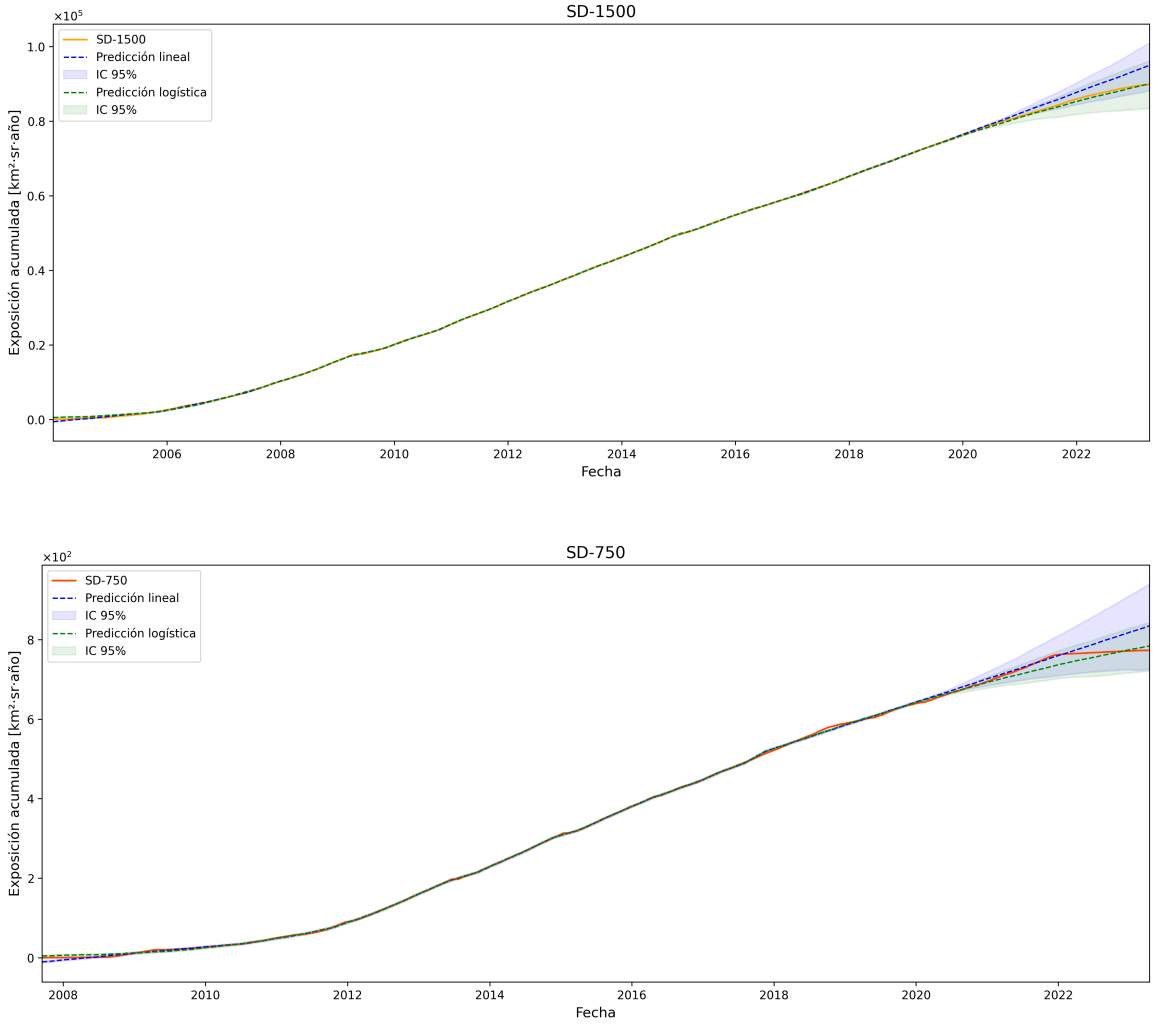

Figura 4.6: Evolución de la exposición acumulada y validaciones de las dos configuraciones de Prophet. *Arriba:* SD-1500 *Abajo:* SD-750

Conclusiones

Es importante destacar que, a pesar del excelente desempeño obtenido por los modelos, no se observan indicios de sobreajuste. Esto se debe a que Prophet no pertenece a la clase de modelos altamente parametrizados, sino que implementa un enfoque aditivo regularizado, donde la tendencia, la estacionalidad y los cambios estructurales son aprendidos de manera controlada mediante parámetros de suavizado.

Además, las bandas de incertidumbre generadas por Prophet reflejan adecuadamente la propagación del error, y el buen ajuste se mantiene en un horizonte temporal corto. Lo que nos lleva a reforzar la validez del modelo como herramienta de extrapolación y no como simple interpolador.

4.5 Extrapolación de la exposición

Se optó por el modelo logístico, en concordancia con la lógica implementada en los códigos anteriores, dado que este enfoque permite capturar de forma más realista el comportamiento de la tendencia de exposición.

Para realizar la extrapolación, se empleó un nuevo modelo logístico previamente ajustado, en concordancia con la lógica implementada en los códigos anteriores, dado que este enfoque permite capturar de forma más realista el comportamiento de la tendencia de exposición. Pero esta vez utilizando toda la serie temporal disponible de los arreglos. Se definió una capacidad máxima equivalente al 120 % del valor final observado, con el fin de limitar ligeramente el crecimiento proyectado. A partir de este ajuste, se generó una predicción extendida hasta el 31 de diciembre de 2028.

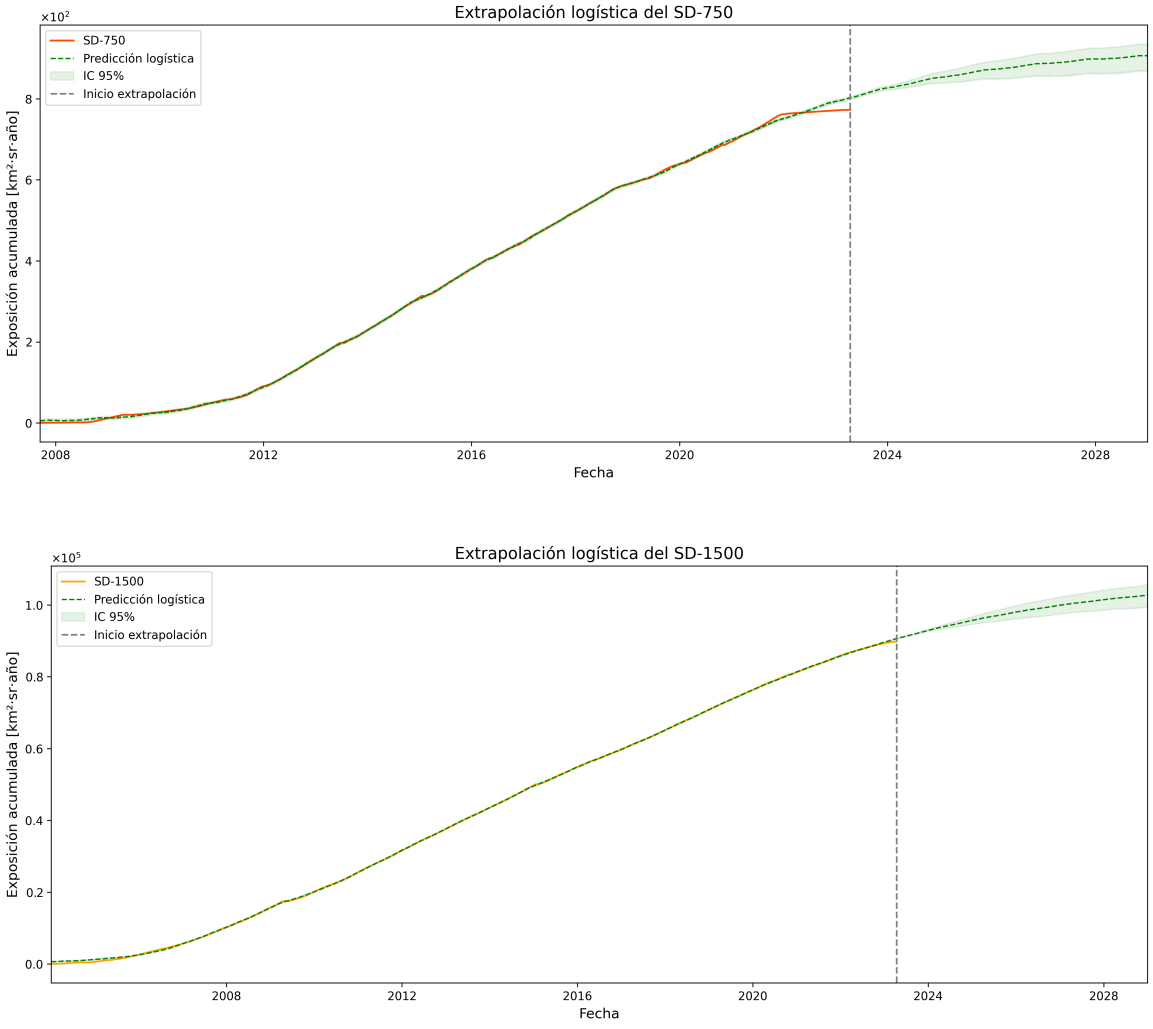

Figura 4.7: Extrapolación. *Arriba:* SD-750 *Abajo:* SD-1500

La figura 4.7 muestra la extrapolación de la exposición acumulada para ambos arreglos. Para el SD-1500, se observa una proyección suave (línea punteada verde) y consistente con

la tendencia histórica (línea naranja), alcanzando una exposición estimada que creció de $89,777\ \mathrm{km}^2 \cdot \mathrm{sr} \cdot \text{año}$ a $102,734.6\ \mathrm{km}^2 \cdot \mathrm{sr} \cdot \text{año}$ al año 2028, lo que representa un incremento del 14.4 %. La línea punteada gris indica el punto a partir del cual inicia la extrapolación.

Por su parte, la figura (4.7 abajo) muestra la extrapolación del SD-750 con un ajuste más marcado en los últimos años, debido a lo que se discutió en la sección anterior. En este caso, la exposición acumulada pasó de $772.7\ \mathrm{km}^2 \cdot \mathrm{sr} \cdot \text{año}$ a $906.34\ \mathrm{km}^2 \cdot \mathrm{sr} \cdot \text{año}$, con un aumento del 17.3 %.

Las bandas de incertidumbre asociadas son estrechas en ambos casos, lo que sugiere un bajo error. Estos resultados permiten proyectar la sensibilidad futura de cada arreglo y en la reconstrucción del espectro de energía de rayos cósmicos.

Capítulo 5

Estimación del espectro de energía

La reconstrucción del espectro de energía requiere considerar eventos detectados en condiciones de eficiencia plena. Aquí, la probabilidad de detección es constante y únicamente determinada por la geometría del arreglo, su ángulo de aceptación y el tiempo efectivo de operación. Esto permite que el número de eventos registrados sea directamente proporcional al flujo incidente, sin depender de factores como la masa del primario o variaciones de eficiencia instrumental, lo que facilita una estimación cruda del espectro a partir de la exposición acumulada.

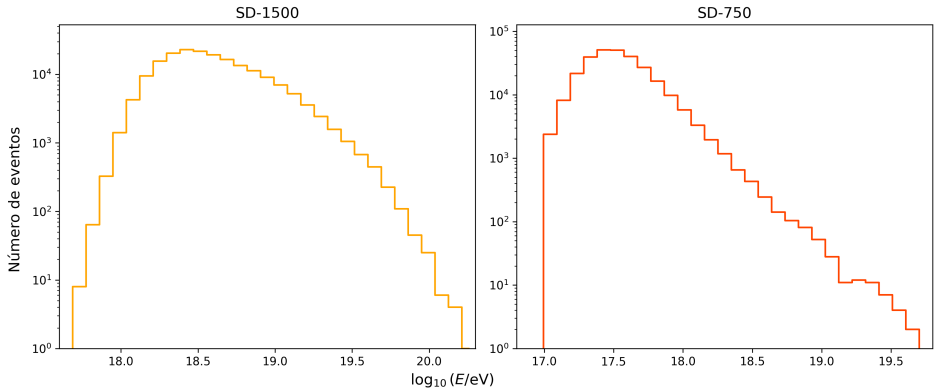

Figura 5.1: Distribución del número de eventos reconstruidos en función de la energía. Los histogramas se muestran en escala log-log.

Con el objetivo de visualizar el umbral energético a partir del cual cada arreglo (SD-1500 y SD-750) alcanza dicho régimen de eficiencia geométrica, se analiza la distribución del número de eventos reconstruidos en función de la energía. La figura 5.1 muestra dicha distribución utilizando un binning uniforme en $\log_{10}(E/\mathrm{eV})$, representada en escala log-log para destacar la transición entre la región de eficiencia creciente y la región de eficiencia plena.

El comportamiento observado en ambas distribuciones sigue un patrón característico: un incremento inicial del número de eventos conforme aumenta la energía, seguido de un

máximo y, posteriormente, una caída exponencial. Esta evolución puede interpretarse de la siguiente manera:

- La región ascendente corresponde al régimen donde la eficiencia del arreglo aún no es total, y el número de eventos detectados crece debido a la mayor probabilidad de disparo del CDAS.

- La región descendente refleja la caída real del espectro físico de rayos cósmicos, regido por una ley de potencias negativa.

El punto de inflexión entre ambas regiones se asocia al umbral energético de eficiencia geométrica. En concordancia con lo reportado por la Colaboración Pierre Auger en [50] y [51], este umbral puede identificarse en:

- $\log_{10}(E/\text{eV}) \approx 18.4$ para el arreglo SD-1500, es decir, $E \geqslant 2.5 \times 10^{18}$ eV,

- $\log_{10}(E/\text{eV}) \approx 17.0$ para el arreglo SD-750, es decir, $E \geqslant 10^{17}$ eV. Sin embargo, se observa que el umbral efectivo en los datos analizados se encuentra en el rango de $\log_{10}(E/\text{eV}) \approx 17.4$, lo cual se le atribuye a los cortes de calidad aplicados, que eliminan eventos cercanos al umbral instrumental, como se menciona en [36].

Estos valores delimitan los rangos de energía a partir de los cuales se puede asumir eficiencia plena en la recolección de eventos. El cuadro 5.1 presenta un resumen estadístico de los eventos después de aplicar los umbrales como criterios de selección para construir la estimación del espectro de energía.

Cuadro 5.1: Distribución porcentual de eventos reconstruidos por intervalo de energía para los conjuntos SD-1500 y SD-750.

Intervalo de energía [eV]	**SD-1500**	**SD-750**
$10^{17.4} \leqslant E < 10^{18}$	0(0.00 %)	187,336 (94.30 %)
$10^{18} \leqslant E < 10^{19}$	110,560 (83.54 %)	11,232 (5.65 %)
$10^{19} \leqslant E < 10^{20}$	21,757 (16.44 %)	85 (0.04 %)
$E \geqslant 10^{20}$	21 (0.02 %)	0 (0.00 %)
Total de eventos	132,338 (100.00 %)	198,653 (100.00 %)

El conjunto de eventos del SD-1500 se redujo en aproximadamente 29.56 % tras aplicar el umbral energético. El conjunto del SD-750 se redujo en aproximadamente 29.04 % por el mismo criterio.

De manera destacada, se identificaron **21 eventos** con energías superiores a 10^{20} eV, los cuales se consideran eventos ultraenergéticos por encontrarse en la región más alta del espectro observado, y representan aproximadamente el **0.02 %** del total. Esta presencia, aunque escasa, es relevante para estudios con el corte GZK y el origen de los rayos cósmicos.

5.1 El espectro de energía a partir de los datos

A partir del cálculo de la exposición acumulada en el capítulo anterior, es posible estimar el espectro de energía mediante un enfoque directo que relaciona el número de eventos detectados, considerando el intervalo energético correspondiente. Bajo esta formulación, el espectro diferencial de rayos cósmicos $J(E)$ se define como [50]:

$$J(E) = \frac{N}{\varepsilon \cdot \Delta E}, \tag{5.1.1}$$

donde N es el número de eventos contenidos en un bin de energía centrado en E, ΔE representa el ancho del bin en unidades de energía, y ε es la ya conocida exposición acumulada del arreglo.

Para construir el histograma energético, siguiendo a [54] se adopta una discretización uniforme en la variable $\log_{10}(E/\text{eV})$, con un ancho de bin de $\Delta \log_{10}(E) = 0.1$. Esta elección responde a la resolución energética estimada del sistema de rango inferior de energías y permite representar adecuadamente la estructura del espectro. En ambos arreglos se define el primer bin en el umbral de energía que se identificó. La implementación correspondiente se encuentra en el código 1, incluido en el Apéndice A.

La figura 5.2 presenta el resultado de esta estimación para ambos arreglos. El espectro obtenido permite observar la caída general del flujo de energía, en particular, la presencia del tobillo (ankle) y la posible supresión a energías ultraaltas. En esta etapa, el espectro se considera crudo, ya que no se han aplicado correcciones sistemáticas adicionales ni ajustes de tipo composición o migración energética.

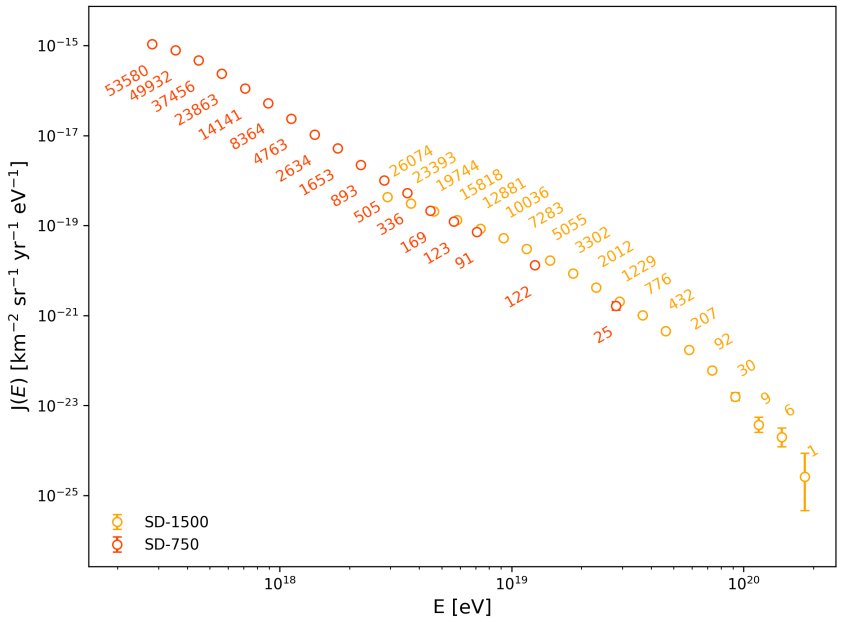

Figura 5.2: Estimación del espectro de energía a partir de los datos del SD-750 y SD-1500. Las curvas representan el flujo $J(E)$ en función de la energía reconstruida E, con barras de error correspondientes a las incertidumbres estadísticas.

Con el propósito de destacar las estructuras características del espectro, se escaló el flujo por E^3 (ver figura 5.3). Esto permite una mejor identificación de la energía correspondiente al tobillo y la supresión GZK.

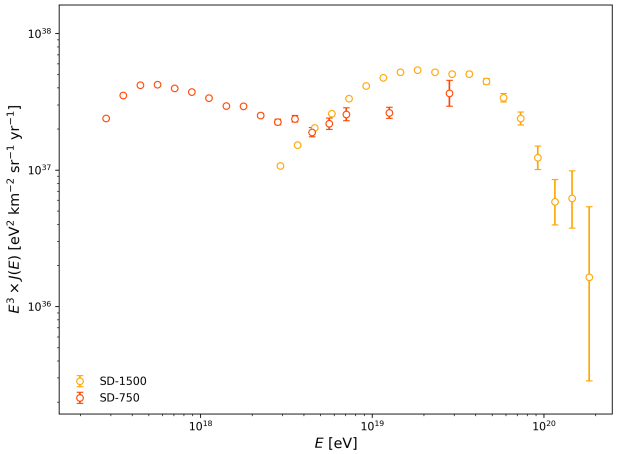

Figura 5.3: El espectro escalado por E^3.

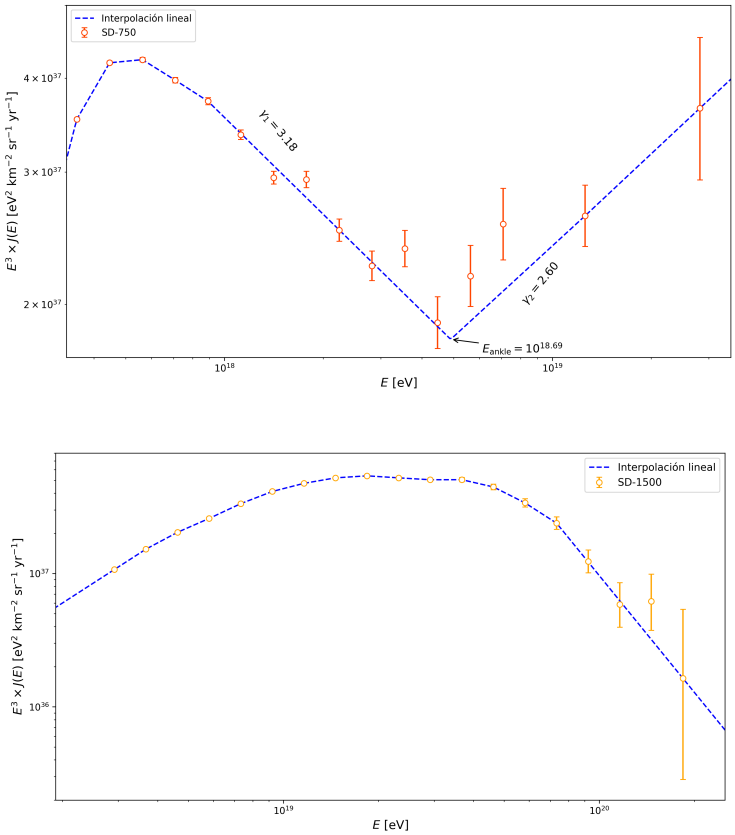

Figura 5.4: Interpolación del espectro escalado $E^3 \times J(E)$. Se observa el tobillo y el cambio de pendiente espectral. Además se observa la supresión del flujo en el régimen de energías ultraaltas. *Arriba:* SD-750 *Abajo:* SD-1500

Se realizó una interpolación lineal en escala logarítmica para cada uno de los arreglos por separado. En la figura (5.4 arriba) se presenta el resultado para el SD-750, donde se observa de forma clara la presencia del tobillo alrededor de $E_{ankle} \approx 10^{18.7}$ eV, así como los cambios en la pendiente espectral. Este comportamiento es compatible con el marco teórico actual, el cual postula una transición en el origen y los mecanismos de propagación de los RC. Las pendientes antes y después del tobillo fueron estimadas como $\gamma_1 \approx 3.18$ y $\gamma_2 = 2.60$, valores que son cercanos a los reportados en [18].

En la figura (5.4 abajo) se muestra el espectro interpolado para el SD-1500, el cual extiende la sensibilidad hacia el régimen de energías ultraaltas. Se identifica una caída del flujo (alrededor de 10^{20} eV), coherente con la supresión de GZK, fenómeno atribuido a la interacción de los RC con el CMB. La interpolación permite visualizar de manera continua los cambios en el flujo de RC sin necesidad de ajustar un modelo específico.

5.2 Predicción del espectro

Aquí se aborda el objetivo central de esta tesis: **evaluar el impacto que tiene la extrapolación de la exposición acumulada sobre la estimación del espectro de energía de rayos cósmicos**. Como se ha discutido, el flujo diferencial $J(E)$ está normalizado por la exposición total del detector, por lo que **un incremento en esta variable reduce proporcionalmente los valores estimados del flujo**, manteniendo inalterada la forma general del espectro.

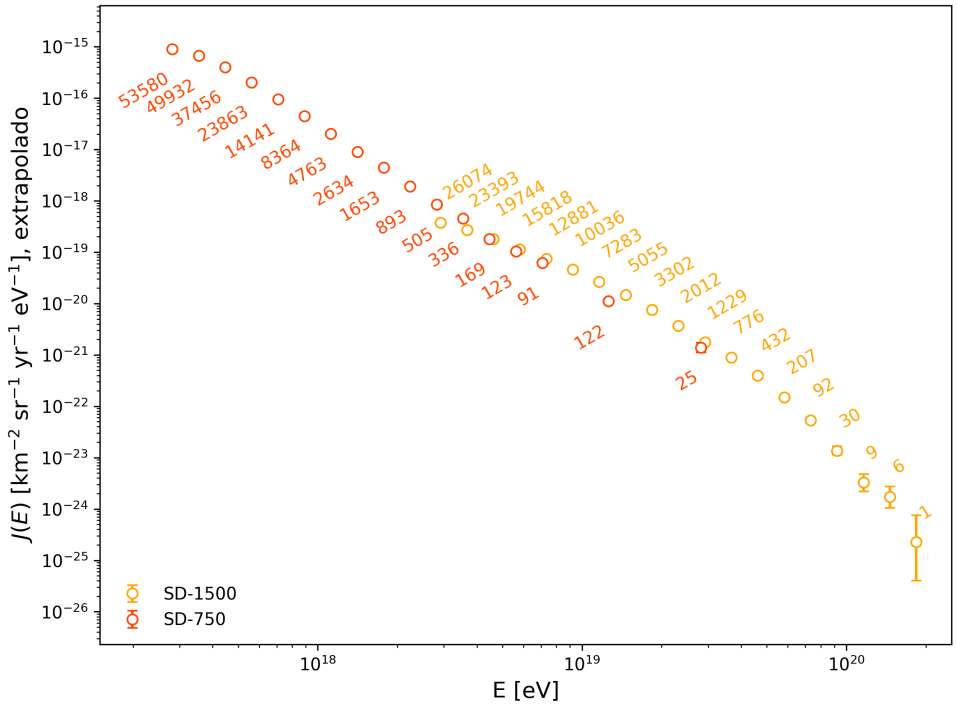

Figura 5.5: Flujo $J(E)$ a partir de la extrapolación $\varepsilon_{1500} = 102734.6 \left[\text{km}^{-2}\text{sr}^{-1}\text{yr}^{-1}\right]$ y $\varepsilon_{750} = 906.34 \left[\text{km}^{-2}\text{sr}^{-1}\text{yr}^{-1}\right]$.

Cuadro 5.2: Comparación entre los valores extremos del flujo obtenido con los datos reales y los valores extrapolados para ambos arreglos.

Arreglo	$J(E)_{max}$**(datos)** $J(E)_{max}$**(extrap.)**	$J(E)_{min}$**(datos)** $J(E)_{min}$**(extrap.)**
SD-750	1.07×10^{-15} 9.09×10^{-16}	1.63×10^{-21} 1.39×10^{-21}
SD-1500	4.31×10^{-19} 3.77×10^{-19}	2.62×10^{-25} 2.29×10^{-25}

Al incorporar la exposición proyectada hasta el 31 de diciembre de 2028, estimada mediante el modelo Prophet, se observa en la figura 5.5 que el espectro reconstruido presenta una **magnitud ligeramente menor** en comparación con el espectro basado exclusivamente en los datos observados hasta el 14 de abril de 2023.

Para cuantificarlo, el cuadro 5.2 resume los valores mínimo y máximo del flujo $J(E)$ obtenidos para ambos arreglos, comparando los casos con y sin extrapolación de la exposición. Esta comparación, aunque aparentemente modesta, muestra que la incorporación de modelos predictivos permite extender la estimación del flujo.

Con el objetivo de complementar esta comparación y analizar el efecto del binning en la resolución espectral, se incluye en el Anexo B una versión del espectro (figura 1) utilizando un ancho de bin de 0.05.

Capítulo 6

Conclusiones

Los resultados obtenidos confirman que la evolución del espectro reconstruido es congruente con las mediciones de [50] en el rango del SD-1500 y [51] en el rango del SD-750. El reescalamiento del flujo por E^3 permitió identificar con mayor claridad la ubicación del tobillo y se logró estimar a groso modo las pendientes en ese rango, consistente con el estudio de [18]. También se identificó la supresión del flujo a ultra altas energías, razonable con el corte GZK. La extrapolación de la exposición permitió extender la sensibilidad del análisis hacia años no vistos, sin modificar la morfología del espectro. Esto sugiere que, dentro de los márgenes de incertidumbre considerados, los efectos observados son compatibles con el marco teórico actual.

El modelo Prophet demostró un desempeño adecuado para intervalos de extrapolación moderados. La predicción a cinco años mostró una tendencia consistente y confiable; sin embargo, al extender la proyección a ocho años (donde se anticipaba una contribución más significativa a la exposición extrapolada en el espectro de energía), las bandas de incertidumbre se ampliaron considerablemente, reflejando una pérdida de precisión y una limitación inherente del modelo en su capacidad de extrapolación a largo plazo.

Una de las principales contribuciones de este trabajo consistió en la aplicación de modelos de Machine Learning para la predicción del flujo del espectro energético, en contraste con los enfoques tradicionales basados en simulaciones tipo Monte Carlo. Si bien los métodos probabilísticos ofrecen un marco detallado y físicamente fundamentado, su implementación implica un alto costo computacional y requiere asumir múltiples condiciones iniciales, distribuciones y parámetros físicos.

En cambio, los modelos implementados, permitieron capturar patrones subyacentes a partir de los datos, capturando tendencias físicas sin necesidad de modelar explícitamente los procesos. Esto los convirtió en herramientas especialmente útiles para la estimación de la exposición acumulada y la proyección preliminar del espectro energético, con una reducción significativa en los tiempos de computo y mayor flexibilidad frente a escenarios complejos.

6.1 Oportunidades de mejora

Aunque los resultados obtenidos con la red neuronal LSTM no fueron satisfactorios en términos predictivos, no se descarta su potencial. Se indentifican diversas oportunidades de mejora, tanto en la representación de los datos como en el diseño del modelo, especialmente si se cuenta con una estimación más robusta de la exposición integrada, que incorpore no solo la acumulación temporal de eventos, sino también la eficiencia operacional del arreglo, posibles períodos de inactividad y las variaciones en la geometría efectiva del detector, permitiría una caracterización más precisa. Esto contrasta con la aproximación utilizada en esta tesis, basada únicamente en la exposición acumulativa.

Por el lado computacional, sería valioso explorar técnicas más avanzadas de modelado secuencial, como *Transformers* adaptados a series de tiempo, modelos híbridos basados en aprendizaje profundo o enfoques *bayesianos* que integren explícitamente la incertidumbre observacional.

En paralelo, se podrían aplicar estas metodologías a la predicción directa del espectro, incorporando como entrada variables experimentales reconstruidas (energía, ángulo cenital, número de estaciones activadas, etc.) y no únicamente la exposición. También se podrían integrar arquitecturas de predicción acopladas.

Desde el punto de vista físico, algo relevante sería contrastar el espectro extrapolado con modelos teóricos de propagación de rayos cósmicos, en particular simulaciones que incluyan efectos de pérdida de energía, composición y estructuras de origen astrofísico. Esto permitiría evaluar si la evolución observada en el espectro es compatible con diferentes escenarios físicos.

Apéndice A

Código 1: Código adaptado a partir del Open Data de la Colaboración Pierre Auger [54] para el análisis del espectro de energía.

```
# CARGA Y PREPARACION DE DATOS
# Cargar archivos
df_sd1500 = pd.read_csv("completo_SD.csv")
df_sd750  = pd.read_csv("completo_SDinfill.csv")

# Umbrales de energia [eV]
energy_threshold_SD1500 = 10**18.4
energy_threshold_SD750  = 10**17.4

# Filtrado por energia y no nulos
df_sd1500 = df_sd1500[
    (df_sd1500["Energy"].notna()) &
    (df_sd1500["Energy"] >= energy_threshold_SD1500)
].sort_values(by="EventId")

df_sd750 = df_sd750[
    (df_sd750["Energy"].notna()) &
    (df_sd750["Energy"] >= energy_threshold_SD750)
].sort_values(by="EventId")

# Energias unicas por evento
energy    = df_sd1500.drop_duplicates("EventId")["Energy"]
energyLE  = df_sd750.drop_duplicates("EventId")["Energy"]

# Exposicion acumulada
exposure    = 89777.007164
exposureLE  = 772.669038

# DEFINICION DE BINS DE ENERGIA
# SD_1500
log_E_min        = np.log10(2.5e18)
E_bins           = 20
E_bin_size       = 0.1
log_E_max        = log_E_min + E_bins * E_bin_size
log_bins         = np.linspace(log_E_min, log_E_max, E_bins + 1)
log_bin_centers  = log_bins[:-1] + 0.05
bins             = 10 ** log_bins
bin_energy18     = 10 ** log_bin_centers
bin_width        = bins[1:] - bins[:-1]
```

```
40
41 # SD_750 (manualmente)
42 binslogP              = np.array([
43     17.4, 17.5, 17.6, 17.7, 17.8, 17.9, 18.,18.1, 18.2,
44     18.3, 18.4, 18.5, 18.6, 18.7, 18.8, 18.9, 19.3, 19.6
45 ])
46 log_bin_centersLE     = binslogP[:-1] + (binslogP[1:] - binslogP[:-1])/2
47 binsP                 = 10 ** binslogP
48 bin_energy18_LE       = 10 ** log_bin_centersLE
49 bin_widthLE           = binsP[1:] - binsP[:-1]
50
51 # HISTOGRAMACION DE EVENTOS
52 h, _    = np.histogram(energy, bins=bins)
53 hLE, _ = np.histogram(energyLE, bins=binsP)
54
55 # CALCULO DE INCERTIDUMBRES ESTADISTICAS
56 alpha, beta = 0.16, 0.16
57
58 # SD_1500
59 lim_low      = h-np.nan_to_num(0.5*scipy.stats.chi2.ppf(alpha,2*h))
60 lim_up       = 0.5*scipy.stats.chi2.ppf(1-beta,2*(h+1))- h
61 cut_nz       = h > 0
62 cut_z        = h == 0
63
64 # SD_750
65 lim_low_LE   = hLE-np.nan_to_num(0.5*scipy.stats.chi2.ppf(alpha,2*hLE))
66 lim_up_LE    = 0.5*scipy.stats.chi2.ppf(1-beta,2*(hLE+1))-hLE
67 cut_nzLE     = hLE > 0
68 cut_zLE      = hLE == 0
69
70 # CALCULO DEL FLUJO DIFERENCIAL
71 # SD_1500
72 normalization    = exposure * bin_width
73 flux             = h[cut_nz] / normalization[cut_nz]
74 flux_lower       = lim_low[cut_nz] / normalization[cut_nz]
75 flux_upper       = lim_up[cut_nz] / normalization[cut_nz]
76
77 # SD_750
78 normalization_LE    = exposureLE * bin_widthLE
79 flux_LE             = hLE[cut_nzLE] / normalization_LE[cut_nzLE]
80 flux_lower_LE       = lim_low_LE[cut_nzLE] / normalization_LE[cut_nzLE]
81 flux_upper_LE       = lim_up_LE[cut_nzLE] / normalization_LE[cut_nzLE]
82
83 # VISUALIZACION DEL ESPECTRO DIFERENCIAL
84 plt.figure(figsize=(8, 6))
85
86 # SD_1500
87 plt.errorbar(
88     bin_energy18[cut_nz], flux,
89     yerr=[flux_lower, flux_upper],
90     fmt="o",
91     label='SD-1500',
92     color="#FFA500",
93     capsize=3,
```

```
    markerfacecolor='white',
    markeredgecolor="#FFA500"
)

# SD_750
plt.errorbar(
    bin_energy18_LE[cut_nzLE], flux_LE,
    yerr=[flux_lower_LE, flux_upper_LE],
    fmt="o",
    label='SD-750',
    color="#FF4500",
    capsize=3,
    markerfacecolor='white',
    markeredgecolor="#FF4500"
)

plt.xscale("log")
plt.yscale("log")
plt.xlim(1.5e17, 2.5e20)
plt.xlabel('E [eV]')
plt.ylabel(r'J$(E)$ [km$^{-2}$ sr$^{-1}$ yr$^{-1}$ eV$^{-1}$]')

# Ajuste de limites verticales para buena visibilidad
plt.ylim(flux[flux > 0].min()*0.01, flux_LE.max()*7)

# Etiquetas de conteo SD_1500
for E, J, count in zip(bin_energy18[cut_nz],flux,h[cut_nz]):
    if count > 0:
        plt.annotate(
            count, (E, J * 2.4),
            rotation=30,
            va='bottom',
            ha='left',
            color="#FFA500"
        )

# Etiquetas de conteo SD_750
for E, J, count in zip(bin_energy18_LE[cut_nzLE],flux_LE,hLE[cut_nzLE]):
    if count > 0:
        plt.annotate(
            count, (E, J * 0.4),
            rotation=30,
            va='top',
            ha='right',
            color="#FF4500"
        )

plt.legend(loc='lower left', frameon=False)
plt.tight_layout()
plt.show()
```

Apéndice B

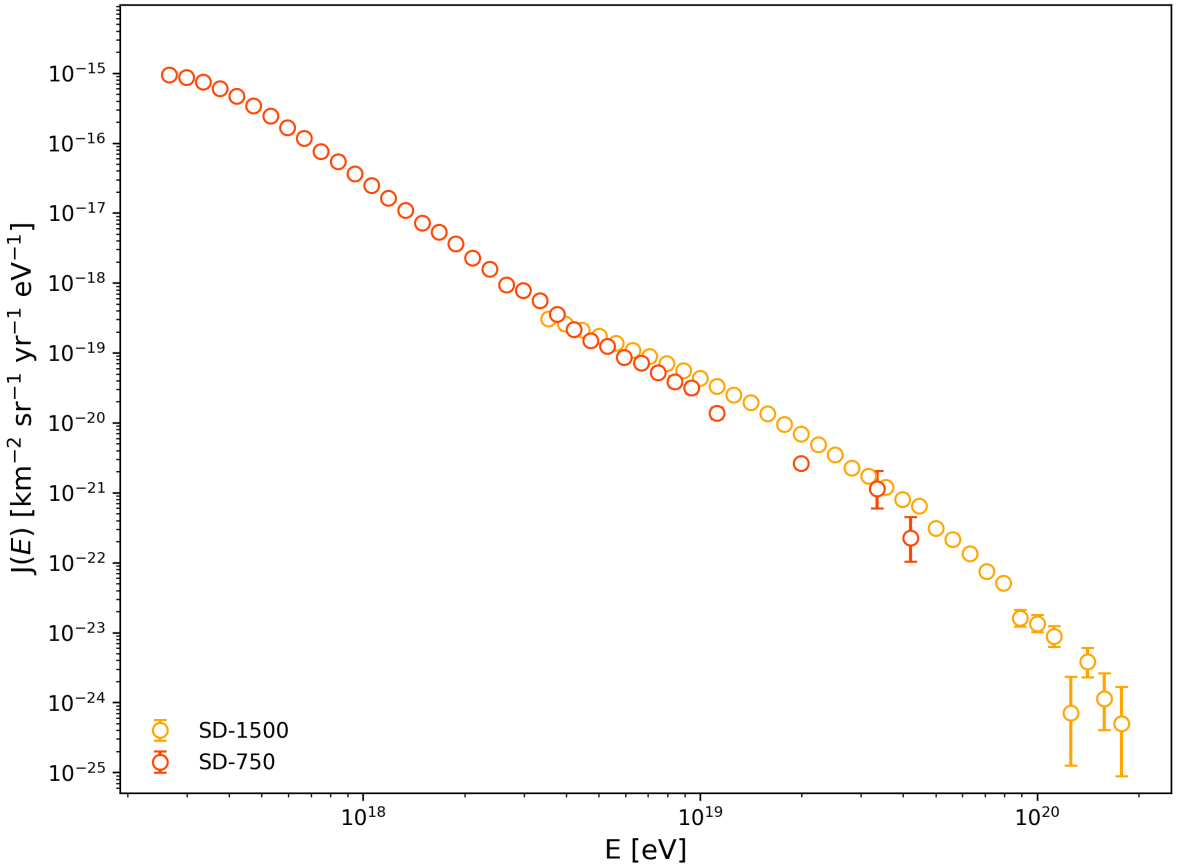

Figura 1: Espectro de energía a partir de la extrapolación de la exposición con un ancho de bin de $\Delta \log_{10}(E) = 0.05$.

Bibliografía

[1] Victor F. Hess. Über Beobachtungen der durchdringenden Strahlung bei sieben Freiballonfahrten. *Phys. Z.*, 13:1084–1091, 1912. 2

[2] D. Skobelzyn. Über eine neue Art sehr schneller β-Strahlen. *Zeitschrift fur Physik*, 54(9-10):686–702, sep 1929. doi:10.1007/BF01341600. 2

[3] H. Geiger and W. Müller. Elektronenzählrohr zur Messung schwächster Aktivitäten. *Naturwissenschaften*, 16(31):617–618, aug 1928. doi:10.1007/BF01494093. 2

[4] W. B. Fretter. Proceedings of echo lake cosmic ray symposium. In *Echo Lake Cosmic Ray Symposium*, 1949. 3

[5] Carl D. Anderson. The positive electron. *Phys. Rev.*, 43:491–494, Mar 1933. URL: https://link.aps.org/doi/10.1103/PhysRev.43.491, doi:10.1103/PhysRev.43.491. 3

[6] Karl-Heinz Kampert and Alan A. Watson. Extensive air showers and ultra high-energy cosmic rays: a historical review. *European Physical Journal H*, 37(3):359–412, aug 2012. arXiv:1207.4827, doi:10.1140/epjh/e2012-30013-x. 3, 12

[7] D. J. Bird, S. C. Corbato, H. Y. Dai, J. W. Elbert, et al. Detection of a Cosmic Ray with Measured Energy Well beyond the Expected Spectral Cutoff due to Cosmic Microwave Radiation. *Astrophysical Journal*, 441:144, mar 1995. arXiv:astro-ph/9410067, doi:10.1086/175344. 4

[8] R. U. Abbasi et al. An extremely energetic cosmic ray observed by a surface detector array. *Science*, 382(6673):abo509, 2023. arXiv:2311.14231, doi:10.1126/science.abo5095. 4

[9] Alexander Aab et al. The Pierre Auger Cosmic Ray Observatory. *Nucl. Instrum. Meth. A*, 798:172–213, 2015. arXiv:1502.01323, doi:10.1016/j.nima.2015.06.058. 4, 5, 15, 16, 17, 18, 19, 20, 22, 41

[10] Kumiko Kotera and Angela V. Olinto. The Astrophysics of Ultrahigh-Energy Cosmic Rays. *Annual Review of Astronomy and Astrophysics*, 49(1):119–153, sep 2011. arXiv:1101.4256, doi:10.1146/annurev-astro-081710-102620. 5, 7

[11] Pasquale Blasi. The Origin of Galactic Cosmic Rays. *Astron. Astrophys. Rev.*, 21:70, 2013. arXiv:1311.7346, doi:10.1007/s00159-013-0070-7. 5, 8

[12] Joerg R. Hoerandel. On the knee in the energy spectrum of cosmic rays. *Astropart. Phys.*, 19:193–220, 2003. arXiv:astro-ph/0210453, doi:10.1016/S0927-6505(02)00198-6. 6

[13] A. M. Hillas. Can diffusive shock acceleration in supernova remnants account for high-energy galactic cosmic rays? *J. Phys. G*, 31:R95–R131, 2005. doi:10.1088/0954-3899/31/5/R02. 6, 8

[14] Carmelo Evoli. The cosmic-ray energy spectrum, oct 2018. doi:10.5281/zenodo.2360277. 6

[15] M. Tanabashi, K. Hagiwara, K. Hikasa, K. Nakamura, Y. Sumino, F. Takahashi, J. Tanaka, et al. Review of particle physics. *Phys. Rev. D*, 98:030001, Aug 2018. URL: https://link.aps.org/doi/10.1103/PhysRevD.98.030001, doi:10.1103/PhysRevD.98.030001. 6

[16] W. D. Apel et al. KASCADE-Grande measurements of energy spectra for elemental groups of cosmic rays. *Astropart. Phys.*, 47:54–66, 2013. arXiv:1306.6283, doi:10.1016/j.astropartphys.2013.06.004. 6

[17] Karl-Heinz Kampert and Michael Unger. Measurements of the Cosmic Ray Composition with Air Shower Experiments. *Astropart. Phys.*, 35:660–678, 2012. arXiv:1201.0018, doi:10.1016/j.astropartphys.2012.02.004. 6, 8, 11, 12

[18] R. U. Abbasi et al. Measurement of the Flux of Ultra High Energy Cosmic Rays by the Stereo Technique. *Astropart. Phys.*, 32:53–60, 2009. arXiv:0904.4500, doi:10.1016/j.astropartphys.2009.06.001. 7, 63, 65

[19] R. Aloisio, V. Berezinsky, and P. Blasi. Ultra high energy cosmic rays: implications of auger data for source spectra and chemical composition. *Journal of Cosmology and Astroparticle Physics*, 2014(10):020, oct 2014. doi:10.1088/1475-7516/2014/10/020. 7

[20] Kenneth Greisen. End to the cosmic-ray spectrum? *Phys. Rev. Lett.*, 16:748–750, Apr 1966. URL: https://link.aps.org/doi/10.1103/PhysRevLett.16.748, doi:10.1103/PhysRevLett.16.748. 7

[21] G. T. Zatsepin and V. A. Kuzmin. Upper limit of the spectrum of cosmic rays. *JETP Lett.*, 4:78–80, 1966. 7

[22] Rafael Alves Batista et al. Open Questions in Cosmic-Ray Research at Ultrahigh Energies. *Front. Astron. Space Sci.*, 6:23, 2019. arXiv:1903.06714, doi:10.3389/fspas.2019.00023. 7

[23] Donald V. Reames. Particle acceleration at the Sun and in the heliosphere. *Space Science Reviews*, 90:413–491, oct 1999. doi:10.1023/A:1005105831781. 8

[24] Hilary V. Cane. Coronal Mass Ejections and Forbush Decreases. *Space Science Reviews*, 93:55–77, jul 2000. doi:10.1023/A:1026532125747. 8

[25] Enrico Fermi. On the Origin of the Cosmic Radiation. *Physical Review*, 75(8):1169–1174, apr 1949. doi:10.1103/PhysRev.75.1169. 8

[26] Malcolm S. Longair. High energy astrophysics. *Cambridge*, pages 564–573, 2011. 9, 12

[27] J. Matthews. A Heitler model of extensive air showers. *Astropart. Phys.*, 22:387–397, 2005. doi:10.1016/j.astropartphys.2004.09.003. 10, 12

[28] Konrad Bernlohr. Simulation of Imaging Atmospheric Cherenkov Telescopes with CORSIKA and sim_telarray. *Astropart. Phys.*, 30:149–158, 2008. arXiv:0808.2253, doi:10.1016/j.astropartphys.2008.07.009. 10

[29] Valdés-Galicia Barrantes, M et al. Atmospheric corrections of the cosmic ray fluxes detected by the solar neutron telescope at the summit of the sierra negra volcano in mexico. *Geofísica Internacional*, 57(4):253–275, oct. 2018. doi:10.22201/igeof.00167169p.2018.57.4.2105. 10

[30] Ralf Ulrich, Ralph Engel, and Michael Unger. Hadronic multiparticle production at ultrahigh energies and extensive air showers. *Physical Review D*, 83(5), mar 2011. doi:10.1103/physrevd.83.054026. 11, 13

[31] Thomas K. Gaisser. *Cosmic rays and particle physics.* Cambridge University Press, 1990. 12

[32] Ralph Engel, Dieter Heck, and Tanguy Pierog. Extensive air showers and hadronic interactions at high energy. *Annual Review of Nuclear and Particle Science*, 61(Volume 61, 2011):467–489, 2011. doi:10.1146/annurev.nucl.012809.104544. 12

[33] Tanguy Pierog, R. Engel, and D. Heck. Impact of uncertainties in hadron production on air-shower predictions. *Czech. J. Phys.*, 56:A161–A172, 2006. arXiv:astro-ph/0602190, doi:10.1007/s10582-006-0152-0. 13

[34] Jakub Vicha and Jiri Chudoba. Data processing at the pierre auger observatory. *Journal of Physics: Conference Series*, 608:012077, 05 2015. doi:10.1088/1742-6596/608/1/012077. 15

[35] J. Abraham, P. Abreu, M. Aglietta, C. Aguirre, et al. The fluorescence detector of the Pierre Auger Observatory. *Nuclear Instruments and Methods in Physics Research A*, 620(2-3):227–251, aug 2010. arXiv:0907.4282, doi:10.1016/j.nima.2010.04.023. 16, 21

[36] Diego Ravignani. Measurement of the Energy Spectrum of Cosmic Rays Above 3x1017 eV Using the AMIGA750 m Surface Detector Array of the Pierre Auger Observation. In *International Cosmic Ray Conference*, volume 33 of *International Cosmic Ray Conference*, page 1762, jan 2013. 17, 60

[37] D. Allard et al. The trigger system of the Pierre Auger Surface Detector: operation, efficiency and stablility. In *29th International Cosmic Ray Conference*, 8 2005. arXiv:astro-ph/0510320. 18

[38] P. A. Cherenkov. Visible radiation produced by electrons moving in a medium with velocities exceeding that of light. *Phys. Rev.*, 52(4):378–379, 1937. doi:10.1103/PhysRev.52.378. 19

[39] John David Jackson. *Classical electrodynamics*. John Wiley & Sons, 2021. 19

[40] I. M. Frank and I. E. Tamm. Coherent visible radiation of fast electrons passing through matter. *Compt. Rend. Acad. Sci. URSS*, 14(3):109–114, 1937. doi:10.3367/UFNr.0093.196710o.0388. 19

[41] A. Aab, P. Abreu, M. Aglietta, E. J. Ahn, et al. Depth of maximum of air-shower profiles at the pierre auger observatory. ii. composition implications. *Phys. Rev. D*, 90:122006, Dec 2014. doi:10.1103/PhysRevD.90.122006. 20

[42] A. Aab et al. Reconstruction of events recorded with the surface detector of the Pierre Auger Observatory. *JINST*, 15(10):P10021, 2020. arXiv:2007.09035, doi:10.1088/1748-0221/15/10/P10021. 22, 23, 24

[43] J. Abraham, P. Abreu, M. Aglietta, et al. Trigger and aperture of the surface detector array of the Pierre Auger Observatory. *Nuclear Instruments and Methods in Physics Research A*, 613(1):29–39, jan 2010. arXiv:1111.6764, doi:10.1016/j.nima.2009.11.018. 22

[44] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. Adaptive Computation and Machine Learning series. MIT Press, 2016. 26, 27, 29, 30, 31

[45] P.J. Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990. doi:10.1109/5.58337. 31

[46] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994. doi:10.1109/72.279181. 32

[47] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 11 1997. doi:10.1162/neco.1997.9.8.1735. 32

[48] Zhenglin Li, Qingxiong Zhu, Dan Zhang, Hao Wu, and Yan Peng. Sea surface temperature prediction enhanced by exploring spatiotemporal correlation based on lstm and gaussian process. *Sensors*, 25(5), 2025. doi:10.3390/s25051373. 33

[49] Sean J. Taylor and Benjamin Letham and. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018. doi:10.1080/00031305.2017.1380080. 34, 35

[50] Alexander Aab et al. Measurement of the cosmic-ray energy spectrum above 2.5×10^{18} eV using the Pierre Auger Observatory. *Phys. Rev. D*, 102(6):062005, 2020. arXiv:2008.06486, doi:10.1103/PhysRevD.102.062005. 43, 46, 60, 61, 65

[51] P. Abreu et al. The energy spectrum of cosmic rays beyond the turn-down around 10^{17} eV as measured with the surface detector of the Pierre Auger Observatory. *Eur. Phys. J. C*, 81(11):966, 2021. arXiv:2109.13400, doi:10.1140/epjc/s10052-021-09700-w. 44, 60, 65

[52] Philipp Meder, David Schmidt, and Darko Veberic. Sd-750 energy spectrum and its features. Auger internal note GAP–2024–40, The Pierre Auger Collaboration, 2024. 46

[53] Ralf C. Staudemeyer and Eric Rothstein Morris. Understanding LSTM – a tutorial into Long Short-Term Memory Recurrent Neural Networks. *arXiv e-prints*, page arXiv:1909.09586, sep 2019. arXiv:1909.09586, doi:10.48550/arXiv.1909.09586. 46

[54] The Pierre Auger Collaboration. Pierre auger observatory open data, 2024. doi:10.5281/zenodo.10488964. 61, 67