FACULTAD DE CIENCIAS
ASTRONÓMICAS Y
GEOFÍSICAS

UNIVERSIDAD NACIONAL DE
LA PLATA

Análisis de Vientos y Mareas en la Mesosfera y Baja Termosfera sobre la Región Patagónica Argentina

Tesis de Grado en Geofísica

Autor
Anasimele, Guillermina Paula
Director
Dr. J. Federico Conte
Co-Directora
Dra. Ángela Érika Gularte Scarone

La Plata, diciembre 2021

Índice general

Índice de figuras

1 Introducción

El estudio de la mesosfera y baja termosfera (o MLT, por su sigla en inglés) resulta primordial para entender el grado de influencia que las capas bajas de la atmósfera terrestre tienen sobre la termosfera/ionosfera. Puntualmente, la ionosfera, medio fundamental para el desarrollo de las telecomunicaciones, se verá afectada, entre otras cosas, por cambios en la dinámica de la MLT. Esto último tendrá un claro impacto sobre la vida moderna. Así, y con el propósito de comprender la dinámica de la MLT, se realizan estudios de vientos y perturbaciones de mareas como parte del análisis general de las ondas atmosféricas, dado que estas últimas constituyen el mecanismo fundamental de intercambio entre las distintas capas atmosféricas.

En este contexto, la Patagonia Argentina se destaca a nivel mundial como una de las regiones más activas en cuanto a dinámica atmosférica, siendo uno de los puntos de mayor interés, la zona en torno a la ciudad de El Calafate (e.g., Trinh et al., 2018). Con el propósito de avanzar en el estudio de esta área, el Instituto Leibniz de Física de la Atmósfera (Institute of Atmospheric Physics, IAP) desarrolló la red de radares de meteoros SIMONe en la provincia de Santa Cruz, Argentina. Esta red permite obtener datos de vientos con una resolución espacio-temporal sin precedentes (Vierinen et al., 2019).

El objetivo propuesto para este trabajo se enfoca en presentar una descripción de las características principales de la dinámica de la mesosfera y baja termosfera sobre la región patagónica argentina a partir de los datos de esta red. En particular, se propone procesar y analizar series temporales a diferentes alturas, de vientos zonal y meridional, a fin de extraer información en altura y tiempo sobre vientos medios y componentes de mareas diurna, semidiurna y terdiurna.

El estudio de vientos y mareas en la mesosfera y baja termosfera basado en mediciones proporcionadas por la red SIMONe Argentina representa una oportunidad única para explorar diferentes técnicas de análisis de datos. Para ello, se aplicarán dos estrategias de acción.

En el primer caso, y basado en los lineamientos del trabajo propuesto por Conte et al. (2017) y Conte et al. (2021), se realizará un ajuste de mínimos cuadrados a series temporales, de vientos zonales y meridionales, para determinadas alturas. Se describirá también la variabilidad diaria y estacional de los vientos medios y mareas estimadas.

La primera estrategia refiere a una técnica tradicional de análisis de datos.

En cambio, la segunda estrategia será la aplicación de un enfoque innovador con aprendizaje automático y minería de datos (Andrienko y Andrienko 2006; Han, Kamber y Pei, 2012). En esta segunda etapa se realizará una preparación y análisis de los datos, luego se elegirá e implementará una técnica de minería para las series temporales de vientos zonal y meridional a distintas alturas. La selección de una técnica de minería de datos deberá responder a un proceso eficiente y que otorgue resultados confiables.

Se describirá la variabilidad diaria y estacional de los vientos medios bajo un proceso automatizado. Finalmente, se compararán los resultados obtenidos con ambas estrategias y se discutirán posibles mecanismos físicos que expliquen los fenómenos objeto de estudio.

A pesar de que las herramientas y técnicas de minería de datos y aprendizaje automático poseen un crecimiento exponencial en cuanto a su aplicabilidad en diferentes campos, aún no se registran trabajos en la temática propuesta para esta tesis.

Esta tesis se ha organizado de la siguiente manera: en el capítulo 2 se introduce la mesosfera y baja termosfera como región de estudio, prestando especial atención a las ondas atmosféricas, en particular, las mareas solares y lunares que tienen gran influencia en la dinámica de este medio. En el capítulo 3 se dan detalles sobre el instrumento y la técnica de medición, el pre-procesamiento de los datos y las características principales del dato recogido. Luego, el capítulo 4 introduce al lector en el modelado realizado por mínimos cuadrados sobre el dato de vientos mesosféricos, con el fin de aproximar los parámetros del modelo de viento propuesto. El capítulo 5 presenta un breve resumen de generalidades de aprendizaje automático (Machine Learning). Esta sección se enfocará en el desarrollo de la técnica sobre series temporales. Luego, en el capítulo 6 se detalla la metodología de ARIMA, el modelo de análisis regresivo a utilizar sobre las series temporales muestreadas, con el fin de estudiar su naturaleza y realizar y un pronóstico de su comportamiento. Finalmente, en el capítulo 7 se presentan los resultados y conclusiones de este trabajo.

2 Física de la región de estudio

El principal interés sobre las perturbaciones de mareas se debe a su influencia sobre la dinámica de la atmósfera de la Tierra. Es el objetivo de este capítulo brindar una introducción a las mareas como ondas atmosféricas, en el marco del medio donde se propagan.

Para esto, primeramente se presenta una descripción general de la atmósfera terrestre en la sección 2.1. Luego, en la sección 2.2 se pone especial atención en la capa objeto de estudio, y se describe el comportamiento de las ondas atmosféricas más importantes para la región. Por último, se resalta la importancia de la zona de estudio en la sección 2.3.

2.1. La atmósfera terrestre

En primer lugar, se define a la atmósfera terrestre como aquella envoltura gaseosa que recubre a la Tierra desde su superficie hasta aproximadamente los 1000 km de altura.

Se considera que el aire que conforma la atmósfera es un agregado de diferentes gases, y es común que se mencionen los porcentajes de concentración para un volumen de aire cercano a la superficie. Entre ellos, predominan el nitrógeno, concentrado en un 78 % y el oxígeno en un 21 %. Además, se encuentran en mucha menor concentración argón, neón, helio, hidrógeno, xenón, dióxido de carbono, metano, ozono, óxido nitroso y los conocidos clorofluorcarburos. La composición del aire cambia constantemente y se renueva conforme estos gases interactúan con la superficie de la Tierra y sus emisiones y, al mismo tiempo, con la radiación solar.

Aunque la atmósfera es un medio de gran dinámica global, pueden bien diferenciarse regiones, de manera que hay variadas clasificaciones según el parámetro de análisis. Es correcto caracterizar capas o regiones principalmente según su estado termodinámico y en función de sus parámetros característicos, a saber, la presión, la densidad y la temperatura. Los gradientes definidos por estos y las interacciones que generan, serán de importancia en el estudio de la atmósfera.

Tanto la presión como la densidad del aire disminuyen con la altura conforme se aleja un punto de estudio de la superficie de la Tierra. La densidad y su variación están relacionadas directamente con la presión y su gradiente. Los máximos valores de presión se encuentran cercanos a la

superficie de la Tierra, en el orden de los 1000 hPa. Se asocia al nivel del mar valores de presión de 1013,25 hPa.

La distribución de presión de la atmósfera suele aplicarse como medida de la altura de manera que es común la utilización de coordenadas definidas a partir de este parámetro. Los mencionados gradientes de densidad y presión poseen mayor variabilidad en la zona baja de la atmósfera, hasta los 50 km, luego varían en menor grado y se normalizan. En cambio, la temperatura del aire en altura varía considerablemente. Es por eso que se presenta la estructura en capas de la atmósfera clasificadas según su estado térmico. Cabe destacar que ésta, junto con la clasificación del estado electromagnético atmosférico, es de las clasificaciones más utilizadas en la bibliografía.

Estructura Térmica de la Atmósfera

Desde la superficie terrestre y hasta aproximadamente los 12 km de altitud, se encuentra la troposfera. En esta capa se aprecia una disminución de la temperatura conforme aumenta la altura. La energía térmica de esta región está íntimamente relacionada al intercambio de calor que se da entre la Tierra misma y la atmósfera. El gradiente térmico es del orden de -6.5°C por kilómetro de elevación. Este valor es ampliamente variable en períodos cortos de tiempo (por ejemplo, de un día al siguiente). Se considera entonces que la troposfera abarca la extensión vertical que culmina donde la temperatura comienza a estabilizarse. Esta zona isotérmica se denomina tropopausa, y su altura, varía según la latitud y según la época del año. Normalmente se encuentra a elevaciones más altas sobre las regiones ecuatoriales, y disminuye hacia los polos, y generalmente, es más alta en verano y más baja en invierno en todas las latitudes. Otros eventos de circulación global también pueden alterar su ubicación. Por estar directamente en contacto con la superficie terrestre, en su mayoría océanos, existe una íntima relación entre las corrientes de aire y las corrientes de masas de agua, de modo que la circulación global de fluidos esta interrelacionada. La troposfera se caracteriza por un alto contenido de vapor de agua, que es considerado un gas invernadero, debido a que absorbe parte de la radiación re-emitida por la Tierra. Este es el componente esencial para que, en este medio, y explicado por sus inestabilidades, ocurran los principales eventos meteorológicos que afectan al ser humano. En general, se considera un importante elemento en el balance térmico del aire, interviniendo por medio de sus cambios de estado en la generación de calor latente y propiciando la circulación de masas de aire en diferentes escalas. También así, la topografía de la superficie terrestre tiene gran incumbencia en los movimientos de las masas de aire, bien propiciando intercambios de energía a nivel molecular, como generando perturbaciones.

Por encima de la tropopausa, se encuentra la estratosfera. Esta capa

se distingue por un cambio pronunciado en el patrón de variación de temperatura. La temperatura comienza a aumentar con la altura, lo que es llamado una "inversión térmica". Este patrón invertido produce estabilidad, de manera que esta es una capa estratificada en sí misma. Existe una disminución de los movimientos que se propagan desde la troposfera, hacia la zona superior, que se acentúa por la isoterma inferior (tropopausa). Por esto, las corrientes de circulación de la capa más baja, en general celdas convectivas variadas, quedan confinadas a la troposfera. A una altitud media en esta capa se encuentra una alta concentración de ozono, muy relacionada al incremento térmico característico de la estratosfera. Este ozono constituye un compuesto de especial importancia en la interacción de la atmósfera con la radiación solar y el balance térmico de la misma. Es posible encontrarlo en las capas bajas de la atmósfera, pero en general, la mayor cantidad (entre el 85 % y el 90 % del ozono atmosférico) está concentrado entre los 15 y los 30 km de altura y precisamente, cerca de los 25 km. En esta región se genera de forma natural por fotólisis, y se estima que no supera un porcentaje mayor del 0,002 % del volumen de aire. Este porcentaje puede variar según la latitud y la época del año. Aunque puede resultar baja la concentración en contraste con otros gases, como el oxígeno, este porcentaje de ozono y su ciclo de recuperación son un factor indispensable en la absorción de radiación de alta frecuencia que recibe la Tierra. Por encima de esta zona rica en ozono, cerca de los 30 km de altura la temperatura alcanza casi los -40°C. En esta altitud es donde se observan los eventos de calentamiento repentinos, también llamados SSW por sus siglas en inglés sudden stratospheric warming, descriptos con más detalle luego. Estos eventos se caracterizan por un incremento considerable de la temperatura en un período de sólo algunos días. Cerca de los 50 km de altura, la temperatura alcanza su máximo, influenciada por la disminución de densidad en este punto. Este máximo define la estratopausa, que delimita la estratosfera de la mesosfera.

Se define la mesosfera como la capa atmosférica que se encuentra por encima de la estratopausa, y por debajo de la mesopausa, que se halla aproximadamente a los 85 km de altura. Esta capa se caracteriza por presión atmosférica baja, con un promedio de alrededor de 0.01 hPa, y por el nivel poco denso del aire comparado con las capas inferiores. Se estima que el 99 % de la masa de la atmósfera se encuentra debajo de la estratopausa, siendo el porcentaje de nitrógeno y oxígeno en la mesosfera aproximadamente el mismo que a nivel del mar. Por otro lado, la temperatura del aire en esta capa disminuye con la altura. Este fenómeno es debido, en parte, a la baja concentración de ozono en el aire, activo en la absorción de radiación solar. En consecuencia, las moléculas, especialmente las que se encuentran en la zona superior de la mesósfera, tienden a perder más energía de la que absorben, lo que resulta en un déficit de energía y un enfriamiento.

Por tanto, el aire en la mesósfera se vuelve más frío con la altura, hasta aproximadamente los 85 km. A esta altitud, la temperatura de la atmósfera alcanza su valor promedio más bajo: -90°C.

Se considera la termosfera como la capa atmosférica superior a la mesopausa, y que se extiende hasta cerca de los 500 km por encima de la superficie terrestre. Esta capa de la atmósfera se caracteriza por un nuevo aumento de la temperatura relacionada a su activa interacción con el Sol, aun cuando la densidad del aire es mínima. Aunque esta densidad baja no permite medir directamente la temperatura, observaciones con satélites permiten determinar que la temperatura aumenta y realizar algunas estimaciones. La intensa radiación solar es absorbida por las moléculas de oxígeno (O_2), de forma tal que la absorción de una pequeña cantidad de energía produce un aumento significativo de la temperatura. Este incremento puede variar diariamente. Es en la termosfera donde abundan las concentraciones de iones y electrones libres y existe actividad electromagnética, por ejemplo, las auroras.

Por encima de los 500 km de altura se debilita la acción gravitatoria sobre las moléculas de aire. La región donde los átomos y las moléculas se dispersan hacia el espacio frecuentemente se denomina exosfera, y representa el límite superior de la atmósfera.

Como se ha descripto, las capas poseen características diferentes y por esto es primordial establecer con certidumbre la región de ocurrencia del fenómeno de análisis. En este trabajo se establece como región de estudio la mesosfera y baja termosfera, que se define aproximadamente entre los 60 y 110 km de altitud, y se denota como MLT, por sus siglas en inglés: mesosphere and lower thermosphere. Aunque el análisis está enfocado en este rango, será de importancia considerar la actividad de capas más bajas para explicar la ocurrencia de los fenómenos que tienen lugar en la MLT.

2.2. Mesosfera y Baja Termosfera

En esta sección se describen con un poco más de profundidad los fenómenos que tienen lugar en la MLT. La sección referida a ondas atmosféricas introduce estas perturbaciones, en particular las ondas de gravedad y planetarias. Ambas están íntimamente ligadas a las perturbaciones de mareas y presentes en la mayoría de la bibliografía sobre el tema. En la siguiente sección se dan más detalles sobre las mareas atmosféricas específicamente. Por último, se introduce el fenómeno de calentamiento estratosférico repentino dado que los datos fueron tomados a días de haber ocurrido este fenómeno en el hemisferio sur.

La mesosfera y baja termosfera constituye la parte superior de lo que a

menudo se denomina atmósfera media (10 a 110 km). Es importante destacar que la atmósfera media se considera como la región de acoplamiento de la zonas baja y alta de la atmósfera. Principalmente, el acoplamiento de la atmósfera refiere a la propagación vertical de perturbaciones. Se considera que son algunas de estas perturbaciones las que conforme se propagan tenderán a crecer, y como resultado, transportarán energía e impulso entre regiones, afectando el comportamiento general en ciertas capas. De manera que, describir la dinámica de la región MLT implica el estudio de las perturbaciones verticales que resultan de gran influencia en esta capa en particular.

Las regiones de origen de estas ondas se encuentran en la atmósfera inferior. Por lo que, la actividad termodinámica en estas capas, troposfera y estratosfera, determinará qué tipo de ondas y con qué energía se generan. Principalmente, interesarán la absorción de radiación, el intercambio de calor influenciado por el vapor de agua y la topografía, entre otros. Las perturbaciones generadas en la zona baja se propagarán hacia las capas altas únicamente a través de un medio estable. Se sabe que determinados factores, como cierta dirección particular del viento de fondo, colaboran para que esta propagación sea efectiva. El impacto de las ondas en la MLT depende entonces, en última instancia, de la variabilidad de la fuente y los efectos variables de propagación en la atmósfera media. (Vincent et al. 2015).

Al alcanzar altitudes como las de la MLT, bajo condiciones óptimas de propagación, estas ondas ya poseen amplitudes considerables. A su vez a esta altitud, las ondas que arriban pueden generar nuevas perturbaciones de un orden similar. En conclusión, la actividad de las ondas atmosféricas se acentúa en la MLT llegando a dominar el campo de viento de esta región.

2.2.1. Ondas Atmosféricas

Las ondas atmosféricas son relevantes como componentes mayores de la circulación total. Como se mencionó antes, son el agente de transporte de energía y momento y se pueden explicar por fluctuaciones de densidad, viento o temperatura, variando en escalas espaciales y temporales. En particular, en la MLT, se reconoce la propagación vertical principalmente de tres tipos de onda: ondas planetarias, mareas y ondas de gravedad.

La diferencia entre los tipos de perturbaciones radica en la fuerza restauradora que explica cada tipo de onda. Para que en un medio se propague una perturbación como una onda es necesario que en el medio actúe una fuerza restauradora. En la atmósfera principalmente se destacan: la conservación de la temperatura potencial en presencia de estabilidad estática positiva, la cual se relaciona a la formación de ondas gravitatorias internas y ondas gravitatorias superficiales; la conservación de la vorticidad potencial en presencia del gradiente medio de vorticidad potencial, que explica el origen de

las denominadas ondas de Rossby o planetarias. Luego la excitación termal y el forzamiento gravitatorio originan las denominadas mareas atmosféricas.

Ondas Gravitatorias

Las perturbaciones que se explican a partir de la fuerza restitutiva positiva que actúa sobre parcelas de aire desplazadas del equilibrio hidrostático son conocidas como ondas gravitatorias, o GWs por su sigla en inglés, gravity waves.

Se consideran de escala media a pequeña, de manera que es posible despreciar efectos de rotación. Poseen longitudes de onda verticales de 5 a 15 km, y períodos que oscilan desde minutos hasta algunas horas.

Entre sus causas de origen se encuentran la orografía, tormentas eléctricas, inestabilidades de cizallamiento, convección, etc. Este tipo de ondas sólo pueden propagarse en una atmósfera estratificada y estable. En una atmósfera estratificada y estable una parcela de aire es capaz de oscilar adiabáticamente. Este medio contribuye a la acción de la fuerza restitutiva, la atracción gravitatoria, para que se generen las oscilaciones.

Se diferencian ondas gravitatorias internas y externas, o evanescentes. Partiendo del análisis de ondas gravitatorias acústicas, se denominan ondas internas en propagación a aquellas ondas cuyos componentes individuales oscilan en la dirección vertical. Su estructura oscilatoria se produce en el interior de un dominio acotado. Por el contrario, las ondas evanescentes o externas reciben este nombre porque se desarrollan en la vecindad del exterior de un límite. También llamadas ondas de borde, se pueden analizar de la misma forma que las ondas superficiales que se propagan en el límite de masas de aguas profundas, donde es posible definir una superficie libre (Lindzen, 1990)

En el estudio de la dinámica atmosférica de ondas, es de especial interés el desarrollo de las ondas gravitatorias internas, debido a que su estructura oscilatoria es un punto de partida para comprender los diferentes tipos de perturbaciones, la turbulencia de la alta atmósfera en general, aunque en sí mismas no representan un fenómeno dominante en la circulación troposférica de latitudes medias (Lindzen, 1990). Particularmente, resulta de interés el análisis sobre propagación. Una vez generadas, las ondas gravitatorias pueden propagarse en dirección vertical y horizontal. A estas ondas de gravedad se las llama ondas primarias. Se dice que la onda es filtrada por el viento de fondo debido a que su propagación está condicionada por el mismo. La onda se propagará si la velocidad de fase tiene sentido inverso respecto al viento zonal de fondo. Las ondas de gravedad se propagan y crecen exponencialmente amplificándose con el decrecimiento de la densidad en altura y cuando alcanzan cierta amplitud se vuelven inestables. Cuando la velocidad de fase de la onda es nula con respecto al viento de fondo se

da este rompimiento, la longitud de onda vertical tiende a cero, la onda se ve imposibilitada de propagarse hacia mayores alturas y se dispersa depositando momento y energía en las capas altas. Este evento puede producir la generación de ondas de gravedad secundarias en la estratosfera o mesosfera. En este punto la nueva perturbación generada puede tener una longitud de onda mayor que la onda primaria que la generó (Vadas et al., 2018).

Cabe destacar que alteraciones importantes en la estratosfera pueden modificar significativamente los flujos de ondas de gravedad y, por lo tanto, pueden provocar cambios significativos en la estructura térmica y dinámica de la MLT. Tales perturbaciones incluyen SSW (Vincent et al. 2015). Se han obtenido resultados que confirman que la actividad de ondas gravitatorias se incrementa antes del SSW y se debilita durante el evento (Schneider, Chau y Stober, 2016).

Ondas Planetarias

Las ondas planetarias, o PWs, por su sigla en inglés, planetary waves, como su nombre lo indica, son oscilaciones de escala global. Se caracterizan por poseer longitudes de onda horizontal de aproximadamente miles de kilómetros, pudiendo alcanzar los 10000 km. Con periodos que varían de 1 a 30 días. La escala de estas perturbaciones trae aparejado el inherente efecto de la fuerza de Coriolis. Esta última fuerza y su variación con la latitud es considerada la fuerza restitutiva de este tipo de onda. Es decir, que en el análisis de ondas planetarias no puede despreciarse el efecto de rotación, de manera que estas ondas pueden encontrarse en bibliografía como ondas rotacionales.

Se considera que las ondas planetarias se generan debido a la conservación de la vorticidad potencial por variación latitudinal de la fuerza de Coriolis. El modelo más sencillo de Ondas de Rossby para una consideración dinámica barotrópica se asocia con el carácter solenoidal de la variación de los movimientos de gran escala. El estudio de estas ondas y la conservación de la vorticidad potencial refleja un balance entre cambios en la vorticidad relativa de una parcela de aire y cambios en la vorticidad planetaria debido a desplazamientos meridionales. Las fuentes de excitación de ondas planetarias pueden atribuirse de forma general a la troposfera, explicadas principalmente por discontinuidades tierra-océano, o a nivel local desencadenadas por inestabilidades baroclínicas.

Como las ondas de gravedad, las ondas planetarias se propagan de forma horizontal y vertical y también son filtradas por el viento de fondo, de manera que estas perturbaciones pueden propagarse en la atmósfera media solo cuando la dirección del viento en la estratosfera es hacia el este.

Este tipo de ondas contribuyen al equilibrio térmico en la circulación global de masas de aire, siendo responsables de la generación de eventos

de variación abrupta de la temperatura en la estratosfera y la mesosfera, y eventos de largo período que se propagan hacia abajo y repercuten en troposfera tales como el jet polar.

2.2.2. Mareas

Las mareas atmosféricas se caracterizan como oscilaciones atmosféricas en períodos que son fracciones de un día solar o lunar.

Las fuentes de mareas, a diferencia de las ondas planetarias y de gravedad, se encuentran en excitaciones de origen térmico, forzadas principalmente por la variación diaria de la absorción de la luz solar. También se encuentran mareas de origen gravitacional, excitadas principalmente por el campo gravitacional de la Luna y, en menor medida, del Sol.

Mareas solares

Las mareas solares comprenden aquellas perturbaciones de marea cuyos períodos son armónicos de un día solar. La excitación termal solar produce las conocidas mareas diurnas y semidiurnas correspondientes a periodos de 24 y 12 horas respectivamente, y que son características notables de la región MLT.

La marea semidiurna suele ser la marea dominante, aunque este predominio depende de varios factores. Principalmente, se observa que el comportamiento de la componente dominante cambia con la latitud, más precisamente, con el ángulo de incidencia de la radiación solar, la predisposición local del medio a la absorción de radiación, la fuente de origen de la perturbación, etc. Este último punto referido a la componente dominante no está del todo entendido aún.

Se cree que el predominio semidiurno se deriva de que la radiación solar a nivel local no permanece de forma directa durante 24 horas. Esto se acentúa debido a que la propagación de la componente diurna se dificulta en determinadas latitudes y es un hecho que posee una longitud de onda vertical más corta que la de la componente semidiurna, esta última de más de 50 km. En general, se considera que el mayor porcentaje de la radiación solar es absorbida por la superficie terrestre, y apenas el 10 % de la absorción se da directamente en la atmósfera. En la troposfera, el factor determinante para la absorción es el vapor de agua. Estudios tempranos sobre las mareas solares revelan que un tercio del forzamiento de la marea semidiurna estudiada en superficie se puede explicar a partir del vapor de agua troposférico. Posteriormente, también se ha determinado que la absorción por el ozono en la atmósfera media termina de explicar los dos tercios restantes de la marea semidiurna observada. Es común así, asociar la marea semidiurna con las variaciones de ozono. Un punto de interés constituye el hecho de que por

encima de los 60 km de altura la concentración de ozono no es comparable con la de la estratosfera y asimismo la absorción, y esto indicaría que no se producen excitaciones de marea en la mesosfera. Con respecto a la marea diurna, datos por encima de los 100 km muestran que en latitudes bajas y a diferentes alturas las oscilaciones de esta naturaleza son iguales o mayores en amplitud a las componentes semidiurnas (Lindzen, 1990). Hay acuerdo respecto a que cerca del ecuador la marea diurna se propaga con mayor facilidad. Estas latitudes asimismo preservan altas concentraciones de vapor de agua. La incapacidad en la propagación vertical de esta componente radicaría en las diferencias en la longitud del día en ciertas latitudes. De cualquier forma, se estima que el 80 % del forzamiento diurno resulta en modos no favorecidos para la propagación. Aunque en la vecindad de las fuentes de estas perturbaciones su efecto no es despreciable. De las propiedades dispersivas de las ondas de gravedad internas, cuyo desarrollo se aplica al estudio de mareas, se deduce que el período largo respecto al resto de las componentes de mareas y la escala de latitud restringida de estas ondas resultan en longitudes de onda verticales relativamente cortas (30 km o menos).

Una observación no menor respecto de las componentes diurnas y semidiurnas, es que las concentraciones de vapor de agua y la circulación del ozono en la estratosfera, aunque relevantes, son dos factores de amplia variación y difíciles de predecir.

Por todo lo anterior, es claro que la tasa de radiación y su absorción lidera el desempeño de las mareas solares. Razonablemente, es esperable que la marea sea sincrónica a la actividad solar. En este sentido, se discrimina cuando las oscilaciones solares siguen el movimiento aparente del Sol. En este caso, las mareas solares se denominan mareas migratorias. Este comportamiento puede quedar de manifiesto cuando se observa la fase de las componentes. Una fase constante suele revelar la acción de una marea migratoria. Sin embargo, algunas ondas de marea solar no son sincrónicas al Sol y son, principalmente, una consecuencia de la liberación de calor latente troposférico, por lo que reciben el nombre de mareas no migratorias (Hagan y Forbes, 2002, 2003).

Mareas lunares

Además de las mareas térmicas solares, también hay mareas observadas en la atmósfera cuyo origen reside en el efecto gravitatorio de otros cuerpos celestes sobre la Tierra, esto es, concretamente, la atracción gravitacional de la Luna y el Sol. El campo gravitacional del Sol juega un papel menor en la generación de estas mareas, por lo que hablar de mareas gravitacionales en general refiere a mareas lunares.

Estas perturbaciones se generan en el movimiento aparente de la Luna

entorno a la Tierra y como resultado de su atracción gravitacional en las regiones más bajas y densas de la atmósfera (Stening y Vincent, 1989). Sin embargo, el movimiento vertical de los océanos en el límite inferior de la atmósfera también puede contribuir a la generación de mareas lunares (Stening y Jacobi, 2001).

En total, hay más de 30 modos diferentes que componen la marea lunar, pero la mayoría tiene amplitudes despreciables. De los modos que alcanzan amplitudes significativas, la marea semidiurna migratoria lunar M2 de 12.420 h es la más importante y alcanza las amplitudes más grandes (Sandford et al. 2006).

Sin embargo, algunos otros modos de marea lunar también pueden alcanzar una amplitud detectable, incluidos los modos O1 y N2 (Winch y Cunningham, 1972). Estas mareas lunares de menor amplitud están cerca del límite de lo que se puede detectar en los datos del viento mesosférico y, por lo tanto, no se consideran en los análisis.

Una diferencia fundamental entre las mareas excitadas por el campo gravitacional de la Luna y las excitadas térmicamente por el Sol, es que en el caso lunar el forzamiento puede especificarse con precisión, mientras que en el caso solar las distribuciones variables de ozono y vapor de agua producen variaciones complejas en la generación de las mareas térmicas y esto contribuye a un alto grado de variabilidad de este tipo de perturbaciones en la región MLT (Sandford et al., 2006).

Un punto de interés y estudio radica en el efecto de los eventos SSW sobre las mareas lunares. Se cree que la amplitud puede incrementarse en el hemisferio contrario al de ocurrencia del evento.

Es importante subrayar que, a partir de las mediciones de estación única terrestres, no es posible diferenciar entre mareas migratorias y no migratorias. Hay ocasiones en que dominan las componentes no migratorias. Nos referimos a las mareas semidiurnas lunares y solares como una representación de la combinación de semidiurnas migratorias y no migratorias.

Propagación de Mareas

El estudio de la dinámica de propagación de las mareas no es trivial y requiere profundizar en un desarrollo complejo. Es posible establecer algunos lineamientos de interés. El desarrollo formal y completo puede encontrarse con mayor detalle en la publicación de Lindzen y Chapman de 1970.

Para poder abordar el desarrollo es necesario realizar algunas aproximaciones que simplifican el análisis; entre las cuales podemos mencionar que las mareas deben considerarse como perturbaciones lineales de un estado medio de la atmósfera. Ademas, se ignorarán variaciones horizontales de

temperatura y presión y también principales movimientos en esta dirección. Se ignorarán también fenómenos de disipación de energía a nivel molecular y similares. Por último, la Tierra se asume como una esfera carente de topografía y se ignorarán variaciones longitudinales de la absorción radiativa.

Cabe destacar que en el estudio de las ondas atmosféricas en general se utiliza el método de las perturbaciones. Este abordaje es descripto para este tema con detalle en (Holton, 2004). Se asume un estado básico independiente del tiempo y de la longitud y una parte perturbada que representará la desviación local del estado básico y que se explicita en las expresiones del viento en sus diferentes componentes y en general en parámetros como la densidad y la presión.

Para la resolución se proponen soluciones exponenciales dependientes de la colatitud θ y la altura z de forma general:

$$f(\theta, z) = e^{(\sigma t + s\phi)} \tag{2.1}$$

Donde $\sigma = 2\pi n$, con n=1,2,... y s representa el numero zonal o el modo. Se toma s = 1 para la componente diurna, 2 para la componente semidiurna, etc. El procedimiento requiere la consideración de las coordenadas de presión en la coordenada vertical. La relación hidrostática y la ecuación de energía se componen en una expresión termodinámica. Se resuelve el sistema de ecuaciones que constituyen esta ecuación termodinámica junto con la linealización de las anteriores ecuaciones de movimiento y la ecuación de continuidad, estableciendo condiciones de borde en el movimiento. En particular se utilizan la linealización de las ecuaciones de movimiento para expresar la divergencia de la ecuación de continuidad. De esta forma, aplicando el método de separación de variables se arriba a la denominada ecuación de mareas de Laplace.

$$\frac{i\sigma}{4a^2\Omega^2} F[\Theta_n] = -\frac{i\sigma}{gh_n}\Theta_n \tag{2.2}$$

Donde,

$$F = \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\frac{\sin\theta}{f - cos\theta^2}\frac{\partial}{\partial\theta}\right) - \frac{1}{f^2 - cos\theta^2}\left(\frac{s}{f}\frac{f^2 + cos\theta^2}{f^2 - cos\theta^2}\frac{s^2}{\sin\theta^2}\right)\Phi' \tag{2.3}$$

Esta ecuación expresa una relación entre modos de mareas y las posibles profundidades equivalentes. Esta profundidad refiere más que nada al número de onda vertical o al índice de refracción de la oscilación libre. La profundidad equivalente de un modo define su número de onda vertical. Principalmente, este resultado plantea un problema de valores y funciones propias. Los autovalores quedan representados por las profundidades equivalentes h_n y las autofunciones Θ_n se conocen como funciones de Hough.

Las funciones de Hough juegan un papel importante en meteorología y oceanografía porque representan clases generales de oscilaciones incluyendo ondas gravitatorias, ondas planetarias y mixtas.

2.2.3. Calentamiento estratosférico repentino (SSW)

Uno de los procesos de acoplamiento atmosférico que pueden destacarse, son los calentamientos estratosféricos repentinos, inducidos por ondas planetarias.

Si bien estos eventos son comunes del hemisferio norte, pueden darse en el hemisferio sur, pero con mucha menor frecuencia. La frecuencia de estos eventos en el hemisferio norte está relacionada con la baja estabilidad del vórtice polar. El hemisferio sur, cuya superficie se recubre principalmente por océanos, sostiene un vórtice polar de mayor estabilidad y menor actividad de ondas planetarias, lo cual dificulta el desarrollo de los SSW.

Los calentamientos estratosféricos repentinos surgen de la interacción de las ondas planetarias con el viento de fondo. Las ondas planetarias pueden propagarse en la atmósfera media sólo cuando el viento de fondo zonal dominante resulta hacia el este, por lo cual el SSW resulta en un fenómeno de estación. Estos eventos pueden clasificarse en diferentes niveles según la magnitud de sus efectos, entre los más conocidos se encuentran los eventos SSW mayores y menores. Según la clasificación de la Organización Meteorológica Mundial (WMO) se entiende por evento SSW menor cuando se produce una inversión del gradiente de temperatura hacia el polo a 10hPa en 60° de latitud, y en caso de producirse también una inversión del viento zonal a 60° se habla de un SSW mayor.

Un evento SSW mayor puede también fraccionar o desplazar el vórtice polar. En el borde del vórtice polar es donde puede encontrarse una elevada actividad de ondas gravitatorias.

Eventos SSW en el hemisferio sur se han registrado y analizado; entre los que se destacan un evento menor en 2010 y eventos mayores en 2002 y septiembre de 2019.

2.3. Zona de estudio: Sur de Argentina y Chile

En esta sección se comenta brevemente la particularidad de la zona de estudio con respecto a la actividad de las capas bajas y su repercusión en la MLT.

En latitudes medias a altas se genera en los polos el llamado vórtice polar. Esta es una zona climatológicamente de alta presión, donde se da subsidencia importante de masas de aire frías y que rotan solidariamente

con la Tierra. La dinámica de esta región mantiene estas masas frías de aire confinadas por lo general en esta zona polar. Convive el vórtice polar al mismo tiempo con la circulación más cálida de latitudes medias. La circulación en troposfera baja, se da con vientos que son predominantemente hacia el este impulsados por la zona latitudinal subtropical y que convergen al cinturón de baja presión subpolar, pudiendo encontrarse focos de baja presión a nivel local.

En latitudes medias australes, el flujo de aire hacia el este en la baja troposfera no encuentra mayor resistencia en la topografía debido a la gran extensión de los océanos. América del sur y la región de Nueva Zelanda representan obstáculos importantes a este flujo. Cuando los vientos del oeste encuentran altos topográficos, como la Cordillera de los Andes, se genera turbulencia en el medio troposférico y nacen perturbaciones que pueden propagarse en altura, mayormente ondas de gravedad. La mesosfera sobre esta zona presenta alta actividad de ondas atmosféricas. Este fenómeno se complementa con la generación de ondas secundarias a alturas mesosféricas. De esta forma, la región MLT sobre el sur de Argentina y Chile es considerada una de las zonas dinámicamente más activa a nivel global.

3 Método de Observación e Instrumental

Como se especificó antes, la dinámica de la MLT está dominada por las perturbaciones ondulatorias, generalmente originadas en capas inferiores, que se propagan en la atmósfera de manera ascendente creciendo en amplitud. Por esta razón, estudiar la dinámica o caracterizar la climatología de esta región, sugiere principalmente estudiar estas perturbaciones. De lo explicado sobre la propagación, es claro que el viento de fondo filtra parcial, o totalmente, estas perturbaciones. La magnitud de los vientos y su dirección nos aportan información de las variaciones de estas ondas. Será de interés, entonces, en el presente análisis hacer un estudio de vientos mesosféricos. Para obtener los datos de vientos mesosféricos se hace uso de datos de radares de meteoros.

Se realiza en este capítulo una revisión de la adquisición de los datos de ecos de meteoros y el proceso de obtención de vientos mesosféricos.

En la sección 3.1 se presenta el sistema de radares utilizado como instrumento de la adquisición. En la sección 3.2 se explican brevemente algunos detalles de la medición. En la sección 3.3 se presenta información del pre-procesamiento que permite obtener datos de viento a partir de la información de meteoros y se especifica cuál es el dato a partir del cual se realiza el ajuste de viento medio y mareas. Por último, en la sección 3.4 se describen algunas características estadísticas del dato.

3.1. Instrumento de medición

En esta sección se dan detalles sobre las ventajas del instrumento de medición utilizado y una presentación de la red de estaciones de la cual forma parte el sistema de radares utilizado en este análisis y sus mejoras respecto a estudios anteriores.

En las últimas dos décadas se ha profundizado en el estudio de la dinámica a gran escala de la MLT mediante la utilización de instrumentos terrestres y satelitales. Datos provistos por estos instrumentos, a menudo son combinados con modelos de circulación global. Entre las herramientas más efectivas para estudiar parte de la MLT, se encuentran los radares especulares de

meteoros (SMRs, por sus siglas en inglés specular metor radars).

Puntualmente, estos radares se utilizan como herramienta para estudiar los vientos en la MLT, y así también, para poder hacer estudios sobre las ondas atmosféricas (e.g., Hocking, 2005; Clemesha et al., 2009; Hoffmann et al., 2010; A. Z. Liu et al., 2013; Laskar et al., 2016; Jia et al., 2018).

La metodología implementad en la mayoría de los estudios previos realizados sobre la MLT, a menudo, presentaba ambigüedades espacio-temporales. Con el objetivo de sortear esta complejidad, se apunta a trabajar combinando radares de meteoros. En este sentido, surgen las redes de radares de meteoros SIMONe.

3.1.1. SIMONe Argentina

SIMONe es el acrónimo de Spread Spectrum Interferometric Multistatic meteor radar Observing Network (Chau et al., 2019; Conte et al., 2021) y refiere a un sistema de radares terrestres de última generación que utiliza dispersión de meteoros multiestática. SIMONe Argentina es uno de los dos primeros SMR multiestáticos operativos y de tecnología de espectro ensanchado. El segundo sistema es SIMONe Perú, cuyo sitio transmisor se encuentra ubicado en el Radio Observatorio de Jicamarca (JRO). Ambos sistemas han sido desarrollados y puestos en funcionamiento por el Leibniz Institute of Atmospheric Physics (Alemania), con la colaboración de la Arctic University of Norway (Noruega) y el MIT Haystack (Estados Unidos).

Los detalles de hardware y software de ambos sistemas, es decir, SIMONe Argentina y SIMONe Perú, se pueden encontrar en (Chau et al., 2021 y Conte et al., 2021).

Características del sistema

En transmisión, la designación de Spread Spectrum refiere a que los sistemas SIMONe utilizan espectro esparcido codificado (Vierinen et al., 2016). Se aplica trasmisión continua de señal, codificada en fase con secuencias pseudoaleatorias. Esta señal se genera y transmite en cada antena de forma independiente. El sistema dispone de configuración interferométrica, y los cinco códigos transmitidos se decodifican simultáneamente en cada sitio receptor. Se utiliza para esto el método de decodificación de detección comprimida (compressed sensing) (Urco et al., 2019).

En cuanto al carácter multiestático de la red, la denominación clásica refiere a que la trasmisión y la recepción se realiza por medio de antenas ubicadas en diferentes sitios. Los sistemas multiestáticos permiten estimar nuevos parámetros como la vorticidad relativa, además de estimar parámetros clásicos (como el viento medio y flujos de momento) con mayor precisión.

El Sistema SIMONe Argentina consta de:

- Una estación transmisora constituida por cinco antenas Yagi (linealmente polarizadas) distribuidas en una configuración de pentágono. Las cinco antenas transmiten en una frecuencia de 32,55 MHz, con una potencia media de 400 W c/u. Cada transmisor utiliza formas de onda codificadas con un código binario pseudo-aleatorio diferente (Vierinen et al., 2016). Para limitar la interferencia entre estaciones, las configuraciones de los generadores de números aleatorios que producen los códigos se seleccionan cuidadosamente para minimizar las correlaciones cruzadas entre todos los códigos.
  La estación transmisora se encuentra ubicada en la localidad Tres Lagos, Santa Cruz (49.6°S, 71.443°O).
- En recepción, una estación SIMONe puede constar de una antena o más. El primer caso, es decir, sólo una antena, que corresponde a la configuración de SIMONe Argentina, se tiene una configuración MISO (Multiple Input - Simple Output). La red posee cinco sitios receptores, a distancias de entre 20 y 270 km al rededor del sitio transmisor, cada uno de ellos constituido por una antena Yagi de doble polarización. Las estaciones receptoras se nombran a continuación:

  1. El Calafate (50,353° S, 72,251° O)
  2. El Chaltén (49,338° S, 72,882° O),
  3. Estancia La Estela (49,786° S, 72,076° O)
  4. Gobernador Gregores (48,751° S, 70,256° O)
  5. Río Gallegos (51,600° S, 69,319° O).

Esta configuración permite estudios tomográficos de la dinámica del viento MLT a escalas que antes no eran posibles. Desde que inició sus operaciones, SIMONe Argentina ha sido capaz de detectar, en promedio, más de 30.000 meteoros por día. Con este volumen de detecciones, se pueden estimar de manera confiable los vientos horizontales con mayor resolución. Es habitual que su uso se enfoque al estudio de ondas, en especial mareas y ondas de gravedad, las cuales tienen un importante impacto en el carácter del viento y la estructura termal de la MLT.

3.2. Medición

En esta sección se comenta brevemente el objeto de medición y el carácter de la adquisición.

El objeto de medición es la estela del meteoro. Se denomina estela de un meteoro a la traza de plasma y polvo que deja detrás el meteoroide, cuando

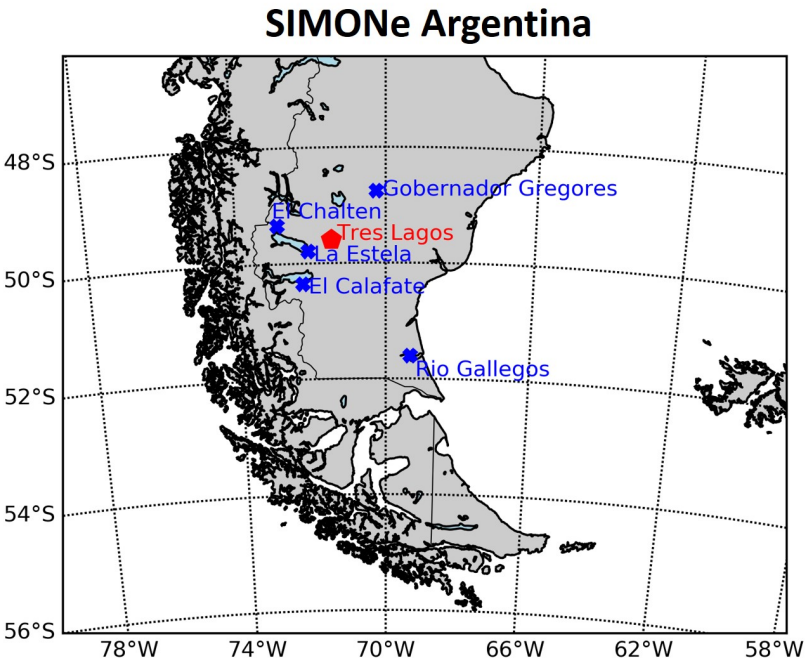

Figura 3.1: Mapa con la ubicación de los radares de la Red SIMONe Argentina. Figura original: Conte et al., 2021

éste interacciona con la atmósfera terrestre. Esta estela de gases y/o polvo imprime un rastro del orden de cientos de metros detrás del meteoroide.

Se estima el desplazamiento Doppler (f) de las estelas de meteoros debido a su deriva con los vientos neutros mesosféricos (Jones et al., 1998).

Se define un sistema centrado en el punto especular del meteoro, constituido por tres ejes independientes con direcciones: Este-Oeste para el eje x, con el sentido positivo hacia el Este; Norte-Sur para el eje y, con el sentido positivo al Norte; y perpendicular a la superficie terrestre para el eje z, con el sentido positivo hacia mayores alturas. La siguiente ecuación relaciona la proyección del vector viento en la dirección radial con el corrimiento Doppler del eco:

$$\vec{u} \cdot \vec{k} = 2\pi f + \zeta \tag{3.1}$$

donde

- $\vec{u} = (u, v, w)$ es el vector de viento neutro, con u, v y w siendo sus componentes zonal (este-oeste), meridional (norte-sur) y vertical (arriba-abajo), respectivamente.
- $\vec{k} = (k_u, k_v, k_w)$ es el vector de onda de Bragg en el sistema de coordenadas centrado en el meteoro (perpendicular a la traza del meteoro);
- f es el corrimiento Doppler;

- ζ es el error de medición del corrimiento Doppler.

Para que esta ecuación sea válida, se debe suponer que los vientos en cada intervalo de altura dado son uniformes durante el período de tiempo seleccionado (método homogéneo). Se asume también que la componente vertical del viento es despreciable con respecto a las horizontales ($w = 0$).

3.3. Pre-procesamiento

A continuación, se detalla el procesamiento aplicado para obtener información de vientos mesosféricos a partir de los datos de meteoros y se especifica cuál es el dato a partir del cual se desarrollan las dos técnicas utilizadas en esta tesis.

Se considera que el viento puede diferenciarse según la dirección en dos componentes: la componente zonal expresa el flujo de movimiento este-oeste y la componente meridional expresa variaciones norte-sur. Se considera al viento zonal y meridional como una componente de estado de base estático a la que se definirá como viento medio, más una componente de perturbación, que involucra a componentes de mareas, y que puede ser representada por una serie de Fourier truncada, donde cada componente de marea queda caracterizada mediante un período, una amplitud y una fase.

Para extraer la información del viento de las mediciones, se debe resolver la ecuación 3.1.

Se lleva a cabo una primera estimación del viento con el fin de eliminar los valores atípicos. Es decir, se resuelve la ecuación 3.1, a través de una técnica de mínimos cuadrados ponderados (WLS) utilizando bins de 1 hora y 1 km (en altura), desplazados media hora y 1 km, respectivamente. La inversa de las incertidumbres de desplazamiento Doppler al cuadrado (ζ en la ecuación 3.1) se utilizan como peso. El WLS se lleva a cabo sólo en aquellos bins que contienen un mínimo de 10 detecciones de meteoros. De este ajuste se obtienen vientos estimados $\tilde{\vec{u}}$. Utilizando esta estimación del viento y el vector de Bragg se determinan corrimientos calculados $\tilde{f}$. Luego se eliminan las velocidades radiales ($2\pi f$) cuyos valores tienen un residuo ($f - \tilde{f}$) correspondiente a más de 3 desviaciones estándar.

Después de eliminar los valores atípicos, la ecuación 3.1 se ajusta nuevamente a las mediciones de desplazamiento Doppler. Para ello, se implementa nuevamente una técnica de mínimos cuadrados ponderados (WLS) utilizando bins de 1 hora y 1 km, desplazados media hora y 1 km, respectivamente.

Se toma una hora de datos de meteoros en una ventana de altura de 1 km de ancho, para asignar un determinado valor de la componente zonal y meridional del viento (u y v, respectivamente) a un punto en una determinada altura y hora. Para el siguiente punto temporal se desplaza la

ventana en tiempo media hora. Se produce un solapamiento de los datos utilizados en el cálculo entre punto y punto.

En resumen, los datos de u y v se obtienen cada media hora y cada 1 km, con una ventana de datos de meteoros de 1h en tiempo y usando una ventana de altura de 1 km. El ajuste permite obtener los vientos horizontales, para altitudes entre 75 y 105 km.

3.3.1. Dato de viento

Se obtiene un archivo de datos que contiene vientos estimados. El período utilizado en este análisis comprende los días entre el 21 de septiembre de 2019 y el 1 de abril de 2020.

El archivo cuenta con la siguiente información para cada punto: DoY (Day of the year) refiere al día del año, altura, valor estimado de viento en componente zonal (u), error de estimación de la componente zonal (σ_u), valor estimado de viento en componente meridional (v), error de estimación de la componente meridional (σ_v).

Se cuenta con 31 alturas. En cada una de ellas, se muestrean vientos cada media hora, es decir, 48 puntos por día, a lo largo de un rango de 224 días.

3.4. Descripción y análisis de los datos de vientos

En esta sección se comentan algunos detalles de primera observación del dato de vientos mesosféricos.

Del dato de viento, se puede realizar un análisis por componente. Además, las series temporales se analizan, en esta tesis por lo general, de forma separada.

Dado que la distribución de meteoros no es regular en el rango de alturas de estudio es de interés analizar como influye esto en la distribución de datos faltantes. Si bien, el radar funciona de forma continua, se observan datos faltantes, se analiza entonces, la distribución de los mismos.

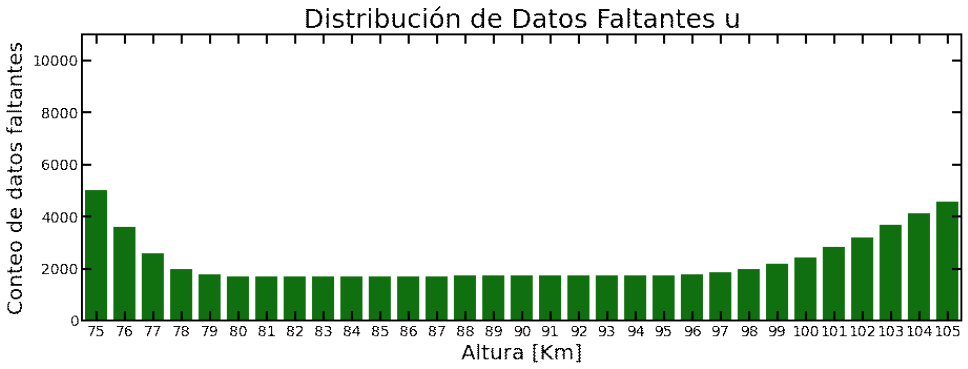

Figura 3.2: Conteo de datos faltantes en u, componente zonal, respecto al número total de muestras 10752 por serie de altura fija.

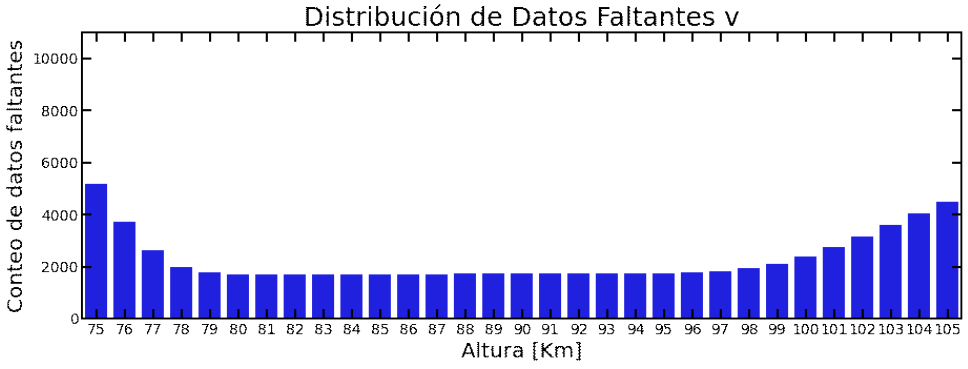

Figura 3.3: Conteo de datos faltantes en v, componente meridional, respecto al número total de muestras 10752 por serie de altura fija.

En las figuras 3.2 y 3.3 se muestra la cantidad de datos faltantes por altura para las componentes zonal y meridional respectivamente. Puede

observarse en las alturas en los extremos del rango, que la cantidad de datos faltantes crece conforme se aproxima la altura al borde. En cambio, en el centro (entre 87 Km y 97 Km, la distribución se mantiene aproximadamente constante. Para la componente v, la distribución es similar.

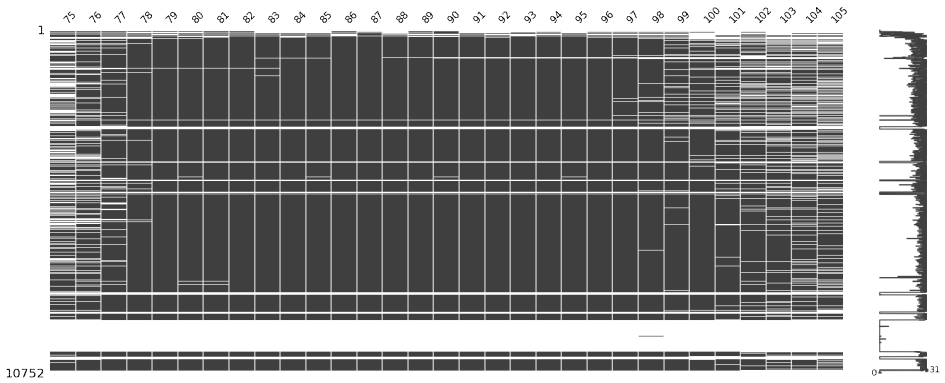

Figura 3.4: Distribución de los datos faltantes (en blanco) y datos útiles (negro), en la componente zonal (u). Las alturas se encuentran representadas en cada una de las barras, y la serie temporal transcurre de arriba hacia abajo en muestras (10752).

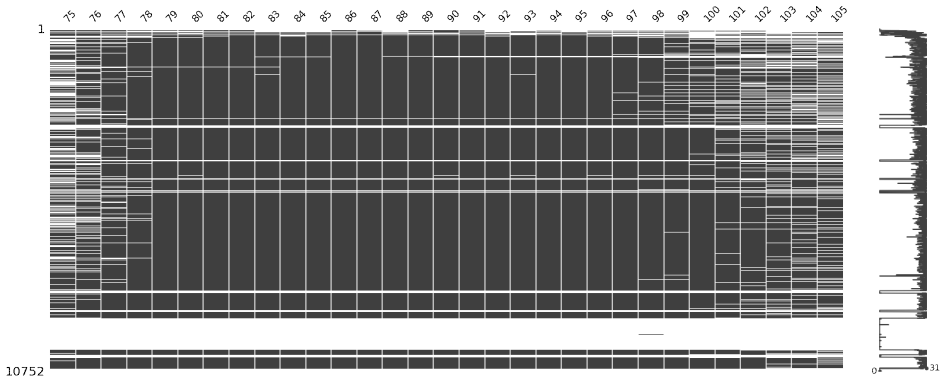

Figura 3.5: Distribución de los datos faltantes (en blanco) y datos útiles (negro), en la componente meridional (v). Las alturas se encuentran representadas en cada una de las barras, y la serie temporal transcurre de arriba hacia abajo en muestras.

En las figuras 3.4 y 3.5 puede verse la distribución de los datos faltantes (en blanco) y datos útiles (negro) en la serie temporal para todas las alturas,

para las componentes zonal y meridional respectivamente. Se subraya, que para las alturas en los límites del rango (como las alturas 75, 76, 103, 104 y 105 km), los datos faltan regularmente en todo el muestreo. No será posible, extraer una parte de la serie con mayor densidad de datos para esas alturas. En cambio, las alturas centrales presentan datos faltantes en todas las alturas por igual. En estos casos, la falta de datos se debe a que el sistema SIMONe no se encontraba operativo. En particular, se observa un intervalo importante al final de la serie donde no hay registro.

Se analizan alturas con mayor cantidad de datos (o menor cantidad de datos faltantes). Entre estas alturas se presentan los histogramas de la componente zonal y meridional del viento en la figuras 3.6 y 3.7, respectivamente.

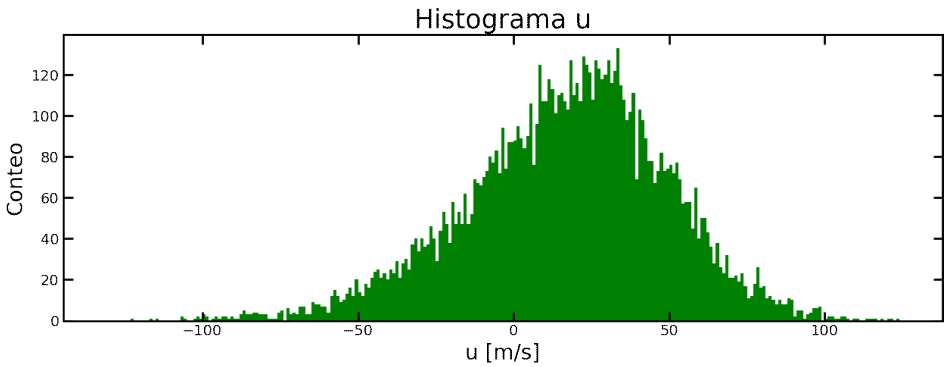

Figura 3.6: Histograma de la componente zonal de los vientos mesosféricos para la altura 90 km.

En ambas componentes se observa que la distribución de los datos no es normal. Esto era esperable, dado que los vientos no tienen un comportamiento aleatorio, sino que siguen patrones estacionales. La moda de los datos es aproximadamente 25 m/s en el caso de la componente zonal y cercanos a 0 m/s para la componente meridional. Es posible observar también que las amplitudes de los vientos oscilan entre valores de 0 y 100 m/s aproximadamente.

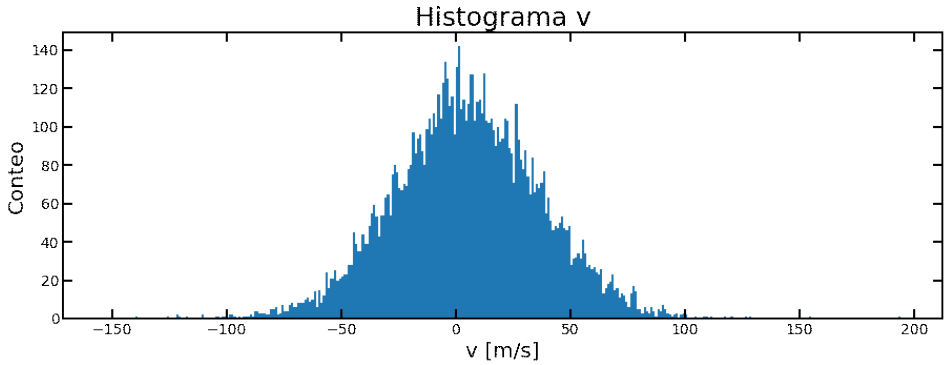

Figura 3.7: Histograma de la componente meridional de los vientos mesosféricos para la altura 90 km.

Otro gráfico de utilidad para analizar el dato es el diagrama de cajas. En él se pone de manifiesto la distribución de un determinado conjunto de datos: los límites de la caja representan los valores que toman el primer y tercer cuartil, la línea divisoria dentro de la caja indica el valor de la mediana y los bigotes aproximan la totalidad de valores que presenta la serie. Los diamantes representan valores atípicos específicos que quedan fuera del rango de los bigotes.

En la figura 3.8 puede verse diagramas de cajas por altura para cada mes, entre los meses de septiembre a diciembre (2019), y en la figura 3.9 se representan diagramas de cajas por altura para cada mes, entre los meses de enero y abril (2020). Se observa en estos diagramas la variabilidad que se hace notoria en altura en los meses de verano diciembre de 2019 a marzo de 2020), e incluso en noviembre. Esta variación en el rango de valores que toma la serie no es tan pronunciada en los meses restantes.

Por último, la figura 3.10 muestra un diagrama de cajas por mes, sobre una altura particular (altura 90 km), en ambas componentes del viento. Este diagrama pone de manifiesto que la variabilidad de un mes a otro es más visible en la componente zonal que en la meridional.

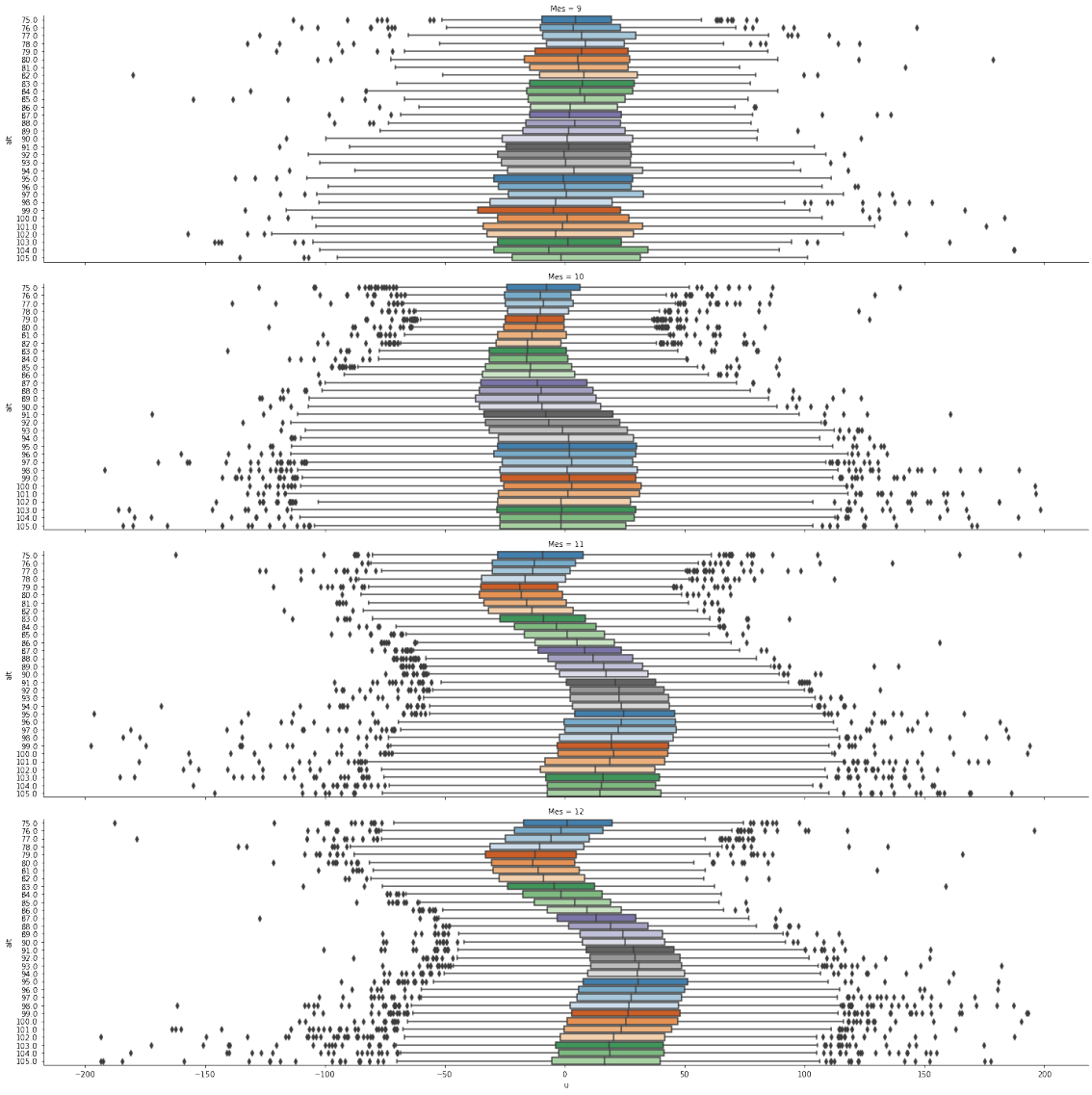

Figura 3.8: Diagrama de cajas, para la componente zonal. Filas: meses septiembre a diciembre (2019).

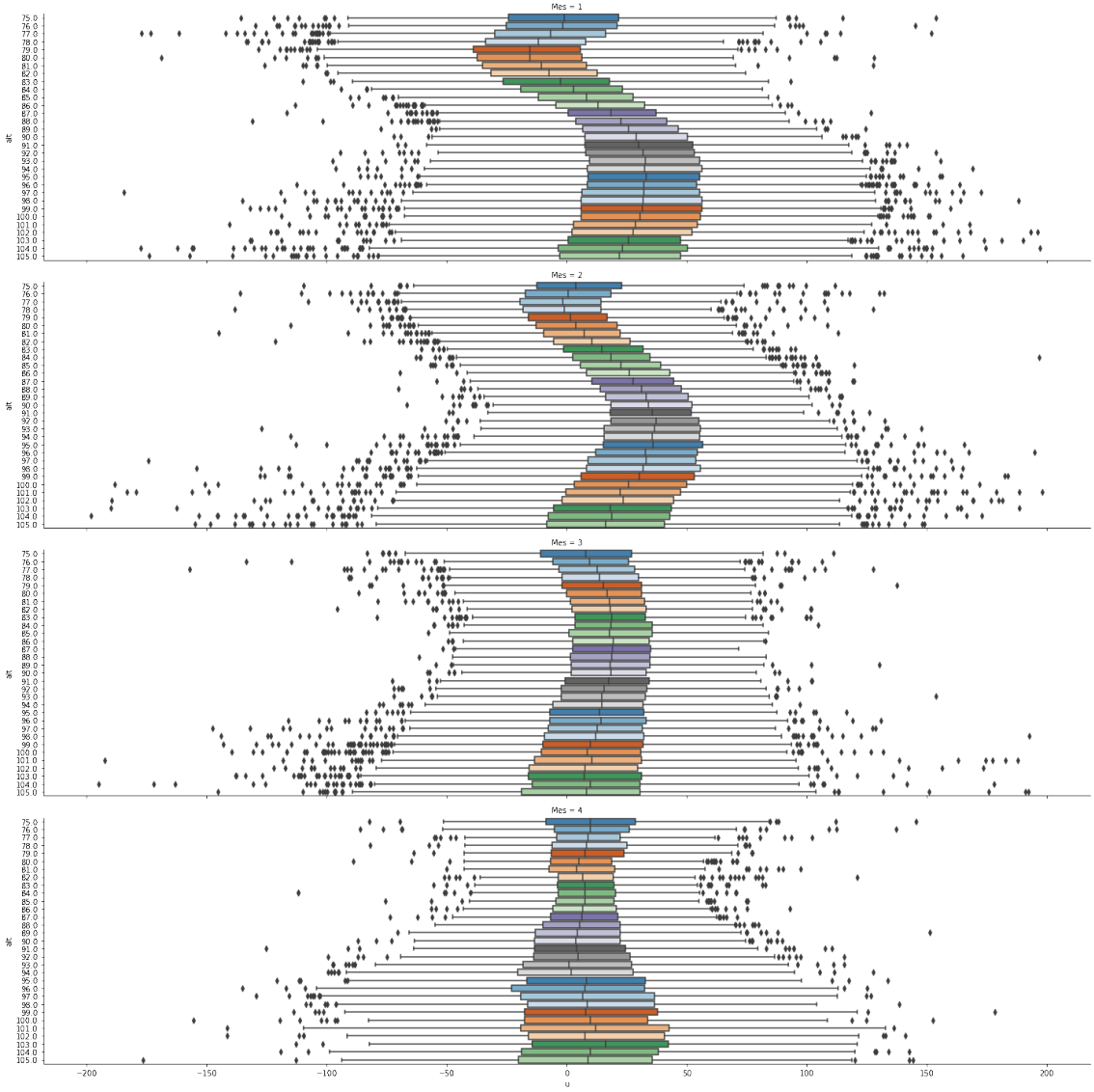

Figura 3.9: Diagrama de cajas, para la componente zonal. Filas: meses enero a abril (2020).

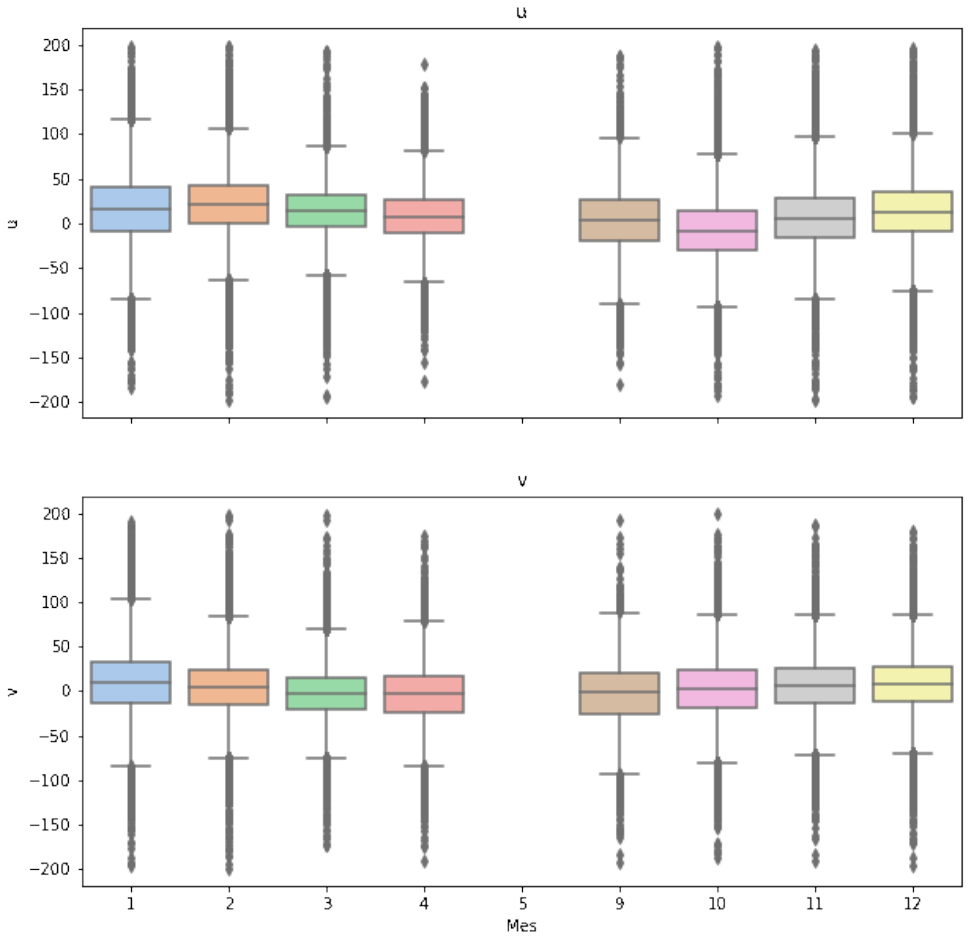

Figura 3.10: Diagrama de cajas en la altura 90 km. En el eje horizontal se representan los meses involucrados en el dato, de septiembre (9) a diciembre (12) de 2019 y enero (1) a abril (4) de 2020.

4 Técnica de Modelado Clásico: Ajuste por Mínimos Cuadrados

Una vez obtenidos los datos de vientos a partir de las observaciones de radar, se procede a modelarlos.

En este capítulo se explica cómo se aplica la técnica de ajuste por mínimos cuadrados para el conjunto de datos de viento. Los objetivos a partir de esta técnica son:

- Encontrar una estimación del viento medio en ambas componentes.
- Caracterizar las diferentes perturbaciones de mareas en ambas componentes.

A continuación, se comenta la elección de la técnica de mínimos cuadrados. Luego, en la sección 4.1 se específica cómo se aplica el problema de mínimos cuadrados en el caso de esta tesis. Finalmente, en la sección 4.2 se detallan algunas decisiones en la implementación.

Elección de la Técnica

Diversas técnicas matemáticas pueden utilizarse para modelar el viento medio y las mareas, como el análisis de mínimos cuadrados o la aplicación de wavelet a los datos de viento para extraer la información de mareas (Stening et al., 1997; Sandford et al., 2006; Conte et al., 2019; He et al., 2017).

En estudios de otros autores, se han investigado también métodos para aislar la marea lunar de las series de datos.(e.g., Stening et al., 1997b; Conte et al., 2017). La complejidad en el modelado de esta componente radica en, la baja amplitud de sus modos y en su período (T_{M2} = 12.42 hs) que se encuentra muy cercano al período de la marea solar semidiurna (T_{S2} = 12 hs), la cual posee mayor amplitud.

Stening et al. (1997b) consideran una adaptación del ajuste de mínimos cuadrados de la suma de mareas solares y lunares en series de tiempo presentadas por (Malin y Schlapp, 1980). Estos métodos son comparables con los métodos basados en el análisis de Fourier (Winch y Cunningham, 1972)].

El método de mínimos cuadrados ofrece una serie de ventajas. En particular, cada punto de datos por hora se trata por separado, por lo que los datos faltantes no son un problema. Se aplica entonces, un enfoque de

mínimos cuadrados para ajustar vientos medios y mareas. Se presenta, a continuación, la técnica de (Stening et al., 1987) basada en la adaptación de mínimos cuadrados.

4.1. Solución para el problema de vientos

Se comienza realizando un análisis por altura, es decir, la serie de tiempo utilizada para el ajuste se toma a una altura fija.

La estimación de parámetros, se realiza sobre una ventana de tiempo móvil, donde j,k denotan el número de muestras de la componente zonal y meridional, respectivamente, de forma que $j = 1, \ldots, N$ y $k = 1, \ldots, M$. Especificaciones de definición de la ventana se revisan más adelante en el texto.

En cada posición r de la ventana, se quieren estimar los parámetros que definen:

$$\vec{\eta}_r = (\eta_r^u, \eta_r^v) \tag{4.1}$$

Bajo la hipótesis de que los vientos horizontales obtenidos son el resultado de la superposición de oscilaciones en torno a un viento medio, se acepta la siguiente expresión como el modelo a ajustar, $\vec{\eta}_r$, en las componentes del viento zonal (u) y meridional (v):

$$(\eta_r^u, \eta_r^v) = (u_0, v_0)_r + \sum_{i=1}^{4} \vec{A}_i \cos\left(2\pi\left(\frac{t - \vec{\varphi}_i}{T_i}\right)\right) \tag{4.2}$$

donde $(u_0, v_0)_j$ son los vientos medios zonales y meridionales en la ventana de ajuste. Y sobre el término de perturbaciones, se identifica a:

- $\vec{A}_i = (A_{ui}, A_{vi})$ son las amplitudes zonales y meridionales para cada componente i de marea.
- t representa el tiempo (UTC), en horas;
- T_i son los períodos, donde i indica la componente de marea;
- $\vec{\varphi}_i = (\varphi_{ui}, \varphi_{vi})$ son las fases zonales y meridionales para cada componente de mareas.

Para este ajuste se utiliza como dato:

1. Las series temporales de vientos mesosféricos horizontales obtenidos a partir de los datos de radar, zonales y meridionales.
2. Los períodos de las distintas componentes de mareas.

Además, se supone que estas mareas son cosenoidales, y las fases y amplitudes constantes en la ventana de tiempo elegida. Se consideran entonces, en el desarrollo las mareas solares diurna, expresada generalmente como D1, con período T_1 = 24 horas; semidiurna solar, nombrada S2, con período T_2 =

12 horas, y terdiurna, también llamada T_3, o ST en esta tesis, con período T_3 = 8 horas, y marea lunar M2 semidiurna, con período T_M = 12,42 horas.

Las componentes zonal y meridional del viento se ajustan por separado. El modelo que explica los datos para una estimación r es:

$$\eta_r^u = (u_0)_r + \sum_{i=1}^{4} A_{ui} \cos\left(2\pi\left(\frac{t_j - \varphi_{ui}}{T_i}\right)\right) \tag{4.3}$$

para las j muestras de tiempo dentro de la ventana r.

Sistema de Ecuaciones Normales y Solución

Para cada estimación y altura (r, h) se observan dieciocho parámetros a ajustar, correspondiendo nueve a cada componente de viento. Los parámetros a estimar concretamente son (u_0, A_i, φ_i) para los componentes de marea $(i = 1, \ldots, 4)$.

La función coseno debe ser modificada para linealizar el ajuste de la fase. Para ello, se utiliza la relación:

$$\cos(a - b) = \cos(a)\cos(b) + \operatorname{sen}(a)\operatorname{sen}(b) \tag{4.4}$$

Se puede escribir, para una estimación r y para la i-ésima componente de marea:

$$\cos\left(\frac{2\pi t_j}{T_i} - \frac{2\pi\varphi_i}{T_i}\right) = \cos\left(\frac{2\pi t_j}{T_i}\right)\cos\left(\frac{2\pi\varphi_i}{T_i}\right) + \operatorname{sen}\left(\frac{2\pi t_j}{T_i}\right)\operatorname{sen}\left(\frac{2\pi\varphi_i}{T_i}\right) \tag{4.5}$$

Luego, se consideran las siguientes funciones:

$$\alpha(x, T_i) = \cos\left(\frac{2\pi x}{T_i}\right) \tag{4.6}$$

$$\beta(x, T_i) = \operatorname{sen}\left(\frac{2\pi x}{T_i}\right) \tag{4.7}$$

De manera que la función queda expresada como:

$$\cos\left(\frac{2\pi t_j}{T_i} - \frac{2\pi\varphi_i}{T_i}\right) = \alpha(t_j, T_i)\alpha(\varphi_i, T_i) + \beta(t_j, T_i)\beta(\varphi_{ui}, T_i) \tag{4.8}$$

Finalmente, la función de ajuste resulta:

$$u_j \cong (u_0)_r + \sum_{i=1}^{4} A_i[\alpha(t_j, T_i)\alpha(\vec{\varphi}_i, T_i) + \beta(t_j, T_i)\beta(\vec{\varphi}_i, T_i)] \tag{4.9}$$

En notación matricial la expresión resulta:

$$\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{pmatrix} \cong \begin{pmatrix} 1 & \alpha(t_1,T_1) & \dots & \alpha(t_1,T_3) & \beta(t_1,T_1) & \dots & \beta(t_1,T_3) \\ 1 & \alpha(t_2,T_1) & \dots & \alpha(t_2,T_3) & \beta(t_2,T_1) & \dots & \beta(t_2,T_3) \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & \alpha(t_N,T_1) & \dots & \alpha(t_N,T_3) & \beta(t_N,T_1) & \dots & \beta(t_N,T_3) \end{pmatrix} \begin{pmatrix} u_0 \\ A_1\alpha(\varphi_1,T_1) \\ A_2\alpha(\varphi_2,T_2) \\ A_M\alpha(\varphi_M,T_M) \\ A_3\alpha(\varphi_3,T_3) \\ A_1\beta(\varphi_1,T_1) \\ A_2\beta(\varphi_2,T_2) \\ A_M\beta(\varphi_M,T_M) \\ A_3\beta(\varphi_3,T_3) \end{pmatrix} \tag{4.10}$$

donde se identifican:

- La matriz de diseño:

$$A = \begin{pmatrix} 1 & \alpha(t_1,T_1) & \dots & \alpha(t_1,T_3) & \beta(t_1,T_1) & \dots & \beta(t_1,T_3) \\ 1 & \alpha(t_2,T_1) & \dots & \alpha(t_2,T_3) & \beta(t_2,T_1) & \dots & \beta(t_2,T_3) \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & \alpha(t_N,T_1) & \dots & \alpha(t_N,T_3) & \beta(t_N,T_1) & \dots & \beta(t_N,T_3) \end{pmatrix}$$

- El vector que contiene los parámetros incógnitas:

$$\tilde{\tilde{\varkappa}} = \tilde{\tilde{\varkappa}}(u_0, A_1, A_2, A_3, A_M, \varphi_1, \varphi_2, \varphi_3, \varphi_M)$$

$$\tilde{\tilde{\varkappa}} = (\tilde{\varkappa}_0, \tilde{\varkappa}_1, \tilde{\varkappa}_2, \dots, \tilde{\varkappa}_8)^T = \begin{pmatrix} u_0 \\ A_1\alpha(\varphi_1,T_1) \\ A_2\alpha(\varphi_2,T_2) \\ A_M\alpha(\varphi_M,T_M) \\ A_3\alpha(\varphi_3,T_3) \\ A_1\beta(\varphi_1,T_1) \\ A_2\beta(\varphi_2,T_2) \\ A_M\beta(\varphi_M,T_M) \\ A_3\beta(\varphi_3,T_3) \end{pmatrix}$$

Hallado $\tilde{\tilde{\varkappa}}$, se resuelve

$$\tilde{\varkappa}_0 = u_0 \tag{4.11}$$

y cada par:

$$\begin{cases} \tilde{\varkappa}_n = A_i\alpha(\varphi_i,T_i) \\ \tilde{\varkappa}_{n+4} = A_i\beta(\varphi_i,T_i) \end{cases}$$

con $n > 0$ y permite despejar los parámetros $(A_1,\varphi_1),(A_2,\varphi_2),(A_3,\varphi_3) y (A_M,\varphi_M)$.

Luego se plantea análogamente la solución para la componente meridional.

4.2. Detalles de Implementación

Luego de haberse presentado la resolución teórica de mínimos cuadrados, es necesario tomar decisiones sobre los parámetros que permiten la implementación práctica de la técnica. El análisis descripto en la sección anterior se implementa mediante un código de procesamiento. En esta sección se comentan algunas particularidades de esta aplicación

Resumen de pasos

En primer lugar se trabaja sobre una componente, luego de realizadas todas las estimaciones para la misma, se procede con la otra componente.

- Cómo ya se dijo antes se procesa altura a altura. Seleccionada una altura, se selecciona un día y se toman 21 días de datos de vientos hacia adelante.
- Se resuelve el sistema para aquellas estimaciones con número de muestras N, distintas de NaN, mayor o igual a 81 (9 parámetros). Si la cantidad de datos para un día, cumple con las condiciones establecidas, se calcula la matriz de diseño y se ajustan los parámetros.
- Estos parámetros, viento medio, amplitudes y fases, se asignan a dicho día.
- Se repite esta estimación para los 224 días del rango temporal de los datos.
- Se selecciona otra altura y se realizan nuevamente los pasos anteriores. Se procesan así las 31 alturas.

Se repiten estas estimaciones en la otra componente.

Se define que las estimaciones de los parámetros se realizan cada 24 hs, de manera que r denota el día al cual se le asigna la estimación. La ventana de datos se define tomando muestras hacia adelante a partir del punto en cuestión. La longitud de la ventana en tiempo está limitada por la suposición fuerte de invariabilidad de los parámetros y por la teoría de muestreo y necesidad de representación de las frecuencias asociadas a los períodos de marea semidiurnos.

La invariabilidad refiere a que, en vistas de realizar el ajuste para un día, se acepta como aproximación que en la ventana de tiempo necesaria, el valor de los parámetros a estimar no varía.

Por otro lado, al hablar de la resolución en frecuencia, se debe recordar que del análisis de la relación entre muestreo o discretización en tiempo y contenido en frecuencia se destacan como reglas generales:

- Según dicta el Teorema de muestreo, para que no se produzca aliasing en tiempo, la frecuencia de muestreo debe ser mayor a dos veces la

frecuencia más grande a representar. Esta frecuencia de muestreo se define como la inversa del espaciado en tiempo de las mediciones. Dicho de otra forma, el contenido de frecuencias de la señal que se puede recuperar está determinado por el muestreo en tiempo.

- Tomar una ventana finita de muestras en tiempo es similar a multiplicar al dato en tiempo por una función cajón, en frecuencia esta operación equivale a convolucionar al espectro con una función sinc. Esto viene acompañado de una pérdida de resolución inevitable, ya que siempre se tendrá una ventana finita de muestreo en tiempo. La pérdida de resolución se relaciona con el ancho del lóbulo principal de la función sinc que integra en la convolución más o menos frecuencias. El ancho de la función cajón está íntimamente ligada al ancho del lóbulo principal de la función sinc de manera inversamente proporcional. Se puede concluir que cuanto mayor es la ventana en tiempo, más angosto será el lóbulo principal de la función sinc, y mayor resolución se tiene en frecuencia.

En este análisis se analizan períodos de mareas de diurnas, semidiurnas y terdiurnas. La resolución en frecuencia que se busca esta condicionada por las mareas semidiurnas que poseen períodos muy similares. Para poder diferenciar ambas componentes para una dada estimación, es necesario que la ventana de tiempo sea lo suficientemente amplia. De manera que se analiza a continuación, la elección del ancho de la ventana.

La resolución en frecuencia viene dada por:

$$\Delta f = \frac{f_s}{N} = \frac{1}{N\Delta t} \tag{4.12}$$

Donde f_s es la frecuencia de muestreo, N es el número de muestras y Δt es el paso del tiempo de muestreo. Los datos de vientos en tiempo están tomados cada 30 min, por lo que, $\Delta t = 30$ min.

Se quiere calcular el número de muestras mínima en la ventana para generar una frecuencia máxima de muestreo igual a la diferencia entre los períodos que se quieren diferenciar. La diferencia entre estas frecuencias, calculadas como las inversas de los períodos tomados en segundos ($T_2 = 12$ hs y $T_M = 12{,}42$ hs) será:

$$\begin{aligned} f_2 - f_M &= \frac{1}{T_2} - \frac{1}{T_M} \\ \Delta f &= 7{,}82x10^{-6} \end{aligned} \tag{4.13}$$

La diferencia entre f_2 y f_M arroja en tiempo un muestreo de:

$$N = \frac{1}{\Delta f \Delta t} \approx 710 \tag{4.14}$$

Considerando que un día tiene 48 muestras, la cantidad de días necesarios en la ventana es mayor a 15 días, para poder diferenciar ambas componentes de mareas.

Por lo tanto, la decisión sobre la longitud de la ventana será tomar 21 días. Esta longitud permitirá obtener resultados comparables con otros trabajos en el tema, con procesamientos de la misma red de estaciones.

Datos faltantes en la ventana de tiempo

Dentro de los datos, hay puntos que tienen asignados como datos de vientos un valor NaN (not a number o valor faltante). Se asigna NaN al cálculo de los parámetros, cuando dentro de la ventana de ajuste, el número de datos de vientos $\vec{Y}_j$ distinto de NaN es menor al cuadrado de las incógnitas.

Para cada componente de vientos, se tiene como incógnitas: cuatro valores de amplitud, cuatro valores de fase y un valor de viento medio. Es decir, nueve incógnitas para cada componente. El número de ecuaciones en la ventana debe ser mayor a ochenta y uno (81) para que la solución sea calculada. Esta condición robustece la solución.

5 Introducción a Aprendizaje Automático

Se realiza en este capítulo una breve introducción al aprendizaje automático, sobre esta técnica se revisan conceptos y vocabulario relevante para esta tesis.

En la sección 5.1 se introduce vocabulario sobre aprendizaje automático. Debido a que es un campo en desarrollo, dicho enfoque es dinámico, es decir, se actualiza atendiendo a los distintos desafíos que se presentan. Se priorizan, a continuación, los conceptos necesarios para abordar esta tesis. Luego en la sección 5.2 se explican brevemente algunos detalles de las series temporales y sus desafíos, y cómo son abordadas por el aprendizaje automático.

5.1. Aprendizaje Automático

El desarrollo de la tecnología en las últimas décadas ha permitido mejoras continuas en la capacidad de las computadoras y, consecuentemente, ha contribuido a la producción masiva de datos. Este concepto ha adquirido el nombre de macrodatos, o Big Data por su nombre en inglés. Los macrodatos se diferencian primeramente por el volumen de la información, pero también es un hecho la diversificación de éstos, la abundancia de nuevas fuentes generan distintos tipos de datos. Además del volumen y la variabilidad, este tipo de dato se caracteriza por la velocidad necesaria de trasmisión y tratamiento de los datos. Debido también a que ha crecido la velocidad a la cual se genera la información, y la disponibilidad de la misma, un nuevo desafío consiste en poder aprovechar la gran cantidad de datos, es decir, contar con equipos y enfoques algorítmicos capaces de hacer uso de los mismos. Este proceso de crecimiento constante implica, no sólo reinventar los dispositivos de almacenamiento, sino también actualizar y mejorar técnicas, flujos de procesamiento y buscar soluciones a las nuevas problemáticas que estos cambios traen.

Se llama Ciencia de Datos, Data Science en inglés, al campo interdisciplinario que articula procesos y sistemas propios del método científico y la estadística, y otros procesos cercanos a estas ramas, con el objetivo de extraer conocimiento o información de grandes volúmenes de datos.

Para construir conocimiento sobre la información, esta se modela en una serie de etapas que comienzan con un pre-procesamiento. En él se lleva a cabo la selección de los datos, su análisis y preparación, como también la búsqueda de satisfacer aquellos requisitos propios de los algoritmos y modelos que se aplicaran sobre los datos. En el desarrollo del modelado pueden aplicarse gran cantidad de algoritmos que son clasificados en dos enfoques principales, los métodos descriptivos y los predictivos.

El primero esta relacionado a encontrar patrones en el dato y hacer un estudio descriptivo del mismo. En este caso, es común utilizar la serie completa de datos en el modelado, y al estudio se lo relaciona con la Minería de Datos, Data Mining en inglés (DM). Esta se define como el conjunto de procesos capaz de clasificar grandes conjuntos de datos para identificar estos patrones descriptivos y establecer relaciones mediante el proceso denominado ”Descubrimiento de conocimiento de Bases de Datos”, mas conocido como KDD por sus siglas en inglés: Knowledge Discovery in Database.

En un enfoque descriptivo pueden utilizarse técnicas de la minería de patrones frecuentes, esto es, el proceso de descubrir patrones en un conjunto de datos que ocurren con frecuencia sobre múltiples objetos de un conjunto. Otras técnicas pueden ser las relacionadas a la minería de relaciones, que permite establecer relaciones entre pares de objetos utilizando cualquiera de las medidas de similitud. Ejemplo de tareas usuales dentro de esta categoría son las tareas de asociación y análisis de correlaciones, etc. Cuando no se sabe a que clase pertenecen los datos, éstos se agrupan para formar clases. Este concepto esta ligado al principio de maximizar la similitud dentro de la clase y minimizar la misma entre clases. Un ejemplo de aplicación de este enfoque son las llamadas tareas de agrupamiento, también llamado Cluster Analysis. Las tareas de esta clase describen la metodología de trabajo y las técnicas utilizadas pueden desarrollarse a partir de diferentes tareas. Es posible realizar modelos descriptivos utilizando como técnica los árboles de decisión, por ejemplo, a partir de una tarea de agrupamiento, o bien a partir de la tarea de asociación.

Pero otro tipo de procesamiento se aplicará si lo que se busca es predecir el comportamiento de los datos y hacer estimaciones a futuro. En este segundo enfoque, es común encontrar todo un conjunto nuevo de problemáticas que afectarán la efectividad de los algoritmos y el proceso y en general, será necesario tener una estrategia de utilización del dato, como la separación en datos de entrenamiento y testeo, y automatización del proceso. Se utiliza Aprendizaje Automatizado, Machine Learning (ML) en inglés, para encontrar el modelo mejor a partir del conjunto de datos en este sentido. Los algoritmos de ML, en general, mejoran su rendimiento de forma adaptativa a medida que aumenta el número de muestras disponibles para el aprendizaje. Es por esto que el disponer de grandes bases de datos

en muchas disciplinas ha impulsado el desarrollo de técnicas ML capaces de aprovechar este volumen de información.

Dentro de los métodos predictivos, se destacan las tareas de regresión y clasificación. El objetivo básico de los métodos de aprendizaje predictivo es aprender un mapeo de las características de entrada (también llamadas variables independientes) a las variables de salida (también llamadas variables dependientes) usando un conjunto de entrenamiento representativo.

Por otro lado, también ha evolucionado la forma en la que el usuario interacciona con las computadoras y el grado de participación de éste, en las decisiones referidas a los flujos de ejecución de tareas. Si bien en el Aprendizaje Automático no se ha logrado aún independencia completa del usuario, existe un grupo de algoritmos que requieren menos intervención y por esto se define otra clasificación (Han, Kamber y Pei, 2012) donde se definen enfoques supervisado, inforce y no supervisado.

El aprendizaje supervisado consiste en flujos de trabajo donde se entrena un modelo con datos de entrada y salida, conocidos, de manera que el modelo pueda predecir resultados futuros, formulando las salidas para determinadas entradas. Implica dos fases importantes: La construcción de un modelo inicial, donde hay que describir las clases de clasificación existentes. El conjunto de datos denominado de entrenamiento o Train en inglés, es el ejemplo de clasificación sobre el cual se construye el modelo de representación, que puede predecir futuros datos. Luego se pone a prueba la calidad del modelo en un conjunto reservado del dato, no utilizado en el entrenamiento, denominado datos de prueba o Test. Sobre este conjunto se realiza una predicción y se contrasta la salida de predicción con el dato real reservado para prueba, con el fin de evaluar la precisión del modelo hallado en la etapa anterior. La precisión se relaciona con la cantidad de datos que son correctamente predichos o clasificados por el modelo. Esta metodología es básicamente un sinónimo de clasificación o aprendizaje regresivo. Al proporcionar la etiqueta de clase en el entrenamiento, el aprendizaje del clasificador se ”supervisa” en el sentido de que se indica a qué clase pertenece cada tupla de entrenamiento. Algunas de las técnicas mas relevantes de aprendizaje supervisado son:

- Técnicas regresivas: es una de las herramientas más básicas en el aprendizaje automático. Al usar la regresión, se ajusta una función a los datos disponibles y se intenta predecir el resultado para muestras futuras. Este ajuste de función tiene dos propósitos generales: Poder estimar los datos faltantes dentro del rango de datos de la muestra (interpolación), o poder estimar datos futuros fuera del rango de datos (extrapolación).
- Árboles de decisión: abarca tanto el enfoque de clasificación como el de regresión, de hecho, los algoritmos de árboles de decisión se

denominan CART o árboles de clasificación y regresión. La importancia de la característica es clara y las relaciones se pueden ver fácilmente. En este procedimiento se consideran todas las características y se prueban diferentes puntos de división utilizando una función de costo. Se selecciona la división con el mejor costo.

- Redes neuronales: están diseñadas para imitar el proceso de clasificación de la información que realiza el cerebro humano, caracterizándose por capas de información y pesos, que realizan un control sobre la muestra. Estas son el enfoque algorítmico más destacado de un subcampo de ML llamado Aprendizaje Profundo (DL, por su nombre en inglés: Deep Learning).

La anterior metodología contrasta con el aprendizaje no supervisado, íntimamente ligado a la agrupación, en el que no se conoce la etiqueta de clase de cada tupla de entrenamiento, y es posible que no se conozca de antemano el número o conjunto de clases que se aprenderán.

En la presente tesis se aplica el aprendizaje supervisado haciendo uso de la Minería sobre datos espacio-temporales con las técnicas de regresión sobre series temporales.

Por último, es conveniente comentar que en una realización ideal donde la intervención del usuario se redujera al máximo, la computadora utilizaría un conjunto de datos de inicio y todo el aprendizaje de la misma se daría a partir de la nueva información recolectada y derivada del análisis de datos que realizaría y actualizaría de manera automática. Esta automatización involucra tanto el procesamiento, como las decisiones sobre la estructura del mismo y la búsqueda de los mejores hiperparámetros. El objetivo de este enfoque es que las computadoras puedan procesar la información como humanos y animales lo hacen: a partir de la experiencia. Esto significaría que se desarrollen procesamientos de datos similares al análisis de la mente humana, para poder imitar los razonamientos humanos y también resolver problemas de una manera más rápida que éste. La disciplina que estudia estos procesos y desarrollos es llamada Inteligencia Artificial y para arribar a ese objetivo, el trabajo principal consiste en analizar la forma en que los ordenadores reciben la información e interactúan con el usuario.

5.2. Series Temporales

En gran parte de las disciplinas científicas, las muestras utilizadas son referenciadas a datos de tiempo o espacio, o ambos. Es decir, las observaciones científicas son inherentemente de naturaleza espacio-temporal.

El estudio de las series espacio-temporales posee cualidades especiales que presentan problemas para las técnicas habituales de minería de datos.

Aquí se describirá cuales son los enfoques para tratar la información espacio temporal y cómo se definen los objetos o instancias.

Hay una variedad de tipos de datos espacio-temporales (abreviado ST) que difieren en la forma en que los parámetros de espacio y tiempo se utilizan en el proceso de recopilación y representación de datos. Los diferentes tipos de datos conducen a diferentes categorías de formulaciones de problemas de Minería de datos espacio-temporales (STMD). Por esta razón, se presentan las diferencias entre cuatro tipos principales de datos ST disponibles: datos de eventos, trayectorias, referencia puntual y ráster.

Se puede caracterizar los mismos de forma simplificada (Alturi et al., 2018) como datos de eventos, si se consideran eventos discretos que ocurren en ubicaciones y tiempos puntuales, o como datos de trayectoria si se miden las trayectorias de cuerpos en movimiento. También, pueden clasificarse como datos de referencia puntual a las mediciones de un campo ST continuo en sitios de referencia ST en movimiento (por ejemplo, mediciones de la temperatura de la superficie recolectada usando globos meteorológicos). Por último se clasifican como datos ráster a las observaciones de un campo ST que se recopilan en células fijas en una cuadrícula o grilla ST. Mientras que los dos primeros tipos de dato (eventos y trayectorias) registran observaciones de eventos y objetos discretos, los siguientes dos tipos de datos (referencia de puntos y rásters) capturan información de campos ST que pueden ser continuos o discretos.

Si un conjunto de datos ST se recopila en un tipo de datos que es diferente del que pretendemos usar, en algunos casos, es posible convertir de un tipo a otro.

En particular, en esta tesis se considera que los datos de meteoros son datos de trayectoria. Realizado el procesamiento y, obtenidos los datos de vientos, se re-categorizan los datos como datos ráster.

Por otro lado, se define una instancia de datos como la unidad básica en la que opera un algoritmo de minería de datos (Alturi et al., 2018). En la configuración clásica de DM, una instancia se representa como un conjunto de características, pueden ser reconocibles por etiquetas supervisadas. Los datos de vientos contienen dos características, dadas por las componentes del viento zonal y meridional.

En el contexto de los datos ST, existen múltiples formas de definir instancias para un tipo de datos dado, cada una de las cuales resulta en una formulación STDM diferente. En la figura 5.1 se observan cinco categorías comunes de instancias ST que se encuentran en problemas STDM, a saber (Alturi et al., 2018): puntos, trayectorias, series de tiempo, mapas espaciales y rásters ST.

Estas instancias de datos forman los bloques de análisis para una amplia gama de problemas y métodos en STDM. Los eventos ST pueden representarse naturalmente como "instancias puntuales". Los datos de trayectoria específicos pueden representarse como una colección de instancias puntuales (lista ordenada de ubicaciones de un objeto en movimiento), una "instancia de trayectoria" o como una "serie temporal" de identificadores espaciales (por ejemplo, coordenadas de ubicación). Los datos de referencia ST comprenden "instancias puntuales", donde cada instancia es un punto de referencia del campo ST en espacio y tiempo. Por último, poniendo especial atención en datos ráster, hay tres formas diferentes de construir instancias para este tipo de datos ST. Primero, podemos definir una instancia como el conjunto de mediciones en cualquier ubicación, si se hace representada como una serie de tiempo, Ti. En segundo lugar, también podemos definir una instancia como el conjunto de medidas en cualquier marca de tiempo, tj, representada como un mapa espacial, Sj. Un tercer enfoque es considerar todo el ráster ST (colección de observaciones sobre toda la cuadrícula ST) como una única instancia de datos.

Es la diversidad de formas para definir instancias lo que resulta en múltiples problemas y métodos. La elección del enfoque correcto para construir instancias de ST depende de la naturaleza de la pregunta que se investiga y de los datos y la familia de métodos de STDM disponibles.

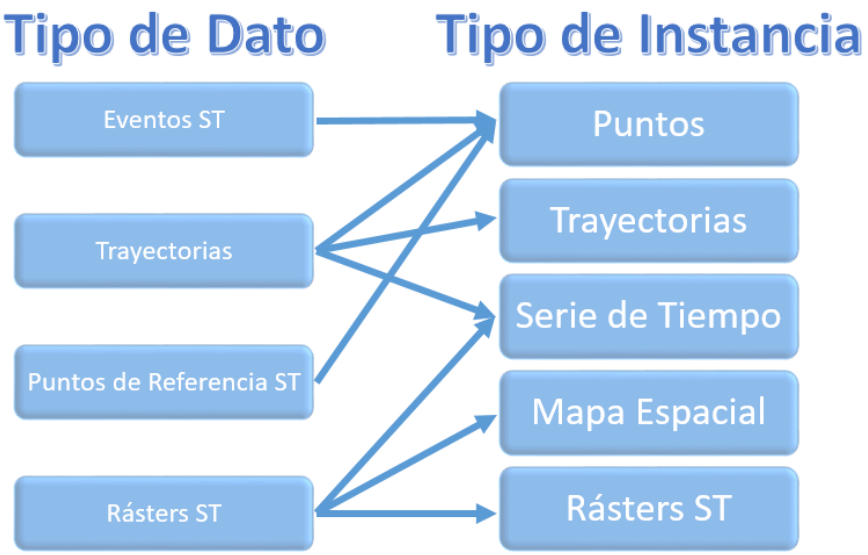

Figura 5.1: Relación entre tipos de datos e instancias en minería de datos espacio-temporales. Fuente original: Alturi et al., 2018

La información de vientos utilizada en esta tesis, se caracteriza como

instancia de tipo ráster, sin embargo se construyen como instancias de series temporales cuando se realiza un tratamiento por altura.

5.2.1. Propiedades de las series ST

Existen dos grandes objetivos que han impulsado el estudio de las series temporales: identificar la naturaleza del sistema que genera la serie de datos, y predecir los valores futuros que pueda tomar dicha serie.

Para desarrollar el primero de estos objetivos, surge la necesidad de identificar las diferentes secuencias que componen la variación del dato. Una serie temporal normalmente se descompone en cuatro elementos, componentes o movimientos principales, a saber (Orallo, 2004): componentes de tendencia, cíclicos, estacionales e irregulares.

Los movimientos de tendencias, que con frecuencia son también llamados tendencias seculares, indican el comportamiento general de la serie en un período largo de tiempo. Ayuda a identificar cuál es el comportamiento que sigue o ha seguido la serie despojada de efectos de corto período. Se llama variaciones cíclicas a movimientos que representan ciclos en las series. Estas variaciones cíclicas pueden o no ser periódicas. Es decir, los ciclos pueden no ser completamente iguales después de períodos de tiempos idénticos. Los movimientos irregulares representan el comportamiento de la serie debido a eventos aleatorios o semi-aleatorios. Por último, los movimientos estacionales son aquellos que se deben a eventos que ocurren con una frecuencia establecida y constante. Específicamente, un movimiento estacional se define formalmente como una dependencia de correlación de orden k entre cada uno de los i-ésimos elementos de la serie y el elemento k-ésimo posterior. Esta dependencia en general se mide mediante la autocorrelación (correlación entre dos valores desplazados de las mismas variables). El valor k se denomina usualmente como lag (retraso). Un movimiento estacional se aprecia visualmente como un patrón que se repite cada k elementos.

El análisis de series temporales a menudo es llevado a cabo mediante la descomposición de estas series en sus cuatro movimientos básicos.

En lo referente al segundo objetivo, predecir los valores futuros, una primera observación es que las series ST siguen por lo general un patrón regular a largo plazo. Por lo cual, la predicción de series de tiempo utilizando una técnica adecuada puede resultar, de hecho en estimaciones muy cercanas a los valores reales.

Para diseñar un modelo adecuado para pronósticos futuros, se espera que la serie de tiempo subyacente en los datos sea estacionaria. Algunos test y pruebas matemáticas como la de Dickey y Fuller se utilizan generalmente

para detectar la estacionariedad en los datos. Cuando esta serie no fuere estacionaria, es posible realizar un tratamiento de heterogeneidad que fuerce la estabilidad estadística en la misma.

Cuanto mayor es el lapso de tiempo de las observaciones históricas, mayor es la probabilidad de que la serie de tiempo presente características no estacionarias. Esta heterogeneidad relacionada a un largo registro puede explicarse, por ejemplo, por la influencia de la tendencia en la regularidad de la serie.

Durante un período de tiempo relativamente corto, se puede modelar la serie utilizando un proceso estacionario. En tales casos, será necesario aplicar diferentes estrategias que controlen la tendencia y la estacionalidad, principales fuentes de heterogeneidad. Las diferencias y las transformaciones de potencia se utilizan para minimizar efectos de la tendencia. El modelado de la componente estacional y su posterior sustracción de la serie es una opción para resolver estacionalidad.

Existen dos propiedades genéricas de los datos ST de especial importancia en relación con la predicción de series temporales: la autocorrelación y la estacionariedad (Alturi et al., 2018). Para aplicar procesos autorregresivos es de importancia hacer un estudio también de estas cualidades sobre la serie objetivo. A continuación se revisan dichos conceptos.

La llamada autocorrelación refiere a la cercanía de las mediciones espaciales y al carácter continuo que subyace en la discretización en tiempo de gran parte de las variables físicas de manera que existe una relación propia en los datos. Esta auto-correlación en los conjuntos de datos ST resulta en una coherencia de las observaciones espaciales y en suavidad en las observaciones temporales. Como resultado, los algoritmos clásicos de minería de datos que suponen independencia entre las observaciones no son adecuados para las aplicaciones ST.

Otra propiedad importante es la estacionariedad. Para definirla es necesario revisar algunos conceptos estadísticos.

Se llama proceso estocástico a la expresión matemática que describe la estructura de probabilidad de una variable aleatoria. Se define también como población al conjunto o totalidad de realizaciones de un proceso estocástico. Se considera que las series ST son procesos estocásticos, en el sentido de que la muestra es una realización particular de la población. Es posible, de esta forma, hacer inferencias sobre la población a partir de la muestra.

Además, los procesos estocásticos pueden ser caracterizados a partir de sus momentos estadísticos. Un tipo de proceso de especial importancia se da cuando todos los momentos se presentan invariantes con respecto al tiempo, es decir, el proceso es estacionario. Los modelos estacionarios asumen que el proceso permanece en equilibrio estadístico con propiedades que no

cambian con el tiempo. Se define esta propiedad como estacionariedad estricta.

Esta hipótesis puede relajarse, sin pérdida de generalidad, cuando el proceso estacionario sigue una distribución normal. Se dice, en general, que un proceso estocástico es débilmente estacionario de orden k, si los momentos estadísticos del proceso hasta ese orden dependen sólo de las diferencias de tiempo y no del tiempo de ocurrencia de los datos que se utilizan para estimar los momentos. Es decir, que un proceso estocástico estacionario de forma débil, de segundo orden, será también estrictamente estacionario cuando la distribución de probabilidad pueda caracterizarse sólo a partir de sus primeros dos momentos, como en el caso de la distribución normal. La estacionariedad débil de segundo orden queda satisfecha cuando se cumplen las siguientes propiedades:

1. Media constante
   $E(z_t) = \mu$
2. Varianza constante
   $var(z_t) = E(z_t - \mu) = \sigma^2$
3. Covarianza en el rezago k, es la covarianza entre z_t y su rezago z_{t+k} solo dependiente del nro de rezagos
   $\gamma_k = E[(z_t - \mu)(z_{t+k} - \mu)]$

Esta cualidad también es denominada simplemente "estacionariedad de segundo orden", "estacionariedad en covarianzas" o proceso estacionario "en el sentido amplio" (Gujarati y Porter, 2010). Este tipo de estacionariedad se considera suficiente para el tratamiento de series temporales, dada la compleja naturaleza de las mismas.

En esta tesis se utiliza la palabra estacionariedad para referir a la estacionariedad débil de segundo orden.

Los conjuntos de datos ST pueden mostrar también heterogeneidad, tanto en el espacio como en el tiempo. Hablar de heterogeneidad puede asociarse directamente con el carácter no estacionario de una serie ST. Resolver esta heterogeneidad resulta complejo y requiere el aprendizaje de diferentes modelos para diferentes series ST.

Los datos de la presente tesis tienen una autocorrelación evidente en su naturaleza temporal y, más precisamente, en el carácter periódico del dato, resultado de la interferencia de diferentes señales con períodos variados (que oscilan desde unas horas hasta alcanzar periodicidades mensuales). Además, las series pueden poseer implícitamente tendencia irregular, no lineal y variación abrupta de la dispersión, lo cual se debe analizar en el procesamiento.

5.2.2. Problemática de las series ST en Minería de Datos

Las metodologías propias del ámbito científico enfocadas a series ST presentan una problemática particular en minería de datos (DM) relacionadas a la autocorrelación y la heterogeneidad.

Muchos métodos tradicionales se basan en el supuesto de que las instancias de datos son estacionarias. Esta suposición básica hecha por las formulaciones clásicas de minería de datos, no siempre puede asegurarse y no es una característica usual de los datos ST. Por lo general, las series de tiempo que muestran tendencias o patrones estacionales, son de naturaleza no estacionaria.

Además, los datos ST se diferencian por la presencia típica de dependencias, como se mencionó antes. Este tipo de instancias están estructuralmente correlacionadas y esta característica limita la efectividad de los algoritmos clásicos de minería de datos. De la misma forma, el acoplamiento de la información espacial y temporal en los datos ST introduce nuevos problemas y desafíos.

Al mismo tiempo, estas dificultades se traducen en más formulaciones de análisis. Se han establecido algunos enfoques principales: El primero consiste en construir un objeto o instancia en una ubicación espacial, es decir, que las características se derivan de la muestra recolectada en una ubicación espacial conforme el tiempo varía. Otro enfoque sería considerar a los puntos en tiempo como una instancia, y que la colección de datos espaciales a un tiempo determinado construyen las características de ese punto discreto en el tiempo. Por último, una tercera formulación se puede definir al tratar un evento como un objeto, y muestras en tiempo y espacio variable construyen las características del evento.

En esta tesis se realizará adquiriendo el primer enfoque.

6 Técnica de Aprendizaje Automático en Análisis Regresivo

Una vez obtenidos los datos de vientos a partir de las observaciones de radar, y habiéndose realizado el ajuste de mínimos cuadrados, se procede a modelar los vientos con otra estrategia. En este caso se utiliza un procesamiento con aprendizaje automático acorde a la anterior técnica.

Este capítulo explica cómo se aplica la técnica de regresión en aprendizaje automático para el conjunto de datos de vientos mesosféricos. Los objetivos de esta técnica serán encontrar un modelo capaz de reproducir los datos de viento y pronosticar valores de las series temporales.

A continuación, se presenta en la sección 6.1 detalles del modelo ARIMA utilizado para este estudio, y de los modelos más simples en los cuales se basa. Luego, se realiza una descripción sobre cada uno de los pasos aplicados en el procesamiento con esta técnica, proporcionando ejemplos de los datos de vientos en la sección 6.2

Elección de la Técnica

Para un segundo enfoque, se requiere una técnica que, aplicando los procesos de ML, sea comparable con el ajuste de mínimos cuadrados. Dentro de los enfoques algorítmicos de ML, las técnicas de análisis regresivos resultan lo más apropiado para un análisis con características similares al anterior procesamiento.

De las técnicas de análisis regresivo, aplicaciones relativamente sencillas resultan de combinaciones de sistemas autorregresivos y de media móvil, para leer variaciones y generar modelos de aproximación. Por todo lo anterior, se selecciona el método autorregresivo de media móvil e integrado (ARIMA), para ajustar los vientos mesosféricos.

La popularidad del modelo ARIMA se debe principalmente a su flexibilidad para representar variadas series de tiempo con simplicidad. La severa limitación de estos modelos es que el mismo asume previamente una forma lineal para la serie de tiempo, y a ella ajusta sus parámetros (Adhikari y Agrawal, 2013). En algunas situaciones un modelo lineal no representa adecuadamente la serie, por lo cual se han definido también modelos no lineales.

El objetivo de este análisis no es realizar un estudio completo de aprendizaje automático para las series de vientos. Específicamente, el objetivo sobre la utilización de esta técnica apunta a una introducción en este tipo de procesamientos, y un primer acercamiento a los desafíos que las series temporales representan en los estudios científicos utilizando conceptos de minería de datos y aprendizaje automático.

6.1. Análisis Regresivo: ARIMA

En esta sección se desarrollan los modelos en los que se basa ARIMA y las características del mismo. La revisión teórica será necesaria para comprender detalles del flujo de procesamiento.

La técnica ARIMA representa la combinación de diferentes tipos de procesos, que se aplican en conjunción para mejorar la flexibilidad del ajuste. El nombre de la técnica hace referencia a sus componentes auto-regresiva (AR, del inglés: auto-regressive), de media móvil (MA, del inglés: moving average) e integrada (I, por Integrated). A continuación, se resume el desarrollo que permite arribar a la expresión del modelo ARIMA. Los siguientes títulos se basan en los conceptos desarrollados en Box et al.,, 2015.

Se comienza considerando muestras que se encuentran equiespaciadas y se representan como los valores de un dado proceso, a saber: $z_t, z_{t-1}, z_{t-2}, \ldots$. A su vez se definen sobre esta serie operaciones lineales como el operador de rezagos B el cual aplicado a una serie consigue transformar la muestra a valores pasados de la misma:

$$Bz_t = z_{t-1} \tag{6.1}$$

De manera que, aplicar m veces el operador resulta:

$$B^m z_t = z_{t-m} \tag{6.2}$$

Con la misma idea, puede definirse el operador de progreso.

Otro operador útil es el operador diferencia, que se define a partir del operador de rezagos como:

$$\nabla z_t = (1 - B)z_t = z_t - z_{t-1} \tag{6.3}$$

Que aplicado n veces resulta:

$$\nabla^n z_t = (1 - B)^n z_t \tag{6.4}$$

Para una serie donde los valores se encuentran altamente correlacionados, z_t puede explicarse como el resultado de aplicar un proceso lineal sobre una

colección de impulsos independientes a_t. Esto se representa como una suma pesada de la forma:

$$z_t = \mu + \psi_1 a_t + \psi_2 a_{t-2} + \dots \tag{6.5}$$

donde se aplica la función de transferencia que se define como aquel operador lineal construido a partir del operador de rezagos como:

$$\psi(B) = 1 + \psi_1 B + \psi_2 B^2 + \dots \tag{6.6}$$

La secuencia construida de esta forma no siempre resulta finita. Si la secuencia es finita (o infinita, pero absolutamente sumable) el proceso, que no es más que un filtro lineal, resulta estable y se dice que el proceso es estacionario. En este caso, se define a μ como la media sobre la cual varía la serie. De la misma forma se definen las desviaciones del valor medio de la muestra μ como: $\tilde{z_t} = z_t - \mu$.

Se establece un proceso autorregresivo puro como aquel donde la desviación de la media de la muestra actual puede explicarse como un agregado lineal, finito y escalado de valores previos del proceso con adición de un componente aleatorio:

$$\tilde{z_t} = \sum_{i=1}^{p} \phi_i \tilde{z}_{t-i} + a_t \tag{6.7}$$

$$\tilde{z_t} = \phi_1 \tilde{z}_{t-1} + \dots + \phi_p \tilde{z}_{t-p} + a_t \tag{6.8}$$

Definiendo ϕ, operador autorregresivo de orden p, en términos del operador de rezagos B:

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p \tag{6.9}$$

El modelo se simplifica:

$$\phi(B)\tilde{z_t} = a_t \tag{6.10}$$

Este modelo representa un caso especial de filtro lineal y tiene p+2 parámetros a determinar, a saber: $(\mu, \phi_1, \phi_2, \dots, \phi_p, \sigma^2)$. Los mismos se determinan a partir del conjunto de datos.

Un modelo de media móvil toma a la desviación como una combinación linealmente dependiente de un número finito de términos previos de carácter aleatorio, como sigue:

$$\tilde{z_t} = a_t - \sum_{i=1}^{q} \theta_i a_{t-i} \tag{6.11}$$

$$\tilde{z_t} = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \dots - \theta_q a_{t-q} \tag{6.12}$$

Al definir θ, como el operador de media móvil de orden q, en términos del operador de rezagos B:

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q \tag{6.13}$$

El modelo puede reescribirse como:

$$\tilde{z_t} = \theta(B) a_t \tag{6.14}$$

Este modelo tiene q+2 parámetros a determinar: $(\mu, \theta_1, \theta_2, \ldots, \theta_q, \sigma^2)$. Los cuales también se determinan a partir del conjunto de datos.

Otro modelo importante para series estacionarias resulta de la combinación de los dos modelos anteriores, ARMA, que amplia la representación de series y puede escribirse:

$$\tilde{z_t} = \sum_{i=1}^{p} \phi_i \tilde{z}_{t-i} + a_t - \sum_{i=1}^{q} \theta_i a_{t-i} \tag{6.15}$$

$$\tilde{z_t} = (\phi_1 \tilde{z}_{t-1} + \cdots + \phi_p \tilde{z}_{t-p}) + a_t + (-\theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q}) \tag{6.16}$$

Que de forma simplificada, se expresa:

$$\phi(B) \tilde{z_t} = \theta(B) a_t \tag{6.17}$$

Este modelo exige estimar $p + q + 2$ parámetros, que incluyen coeficientes, término medio y error.

Para una serie estacionaria, se espera que una representación de este tipo no demande un orden de p y q demasiado grande.

Es de utilidad escribir este modelo como un filtro lineal identificando en 6.17:

$$\tilde{z_t} = \phi(B)^{-1} \theta(B) a_t = \psi(B) a_t \tag{6.18}$$

De manera que la función de transferencia será:

$$\psi(B) = \phi(B)^{-1} \theta(B) \tag{6.19}$$

Determinadas series exhiben un comportamiento no estacionario. En particular, sus valores no varían en torno a una media, pero podrían hacerlo si a la serie se le aplica alguna transformación. Más precisamente, puede diferenciarse a fin de conseguir estacionariedad.

El comportamiento heterogéneo, no estacionario, frecuentemente se representa con un modelo en términos de un operador AR, $\varphi(B)$, en donde uno o más ceros del polinomio (una o más de las raíces de $\varphi(B) = 0$) se encuentran sobre el círculo unidad (raíz unitaria).

Cuando se tienen, específicamente, d raíces unitarias, puede expresarse mediante:

$$\varphi(B) = \phi(B)(1-B)^d \tag{6.20}$$

donde ϕ es el operador AR estacionario, definido anteriormente.

Finalmente, el modelo que expresa la manifestación de la no estacionariedad en la muestra se escribe como:

$$\varphi(B)z_t = \phi(B)(1-B)^d z_t = \theta(B)a_t \tag{6.21}$$

Y a partir del operador lineal diferencia presentado con anterioridad, la expresión de ARIMA puede encontrarse en bibliografía referida al tema como sigue:

$$\phi(B)\nabla^d z_t = \theta(B)a_t \tag{6.22}$$

Esto nos permite arribar al proceso estacionario w_t:

$$w_t = \nabla^d z_t = \phi(B)^{-1}\theta(B)a_t \tag{6.23}$$

Se estima que en general no es necesario diferenciar más de dos o tres veces.

La importancia de esta modificación del modelo, radica en que d es la cantidad de veces que es necesario tratar o diferenciar la serie para que alcance calidad de estacionaria.

Al término dado por $\nabla^d z_t$ se lo asocia entonces con el proceso w_t estacionario en este sentido: z_t representa a la serie de valores integrada o sumada a partir de w_t, esto es, afectada por el operador suma S:

$$\nabla^d z_t = w_t \tag{6.24}$$

Con $S = \nabla^{-1}$ se tiene:

$$z_t = S^d w_t \tag{6.25}$$

Luego, el modelo ARIMA se genera en la integración de un ARMA estacionario (w_t), en d veces. Este modelo se escribe de forma simplificada como $ARIMA(p,d,q)$.

Finalmente la expresión a ajustar para cada modelo ARIMA luego de la diferenciación, es la siguiente:

$$\nabla^d z_t = \mu + \sum_{i=1}^{p} \phi_i \nabla^d z_{t-i} - \sum_{i=1}^{q} \theta_i a_{t-i} + a_t \tag{6.26}$$

Como puede observarse, luego de realizar la diferenciación, se obtiene un modelo ARMA(p,q) sobre la serie diferenciada.

Los parámetros a ajustar son, por un lado, el término contante μ y el término de error a_t, los coeficientes de las respectivas series autorregresivas $(\phi_1, \phi_2, \ldots, \phi_p)$ y de media móvil $(\theta_1, \theta_2, \ldots, \theta_q)$, que a su vez determinarán el orden p y q, y el grado de diferenciación necesario.

Para el pronóstico de series de tiempo con componentes estacionales subyacentes, el modelo ARIMA rara vez será suficiente para la predicción de datos fuera de la muestra. Se aplica en ese caso, una variación bastante exitosa del modelo ARIMA, a saber, el modelo ARIMA estacional, mas comúnmente llamado SARIMA por sus siglas en inglés de Seasonal ARIMA.

6.2. Flujo de Procesamiento aplicado

En esta sección se presenta el análisis a realizar sobre una serie temporal a una altura fija, en una componente determinada, con el objetivo de introducir algunos conceptos particulares del cual se vale el procesamiento. El siguiente análisis es descripto para alturas de referencia con fines explicativos, el procesamiento se aplica a cada una de las 31 series temporales a altura fija de forma independiente.

6.2.1. Flujo de Box y Jenkins

Los estadísticos George Box y Gwilym Jenkins han desarrollado y estudiado métodos diversos de análisis de carácter regresivo sobre series temporales. En particular, para los modelos ARMA, ARIMA y variantes, desarrollan un enfoque práctico, que puede tomarse como flujo de procesamiento para identificar el mejor modelo que pueda explicar una dada serie temporal y pronosticarla. Este flujo es utilizado con regularidad en gran cantidad de estudios en el tema, y en general estos autores son de relevancia en el área.

El flujo en si mismo no asume un patrón particular en la serie temporal, es más bien, un proceso iterativo de varios pasos. Se propone primero un examen de la serie para determinar la estacionariedad de la misma. Esto se logra calculando la función de autocorrelación (FAC) y la función de autocorrelación parcial (FACP), o mediante un análisis formal de raíz unitaria. Si la serie de tiempo es no estacionaria, debe diferenciarse una o más veces para alcanzar la estacionariedad. A continuación, como primer paso del proceso iterativo, se realiza un análisis de la ACF y la PACF de la serie y se determinan los valores de p y q en el proceso que se va a ajustar. En esta etapa, el modelo ARIMA(p,d,q) seleccionado es tentativo. Como segundo paso, entonces, se estima el modelo tentativo, es decir, sus coeficientes. Por último, se examinan los residuos de este modelo tentativo para establecer si son de ruido blanco. Si lo son, el modelo tentativo es quizás una buena aproximación al proceso estocástico subyacente. Si no lo son, el proceso se inicia de nuevo. Finalizados estos pasos el modelo finalmente seleccionado sirve para pronosticar.

Como puede observarse, este proceso esta esquematizado sólo para la determinación del modelo mas apropiado, el pronóstico se realiza a posteriori.

Se aplica conceptualmente el esquema de pasos en el flujo de procesamiento de alturas para identificar el orden del modelo, aunque se utilizan los test de raíces unitarias en lugar de evaluaciones de la función de autocorrelación simple y parcial.

6.2.2. Resumen de pasos

A continuación, se describen de forma resumida los pasos del procesamiento a aplicar sobre las series temporales de alturas fijas. Seguidamente, se dan detalles de cada paso en particular.

Sobre la serie temporal se procede en dos fases o etapas bien diferenciadas: entrenamiento y testeo. Por lo cual, primeramente se realiza una separación de la muestra en entrenamiento y testeo

El objetivo de la fase de entrenamiento es hallar los parámetros que definen el modelo y comprobar si dicho modelo representa la muestra al punto de ser capaz de reproducirla dentro del rango de entrenamiento. El entrenamiento consiste entonces en aproximar la serie muestreada utilizando parte de la muestra (muestra de entrenamiento) con la cual el modelo aprende y se informa de las variaciones que este presentara y utilizando una técnica, en este caso, de contraste auto-regresiva de media móvil e integrada (ARIMA). Para llevarlo a cabo, se divide el procesamiento en pasos, descriptos a continuación.

El primer paso es un sondeo de estacionariedad. Como otras técnicas, el modelado con ARIMA presenta requisitos que es conveniente asegurar previamente. Como condición necesaria en este análisis se requiere estacionariedad de la serie temporal. Por lo cual, se realizan evaluaciones o test de raíces unitarias para examinar el estado de la serie.

Como segundo paso, se procede a buscar una serie estacionaria cuando esta no lo sea, en lo que se denomina tratamiento de heterogeneidades. Esto se logra aplicando sobre la serie las transformaciones apropiadas, es decir, se arriba a una serie equivalente a la serie estacionaria que subyace en la muestra. Una vez satisfecha esta exigencia, el procedimiento continua con otras dos grandes tareas.

El siguiente paso importante es la identificación del modelo. Este paso consiste en hallar los parámetros óptimos, éstos son: contantes y coeficientes del modelo ARIMA capaz de aproximar la serie de la muestra. (Refiere al primer y segundo paso del proceso de Box y Jenkins)

Finalmente, los últimos pasos del entrenamiento son el diagnóstico y el análisis de residuos. Se denomina diagnóstico, al procedimiento por el cual se contrasta la muestra de entrenamiento con la serie ARIMA generada.

Este contraste permite analizar la calidad del modelo dentro del rango de la muestra seleccionada para el entrenamiento. (Refiere al tercer paso del proceso de Box y Jenkins)

La segunda etapa, que sigue al entrenamiento es el testeo. El objetivo de esta fase es evaluar cuan acertado es el pronóstico generado por el modelo fuera del rango donde el modelo se entrenó. Consiste en la realización de un pronóstico y el consecuente análisis de error.

Una vez hallado el orden de complejidad en el entrenamiento, es acertado proponer el modelo ARIMA como una estimación de la realización de la serie, que puede "evaluarse" sobre un determinado rango temporal. Por consiguiente, se realiza un pronóstico sobre el rango reservado para testeo.

Luego, se realiza también, un análisis de errores sobre el rango de la muestra reservada para testeo o prueba.

6.2.3. Separación en entrenamiento y testeo

El primer paso para comenzar a procesar la información es la determinación de los intervalos de entrenamiento y testeo. La selección de esta partición debe realizarse teniendo en cuenta que la información presente en ambos intervalos no sea sustancialmente distinta.

El intervalo de entrenamiento provee al algoritmo de una muestra que debe ser representativa de la serie temporal que se quiere predecir. De lo analizado en esta tesis se notó y se recomienda para otras aplicaciones que la selección del intervalo de entrenamiento se realice cuidando:

- Que la tendencia en este intervalo sea representativa de la tendencia general de la serie, esto es, que en el intervalo de testeo no haya un cambio significativo de la tendencia.
- Que estén presentes ciclos completos de todos los períodos presentes en la serie.
- Que en este subconjunto del dato se encuentre la mayor variación de la dispersión.
- Si existe estacionalidad, este intervalo debe contener al menos un ciclo completo.

En el caso de la presente tesis, existe un intervalo de datos faltantes, un gap grande, en el mes de marzo y abril, específicamente, este ocupa los días entre el 28 de marzo y el 18 de abril de 2020. Por lo tanto, primero se acorta el rango de la serie al intervalo anterior al inicio de dicho gap grande. Luego, sobre este intervalo se realiza la partición determinando una fecha de corte. Se calcula la cantidad de muestras de ese intervalo, sin tener en cuenta las muestras de datos faltantes. Se determina así una fecha y hora de corte, correspondiente al 95 % de la cantidad de muestras, en lo que se

denomina intervalo de entrenamiento y el 5 % restante se destina a testeo. Esta fecha de corte varía para cada serie.

A continuación, se ve la partición realizada para la serie correspondiente a la altura 79 km en la Figura **??** y la altura 87 km en la figura 6.1, donde puede comprobarse que los intervalos no varían demasiado de una serie temporal a otra.

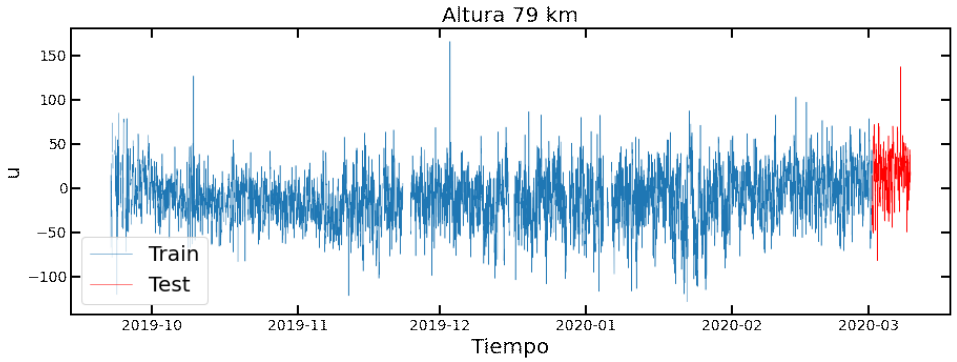

Figura 6.1: Partición en rango de entrenamiento y testeo: Altura 79 km. El corte se da en el punto correspondiente al 1 de marzo, en 20:00 hs

6.2.4. Análisis de Estacionariedad

En este paso se busca determinar si la serie temporal es estacionaria. La determinación de la estacionariedad es el inevitable problema que presentan los anteriores modelos regresivos, y este análisis debe realizarse previamente a cualquier aplicación. Con el fin de identificar heterogeneidades se presentan a continuación herramientas estadísticas, comúnmente utilizadas en la aplicación de técnicas regresivas a series temporales.

La decisión sobre la necesidad de diferenciar series de tiempo se podría basar en las características de la gráfica de z_t en su función de autocorrelación. La autocorrelación se refiere a la correlación de una serie de tiempo con sus valores pasados. Para analizar la relación entre elementos adyacentes en una muestra, se debe calcular la autocorrelación tomando como lag el valor k=1. Si se tomase k=2, se estudia la relación entre los elementos separados por un valor, y así sucesivamente.

Se llama ACF al gráfico de la Función de Autocorrelación, que mide el nivel de dependencia entre los puntos, hasta e incluyendo la unidad de retraso. Para graficarlo se debe especificar un valor de lag k máximo (k_{max}), y calcular la autocorrelación para cada valor entre 1 y k_{max}. En ACF, el coeficiente de correlación está en el eje vertical mientras que el

número de rezagos se muestra en el eje horizontal. En este diagrama no se contempla que existen dependencias entre autocorrelaciones de lags contiguos. Es decir, que si existe correlación de una muestra y sus rezagos k y k-1, inevitablemente existirá alguna relación entre estas últimas dos muestras. Este hecho si es abordado por la función de autocorrelación parcial.

La autocorrelación parcial da cuenta de la relación entre una observación y muestras anteriores en una serie de tiempo dada, con las relaciones de observaciones intermedias eliminadas. Es decir, la autocorrelación parcial en el retraso k es la correlación que resulta después de eliminar el efecto de cualquier correlación debido a los términos en los retrasos más cortos $(k-1, k-2, \dots)$. Después de trazar el gráfico ACF, es recomendable realizar gráficos de la Función de Autocorrelación Parcial (PACF).

La lectura de los correlogramas, y la decisión sobre el modelo a partir de éstos, resulta complejo cuando no se está familiarizado con este tipo de análisis. Criterios para decidir sobre los resultados de los mismos se establecen de forma general, pero no deberían tomarse como una regla inflexible a seguir. Entre estos criterios podría mencionarse:

- En general, el PACF se asocia con el componente autorregresivo. Si el gráfico PACF cae abruptamente dentro de la banda de confianza en el retraso n, entonces se debería usar un modelo AR (n)
- El término de media móvil se representa mejor en la ACF. Cuando ésta cae bruscamente después de algunos retrasos y PACF disminuye más gradualmente, se considera oportuno una componente MA coherente con los rezagos de ACF.
- Si en el correlograma simple no existe decaimiento y, por el contrario, se observan patrones periódicos que escapan de la banda de confianza, se puede sospechar de problemas de estacionalidad.

Si bien existe la posibilidad de realizar esta evaluación únicamente a partir de correlogramas, en una aplicación automatizada esta tarea resulta poco práctica.

Una alternativa suficientemente confiable son los Test de Raíces Unitarias, que permiten evaluar más a fondo utilizando pruebas formales en el operador autorregresivo del modelo. A continuación, se resumen brevemente las implicaciones de estos test, en dos de sus aplicaciones más conocidas: el Test Dicky Fuller Aumentado, abreviado ADF (Said y Dickey, 1984), y el Test Kwiatkowski–Phillips–Schmidt–Shin, abreviado KPSS (Kwiatkowski et al., 1992). El Test ADF es mucho más utilizado y en general, se encuentra como procesamiento de rutina en modelos del estilo ARIMA. El test KPSS, al igual que el ampliamente conocido Test de Phillips-Perrón (Phillips y Perron, 1988), intentan mejorar la estimación de ADF. Entre estos, el test

KPSS se acepta como el más robusto para determinar la estacionariedad de una serie.

El objetivo de la prueba de prueba Dickey Fuller es analizar la hipótesis nula (Dickey y Fuller, 1979):

- H_0: (Hipótesis Nula) La serie posee una raíz unitaria.
- H_1: (Hipótesis Alternativa) La serie es estacionaria.

De manera que, el caso favorable y de certeza se considera el rechazo de la hipótesis nula, equivalentemente, no existe raíz unitaria, y la serie es estacionaria. Esto ocurrirá si el valor p es bajo ($p < \alpha = 0{,}05$) y la estadística de prueba es menor que los valores críticos a niveles de significancia del 5 %.

Los fundamentos del test se derivan del siguiente razonamiento, se considera un proceso de raíz unitaria definido como sigue:

$$z_t = \rho z_{t-1} + b_t \tag{6.27}$$

con $-1 \leqslant \rho \leqslant 1$ y b_t un término de error de ruido blanco. Si $\rho = 1$ entonces el proceso es no estacionario, por ser ésta la expresión de definición de una caminata aleatoria, es decir, un proceso donde existe correlación hacia las muestras anteriores.

Luego, una forma útil de analizar si existe estacionariedad en una serie es hacer la regresión de z_t en un valor rezagado una muestra y comprobar si ρ es estadísticamente de valor uno, en cuyo caso el proceso es no estacionario.

En la práctica, la prueba estadística usual para comprobar si el valor es efectivamente 1 presenta problemas en el caso de raíz unitaria y se utiliza la expresión:

$$z_t - z_{t-1} = (\rho - 1) z_{t-1} + b_t \tag{6.28}$$

donde equivalentemente se deduce que:

- Si $\rho - 1 = 0$, resulta $\rho = 1$. Se tiene una raíz unitaria y se concluye entonces que z_t es no estacionaria.
- Si el término es negativo, se infiere que z_t es estacionaria. Para el caso estacionario resultaría entonces $|\rho| < 1$.

La prueba admite una constante representativa de una tendencia estocástica, la media (μ), o un término de tendencia determinista, representada con el parámetro t, o admite ambas. En cuyo caso la prueba se realiza sobre un modelo más complejo dado por:

$$z_t = \mu_1 + \mu_2 t + \rho z_{t-1} + b_t \tag{6.29}$$

La anterior prueba consideraba que b_t estaba no correlacionado. Más tarde se desarrollaría la prueba que incluía aquellos casos donde el término b_t sí está correlacionado, la cual se conoce como prueba Dickey-Fuller Aumentada (ADF). Esta prueba implica "aumentar" la ecuación 6.29 mediante la adición de los valores rezagados de la variable dependiente z_t.

Se plantea un modelo con una raíz unitaria en un proceso AR generalizado y se arriba a la condición para la serie no estacionaria.

A partir de la siguiente regresión generalizada de orden p+1:

$$z_t = \sum_{j=1}^{p+1} \varphi_j z_{t-j} + a_t \tag{6.30}$$

Que también puede escribirse como:

$$\varphi(B) z_t = a_t \tag{6.31}$$

Donde a_t es un término puramente aleatorio, y $\varphi(B) = \phi(B)(1 - B)$ tiene una raíz unitaria si $\phi(B)$ es el componente autorregresivo estacionario (sin raíz unitaria) que se expresa como:

$$\phi(B) = 1 - \sum_{j=1}^{p} \phi_j B^j \tag{6.32}$$

Reemplazando en la ecuación 6.31 :

$$\phi(B)(1 - B) z_t = a_t \tag{6.33}$$

Luego, utilizando la expresión 6.32:

$$\left(1 - \sum_{j=1}^{p} \phi_j B^j\right)(z_t - z_{t-1}) = a_t \tag{6.34}$$

Distribuyendo y reordenando términos:

$$z_t = z_{t-1} + \left(\sum_{j=1}^{p} \phi_j B^j (z_t - z_{t-1})\right) + a_t \tag{6.35}$$

Y se considera que, es equivalente probar que existe una raíz unitaria en $\varphi(B)$ a probar que se cumple $\rho = 1$ en la expresión:

$$z_t = \rho z_{t-1} + \left(\sum_{j=1}^{p} \phi_j B^j (z_t - z_{t-1})\right) + a_t \tag{6.36}$$

En la práctica, de igual forma se plantea que la condición es $\rho - 1 = 0$ para:

$$z_t - z_{t-1} = (\rho - 1)z_{t-1} + \left(\sum_{j=1}^{p} \phi_j B^j (z_t - z_{t-1}) \right) + a_t \tag{6.37}$$

donde, recordando al operador diferencia $\nabla z_t = z_t - z_{t-1} = w_t$

$$w_t = (\rho - 1)z_{t-1} + \left(\sum_{j=1}^{p} \phi_j B^j w_t \right) + a_t \tag{6.38}$$

- Si $\rho - 1 = 0$, resulta $\rho = 1$. Se tiene una raíz unitaria y se concluye entonces que z_t es no estacionaria.
- Si el término es negativo, se infiere que z_t es estacionaria. Para el caso estacionario resultaría entonces $|\rho| < 1$.

Nuevamente la prueba se puede realizar sobre un modelo más complejo dado por:

$$w_t = \mu_1 + \mu_2 t + (\rho - 1)z_{t-1} + \left(\sum_{j=1}^{p} \phi_j B^j w_t \right) + a_t \tag{6.39}$$

Fuller probó que un estadístico $\tilde{\tau}$ formado por la regresión de w_t en z_{t-1} como se describe en 6.39 se puede usar para testear raíces unitarias en el proceso AR(p+1) expresado como $\varphi(B)z_t = a_t$.

Por su parte, el test de Kwiatkowski-Phillips-Schmidt-Shin (KPSS) es otra prueba para determinar estacionariedad. Las hipótesis de esta prueba a primera vista son contrarias a las del test ADF, aunque no son estrictamente opuestas. A diferencia de otras pruebas de estacionariedad, la afirmación favorable se encuentra en la hipótesis nula.

Este test intenta comprobar la veracidad de las hipótesis estadísticas:

- H_0: (Hipótesis Nula) El proceso es de tendencia estacionaria.
- H_1: (Hipótesis Alternativa) La serie posee una raíz unitaria (La serie no es estacionaria)

En este caso, la hipótesis nula Ho se rechaza si el valor p es más bajo que el nivel de significancia o el estadístico de prueba es mayor al valor crítico para algún nivel de significancia.

De manera que, el no rechazo de la hipótesis nula será el caso favorable, equivalentemente, la serie muestra posibilidad de ser estacionaria en tendencia. Una serie temporal puede ser no estacionaria sin presentar raíces unitarias. La serie estacionaria en tendencia puede transformarse en estacionaria, su condición se relaciona con el hecho de tener una media que

varía a grandes rasgos de forma monótona (siempre creciente o siempre decreciente) que puede corregirse.

Es recomendable utilizar esta prueba como complemento de la prueba de ADF. La diferencia entre una prueba ADF y una KPSS es que ésta última no necesariamente arrojará como resultado que la serie es no estacionaria cuando esta varíe en torno a una tendencia determinista. Es decir, es capaz de evaluar la estacionariedad aún en presencia de este tipo de tendencia. Este detalle no siempre puede ser resuelto por el test ADF aunque cuente con una formulación que incluya una tendencia determinista. Lo cierto, es que el test KPSS estrictamente examina estacionariedad en tendencia.

Para determinar si las series a alturas fijas son o no estacionarias en esta tesis, se aplican los test de Dickey Fuller Aumentado y KPSS. Cada uno de ellos arroja un valor determinado de d (parámetro de ARIMA), en otras palabras, un valor aproximado de la cantidad de veces que es necesario diferenciar la serie. Los test se establecen con un nivel de confiabilidad de $= 0{,}05$ lo cual equivale a un nivel de confiabilidad del 95 %. Además, tienen como limitación un valor máximo igual a 4 para el parámetro d. Entre ambos resultados se selecciona el mayor.

En la figura 6.2 pueden observarse los resultados de los test de raíces unitarias, ADF y KPSS, se cuenta con el estadístico, el valor critico del 5 % y el p-valor. En el caso del test ADF todos los p-valores para las series allí descriptas se presentan menores al 0.05. En cambio, el valor absoluto del estadístico no resulta menor al valor crítico. En el caso del test KPSS los estadísticos resultan mayores a los valores críticos. Es posible sospechar que las series son estacionarias en tendencia, y fácilmente con una primera diferencia sería posible arribar a una serie estacionaria.

Altura	Est_adf	pValor_adf	CritV[%5]_adf	Est_kpss	pValor_kpss	CritV[%5]_kpss
79.0	-11.806842	9.053869e-22	-2.861957	0.933844	0.010000	0.463
87.0	-7.746070	1.031211e-11	-2.861954	9.025687	0.010000	0.463
93.0	-9.167700	2.435368e-15	-2.861957	9.041702	0.010000	0.463

Figura 6.2: Resultados de estadístico (Est), valor crítico 5 % (CritV[%5]) y p-valor de Test de Raíces unitarias de Dicky Fuller Aumentado (adf) y KPSS.

6.2.5. Tratamiento de Heterogeneidades

Las heterogeneidades dependen de las componentes presentes en la serie que se esté analizando. Una determinada tendencia puede resultar un

problema para una dada serie pero ser prácticamente irrelevante en otra. Así es que, no siempre es necesario realizar este paso.

Como se dijo en el anterior capítulo, las dos fuentes más grandes de heterogeneidad son la tendencia y la estacionalidad. Siempre el objetivo será disponer de una serie estacionaria equivalente a la serie objetivo.

Como se ha visto, uno de los componentes en los cuales se descompone una serie temporal es la tendencia. Las ocasiones en las que es de interés remover esta tendencia tienen lugar cuando la misma no es suave o, varía de forma no lineal. La mayoría de las metodologías implican poder identificarla primero con certeza, para luego poder removerla. Esto se puede llevar a cabo descomponiendo la serie en sus movimientos y minimizando el efecto de las otras componentes.

Probablemente, la técnica más popular para identificar tendencias de series temporales es suavizar la serie temporal mediante el cómputo de la media de un intervalo que se sustrae luego en un paso. En la práctica, esto se analiza mediante ventanas deslizantes. La longitud de la ventana determinará las frecuencias que se conservan en la variación de la serie. Una vez promediados los valores dentro de la ventana se hará evidente la tendencia general en la serie.

La tendencia puede removerse también realizando las llamadas diferencias. Una primera diferencia consiste en realizar la resta de una muestra con su rezago a lag = 1. Esto permite ir removiendo la variación del valor medio. Si aun así, no se estabiliza la media y la varianza, se vuelve a realizar otra diferencia. Es posible que con esta metodología no se obtenga una dispersión constante si la serie no la presentaba en primer lugar de forma aproximada. Existen aplicaciones para corregir también la dispersión en caso de ser necesario.

En el caso de la estacionalidad, el primer paso es determinar si en la serie existen movimientos estacionales. Esto puede bien determinarse mediante los correlogramas donde la existencia de patrones estacionales en la serie se pondrá de manifiesto en el diagrama de autocorrelación simple.

Una vez que se hace evidente una componente estacional, de período k, es posible transformar la serie para remover el efecto del movimiento estacional. Las principales metodologías que corrigen estacionalidad lo hacen diferenciando.

Aquí diferenciar refiere a realizar una diferencia entre las muestras y sus rezagos a lags iguales al período de la componente estacional. La diferenciación remueve a la vez parte de la tendencia.

Eliminar el efecto de un movimiento estacional permite que se puedan descubrir otro tipo de movimientos estacionales secundarios ocultos inicialmente. En caso de persistir la heterogeneidad, es decir, no obtenerse una

serie estacionaria luego de diferenciar, se recurre a una segunda diferenciación con otros lags de otras posibles componentes estacionales.

Para el método ARIMA en esta tesis, se toma el mayor valor del parámetro d obtenido mediante los test de raíces unitarias y se realiza una diferenciación de forma automática repetida d veces.

6.2.6. Identificación del modelo

El objetivo de esta etapa es reducir el conjunto de modelos posibles para una dada serie, a una determinada cantidad de modelos óptimos, pudiendo encontrarse uno definitivamente más conveniente. En otras palabras, el resultado ideal consiste en encontrar el conjunto de coeficientes de las series autorregresiva y de media móvil óptimo para el conjunto de datos de entrenamiento. Como consecuencia, se determinan los ordenes (p,d,q) de las series que componen el modelo. Estos parámetros se determinan a partir de las observaciones.

Un proceso estacionario puede aproximarse tanto como uno quiera con un proceso ARMA, es por esto que es posible que variadas tuplas de los parámetros de orden sean, en mayor o menor medida, adecuados para la serie temporal.

Un procedimiento usual es establecer un valor máximo de p y q y se prueban todas las combinaciones de (p,d,q): con d fijo en este caso y determinado en el paso anterior, pero variando p,q de o al valor máximo.

Para cada una de las tuplas se determinan los coeficientes de los polinomios autorregresivos y de media móvil. Luego se dispone de varios modelos estimados con coeficientes y orden. Estos modelos se someten a prueba para definir cual es el mas adecuado o el que conserva mas información sobre la serie. El indicador para determinar cual es la mejor tupla será alguno de los criterios de información.

Los criterios de información utilizan parámetros determinados en los modelos y realizan una evaluación de los mismos. Esta evaluación consiste en adicionar penalizaciones al modelo por la pérdida de grados de libertad.

Entre los criterios conocidos y ampliamente utilizados se encuentran (Box y Jenkins, 2015):

- El Criterio de Información Bayesiano de Schwarz Normalizado (BIC):

$$BIC = \frac{-2\ln\mathcal{L} + k\ln N}{N} \tag{6.40}$$

$$BIC = \ln\hat{\sigma}^2 + \frac{k}{N}\ln N \tag{6.41}$$

- Y el Criterio de Información de Akaike Normalizado (AIC):

$$AIC = \frac{-2\ln\mathcal{L} + 2k}{N} \tag{6.42}$$

$$AIC = \ln\hat{\sigma}^2 + \frac{2k}{N} \tag{6.43}$$

donde $k = p + q + 1$ representan la totalidad de parámetros del modelo ARIMA más una constante, $\mathcal{L}$ es la función de verosimilitud.

Ambos criterios se normalizan por el tamaño de la muestra N, el primer término se corresponde con $-\frac{2}{N}$ veces el logaritmo de la probabilidad maximizada, que también se expresa en función de la estimación de la varianza del residuo. El segundo término es el término de penalización por la inclusión de parámetros adicionales.

Cabe destacar que los resultados arrojados por los criterios de información para cada modelo, no son comparables entre si, debido a que penalizan a los modelos de formas diferentes. Luego, el mejor modelo se elige a partir de uno de ellos, el que se muestre más claro o diferencie mejor los modelos entre si para una misma serie temporal.

Entre los modelos para una misma serie temporal, el mejor según estos criterios es aquél que recibe menos penalizaciones, es decir, el modelo que presente el menor BIC o AIC.

En la presente tesis, se establecen como límites de los parámetros: $0 \leqslant p \leqslant 4$ y $0 \leqslant q \leqslant 4$. Para la selección del mejor modelo se utiliza el BIC. En general, se espera que los ordenes del modelo, (p,q), de cada serie temporal a altura fija no tomen valores mucho mas altos que 4 o 5 términos. A lo largo del procesamiento se ha observado que si la serie es diferenciada ($d \neq 0$), los parámetros resultan en valores más chicos que si no se diferencian.

Cuando determinado BIC, en uno de los modelos es claramente de menor valor con respecto a los demás, se elige ese modelo, sin importar el orden de las series. En este caso, observando la figura 6.3 donde se expresan los BICs correspondiente a algunas combinaciones de modelos que pueden realizarse a partir de p y q tomando valores de 0 a 4, el modelo de ARIMA (0,1,4) para la altura 79 km muestra que esta diferenciado una vez, con cuatro términos de media móvil y con intercepto, o valor constante, y es el de menor BIC. Sin embargo, el valor de BIC no varía demasiado entre un modelo y otro.

6.2.7. Validación del modelo

Habiendo transcurrido las etapas anteriores se cuenta con un modelo para cada serie de tiempo a cada altura fija, que debe ser validado.

En principio, la validación del modelo implica estudiar que tan acertado es el modelo dentro y fuera de la muestra de entrenamiento. Esto implica

```
Performing stepwise search to minimize bic
 ARIMA(2,1,2)(0,0,0)[0] intercept   : BIC=inf, Time=3.45 sec
 ARIMA(0,1,0)(0,0,0)[0] intercept   : BIC=67075.469, Time=0.32 sec
 ARIMA(1,1,0)(0,0,0)[0] intercept   : BIC=66576.466, Time=0.41 sec
 ARIMA(0,1,1)(0,0,0)[0] intercept   : BIC=66178.767, Time=0.88 sec
 ARIMA(0,1,0)(0,0,0)[0]             : BIC=67066.568, Time=0.08 sec
 ARIMA(1,1,1)(0,0,0)[0] intercept   : BIC=inf, Time=2.56 sec
 ARIMA(0,1,2)(0,0,0)[0] intercept   : BIC=65873.342, Time=1.22 sec
 ARIMA(1,1,2)(0,0,0)[0] intercept   : BIC=inf, Time=4.24 sec
 ARIMA(0,1,3)(0,0,0)[0] intercept   : BIC=65831.401, Time=2.07 sec
 ARIMA(1,1,3)(0,0,0)[0] intercept   : BIC=inf, Time=7.18 sec
 ARIMA(0,1,4)(0,0,0)[0] intercept   : BIC=65769.986, Time=2.69 sec
 ARIMA(1,1,4)(0,0,0)[0] intercept   : BIC=inf, Time=8.78 sec
 ARIMA(0,1,4)(0,0,0)[0]             : BIC=65761.084, Time=1.23 sec
 ARIMA(0,1,3)(0,0,0)[0]             : BIC=65822.499, Time=0.88 sec
 ARIMA(1,1,4)(0,0,0)[0]             : BIC=inf, Time=4.05 sec
 ARIMA(1,1,3)(0,0,0)[0]             : BIC=inf, Time=2.92 sec

Best model:  ARIMA(0,1,4)(0,0,0)[0]
Total fit time: 42.975 seconds
```

Figura 6.3: Identificación del modelo para la serie correspondiente a la altura 79 km. Los diferentes modelos ajustados con sus respectivos valores de bic y el tiempo de ejecución. El mejor modelo se muestra debajo, en este caso, ARIMA(0,1,4) con el menor valor de BIC.

comprobar que los parámetros determinados sean confiables y que el residuo del modelo ajustado respecto de la muestra conserve una distribución normal.

Al análisis realizado In-Sample, dentro de la muestra o sobre el entrenamiento se le denomina Diagnóstico. Frecuentemente, se considera validación del modelo sólo a esta mitad del análisis. Este paso implica, evaluar matemáticamente el modelo sobre la muestra de entrenamiento y verificar la aleatoriedad del residuo.

Se llama pronóstico a la prueba de la predicción que se realiza con el modelo fuera de la muestra de entrenamiento, en la muestra de testeo u Out-Sample. En este paso se realiza un análisis de error para comprobar el desempeño del modelo.

Ambas etapas son importantes para evaluar el modelo. Como regla general, si el modelo dentro de la muestra de entrenamiento arroja resultados que promueven la sospecha de no ser adecuado, se recomienda revisar los parámetros e, incluso, buscar un nuevo modelo, antes de pasar a la siguiente etapa de Pronóstico.

Diagnóstico

El objetivo de la etapa de Diagnóstico es evaluar el desempeño del modelo dentro de la muestra de entrenamiento, con el fin de hallar las deficiencias del modelo y así poder mejorarlo. Este proceso se deriva en tareas definidas. La primera de ellas se relaciona con la observación visual de la curva (y de un mapa en general para todas las series) determinada por el modelo ARIMA. En segunda instancia, se realiza un análisis sobre los coeficientes, el orden y su significancia. Por último, se realiza un análisis del residuo determinado por el modelo.

A continuación, se presenta el diagnóstico realizado sobre la componente zonal, en particular, sobre una serie, la correspondiente a la altura 79 km. Se muestra en la siguiente figura la grafica de la serie modelada y la muestra Train, para algunos de los primeros días del mes de octubre.

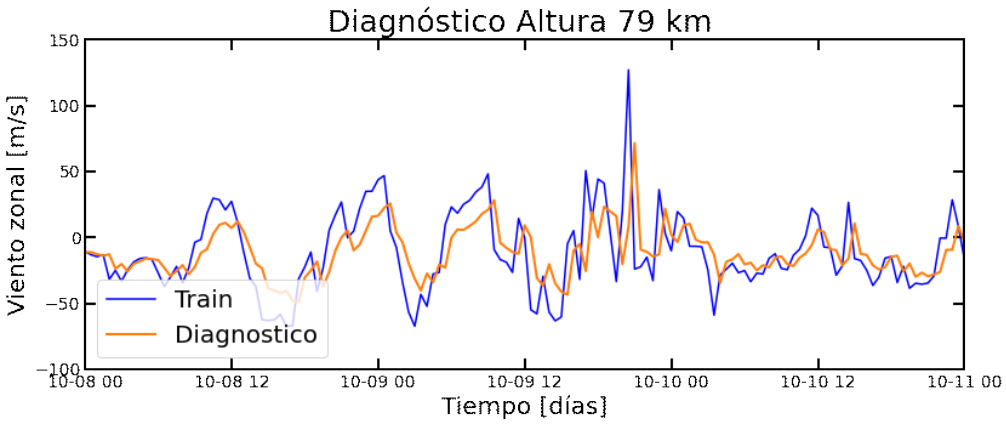

Figura 6.4: Diagnóstico para la serie correspondiente a la altura 79 km. Curva para los días 8 al 11 de octubre, donde se presenta la muestra de entrenamiento y el modelo generado por ARIMA(0,1,4) evaluado en ese rango temporal.

La observación visual arroja que, en intervalos como este, donde los datos faltantes no afectan el procesamiento, es posible encontrar un conjunto de parámetros tal que el modelo ARIMA aproxima la curva. El caso contrario se observa en la figura 6.5, para la serie de la altura 93 Km, donde los datos faltantes producen apartamientos de la curva sobre, y en las cercanías, del gap de datos faltantes.

La determinación del modelo incluye la definición de los parámetros, y con ellos al mismo tiempo el orden de ARIMA. Como ya se dijo antes, además se realiza una evaluación de significancia individual sobre estos parámetros. Estas pruebas se utilizan para determinar si la estimación

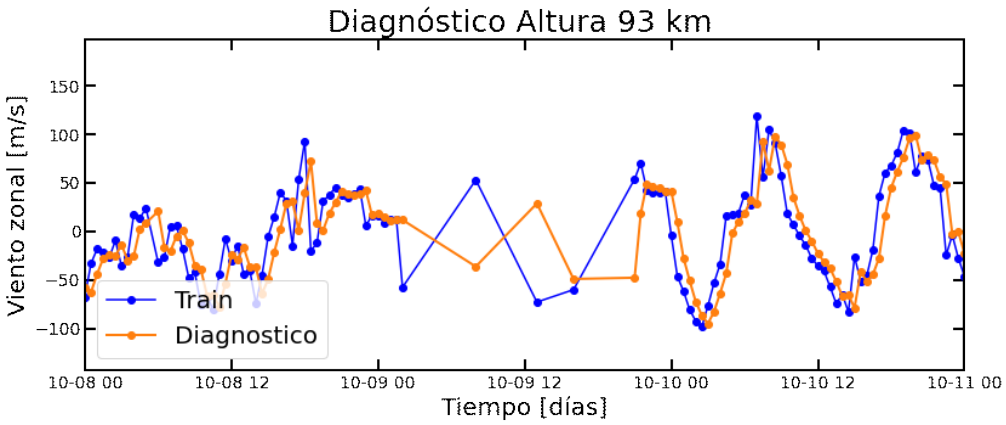

Figura 6.5: Diagnóstico para la serie correspondiente a la altura 93 km. Curva para los días 8 al 11 de octubre, donde se presenta la muestra de entrenamiento y el modelo generado por ARIMA(0,1,3) evaluado en ese rango temporal.

de un coeficiente, que acompaña una dada variable independiente, tiene importancia suficiente en la explicación de la variable dependiente. En este caso, la prueba es llevada a cabo mediante el Test de Wald (Wald, 1943). Este test determina un estadístico z y el valor de probabilidad p de z $(p > |z|)$ asumiendo un determinado nivel de significancia (α). El valor de p definido de esta forma como valor crítico para $\alpha = 0{,}05$ o del 5 %, como se asume en gran mayoría de los casos, lo que denota una confiabilidad del 95 %. Se dice que el coeficiente es significativo cuando el valor de p es menor o igual que el valor del nivel de significancia α.

En la figura 6.6 se ven los coeficientes calculados en este caso, dentro de los cuales se encuentran cuatro correspondientes a la serie de media móvil (ma.Li) y la desviación (sigma2), los cinco parámetros se presentan acompañados con su correspondiente desviación estándar y análisis de test de Wald (estadístico, z, y p-valor). Como se observa, el modelo es ARIMA(0,1,4) y puramente de media móvil. El p-valor es de 0.00 en todos los casos, por lo que se asume que son valores significativos.

El último paso del diagnóstico es una evaluación de residuos. Este análisis apunta a determinar la calidad del residuo y cuánto este se aparta de la distribución normal ya que que si el modelo logra explicar la serie, se tendrá como resultado un residuo aleatorio, de media igual a cero y varianza constante.

Para la verificación del modelo se puede realizar una evaluación sobre las gráficas de los residuos y un análisis estadístico. De nuevo, en el proceso automatizado la inspección visual resulta poco práctica, aunque se utiliza para revisar los resultados, al procesar todas las series se opta por el análisis

	coef	std err	z	P>|z|
ma.L1	-0.4130	0.009	-47.302	0.000
ma.L2	-0.2226	0.010	-21.880	0.000
ma.L3	-0.0908	0.011	-8.352	0.000
ma.L4	-0.1084	0.010	-10.501	0.000
sigma2	449.9364	5.069	88.758	0.000

Figura 6.6: Estimación de parámetros para la serie correspondiente a la altura 79 km. El cuadro presenta los parámetros que definen el modelo ARIMA(0,1,4) que explica la serie temporal de esta altura particular, con ma.Li, i = 1,...,4 representando los coeficientes de la serie de media móvil, y sigma2 representando el error. Cada parámetro está acompañado de su desviación estándar (std err) y del estadístico (z) y la probabilidad ($P > |z|$) del test de significancia individual.

estadístico.

Un análisis visual del comportamiento del residuo ayudará a una primera impresión de la calidad del modelo y puede indicar rápidamente que el modelo no esta funcionando adecuadamente. Motivará, además, un análisis más a fondo sobre los resultados de los valores de los test estadísticos aplicados sobre el mismo.

A continuación se detallan los gráficos utilizados en el siguiente análisis:

- Se obtiene un gráfico del residuo normalizado, que permitirá observar aproximadamente la media y la varianza y su variabilidad. Se espera que la media sea cero y la varianza relativamente constante.
- El segundo gráfico corresponde a un histograma y una gráfica de la distribución de probabilidad del residuo, estimada por densidad de Kernel (KDE) contrastada con una distribución normal de media cero y varianza constante.
- El tercer gráfico corresponde al llamado Q-Q Plot o gráfico de cuantiles-cuantiles. Este gráfico es una comprobación que rigurosamente no constituye una prueba determinante, pero provee una rápida visualización del comportamiento de la distribución, en este caso del residuo. Cuando se compara con una distribución teórica, se elige una dada cantidad de cuantiles coherentes con el tamaño de la muestra y se grafican los cuantiles de la muestra estimados, en contraposición con los cuantiles teóricos de la distribución que se quiere comprobar. Si la muestra preserva la distribución teórica objeto de prueba los puntos se ubicarán sobre una recta, por lo menos en gran parte de los puntos en el rango central. Apartamientos pronunciados de la recta podrían indi-

car que no seria acertado tomar esta distribución como la distribución que sigue la muestra.

- Por último, se realiza un correlograma del residuo con el objeto de ver si este se mantiene amortiguado o si mantiene dependencias o muestra alguna estacionalidad o patrón periódico. Para el caso favorable de un modelo exitoso, este gráfico debería reflejar que el residuo no presenta patrones en el correlograma y la correlación decrece en los lags manteniéndose dentro de los límites de la banda de confianza.

En el siguiente figura, se ven los gráficos obtenidos para la serie de altura 79 km.

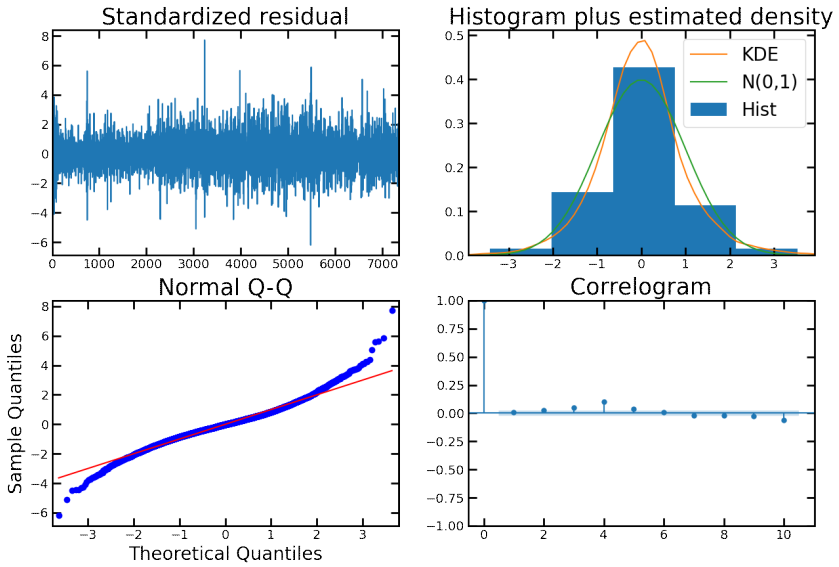

Figura 6.7: Residuo del diagnóstico: Altura 79 km. Inspección visual. En la primera fila se aprecia el gráfico normalizado (Izq) y el histograma junto con la curva KDE vs. una curva ideal de la distribución N(0,1) (Der). En la segunda fila, se presenta el Q-Q Plot (Izq) y el Correlograma del residuo (Der).

En las figura 6.7 el histograma del residuo parece estar centrado en cero y la curva producto de la distribución estadística aproxima a la distribución ideal N(0,1). El diagrama Q-Q también parecería indicar que los residuos siguen mayormente la distribución normal. Sin embargo, el correlograma y el hecho de que se observan aun algunos picos en los gráficos normalizados de las funciones de residuos, generan sospechas de que existe una componente no modelada entre los movimientos del dato.

Para el análisis estadístico del residuo se realizan test que evalúan la normalidad del residuo como se describirá a continuación. El test de Jarque

Bera (abreviado JB), es una prueba de ajuste de bondad que implica calcular ciertos parámetros de forma de la distribución de la muestra (los momentos tercero y cuarto) a fines de aproximar el estadístico JB. De manera que, se contrasta de forma conjunta la asimetría y la curtosis de la distribución. Se utiliza especialmente en regresiones para hacer pruebas de normalidad de residuos.

Dado el estadístico (Jarque y Bera, 1987):

$$JB = n\left(\frac{S^2}{6} + \frac{(K-3)^2}{24}\right) \sim \chi_2^2 \tag{6.44}$$

donde S es la asimetría y K representa la curtosis estimadas a partir de la varianza y el tercer y cuarto momento, respectivamente, se afirma que el estadístico sigue una distribución Chi cuadrado con 2 grados de libertad.

El objetivo de la prueba es analizar las hipótesis:

- H_0: (Hipótesis Nula) S = 0 y K - 3 = 0. Luego, la distribución es normal.
- H_1: (Hipótesis Alternativa) La distribución no es de tipo normal.

La hipótesis nula es rechazada si:

$$|JB| > (\chi_2^2)_{\alpha=0,05} \tag{6.45}$$

En otras palabras, si el valor absoluto del estadístico, que representa el apartamiento de una distribución normal, presenta un valor mas grande que el valor crítico estimado para la distribución Chi cuadrado con 2 grados de libertad dado un nivel de significancia de $\alpha = 0{,}05$, se rechaza la hipótesis. Este valor en tablas corresponde a 5.99, luego en caso que $|JB| > 5{,}99$ la distribución no sería de tipo normal. Equivalentemente se rechaza, si el p-valor (aquí representando la probabilidad de cometer un error asumiendo la falsedad de Ho) es menor al nivel de significancia.

Un segundo test, el test Ljung-Box-Pierce (abreviado LBP o LJ), puede aplicarse con el mismo objetivo y con hipótesis, a fines conceptuales, opuestas. Este test es una prueba conjunta para evaluar la nulidad en la autocorrelación de una determinada cantidad de rezagos menores al rezago m. Su mayor ventaja es que reemplaza una prueba individual de significancia para coeficientes de autocorrelación y se utiliza, en líneas generales, para comprobar que una dada serie se corresponde con ruido blanco.

El estadístico se define como (Ljung y Box, 1978):

$$LJ = n(n+2)\sum_{k=1}^{m}\left(\frac{\hat{\rho_k}^2}{n-k}\right) \sim \chi_m^2 \tag{6.46}$$

donde $\hat{\rho_k}^2$ es el coeficiente de correlación simple y se compara con una distribución Chi cuadrado con m grados de libertad. Cabe destacar que esta

prueba es una variante de la Prueba Q de Box-Pierce (Box y Pierce, 1970) y equivalente a la misma.

El objetivo de la prueba es analizar las hipótesis:

- H_0: (Hipótesis Nula) $\rho_i = 0$ para $i = 1, \ldots, m$.
- H_1: (Hipótesis Alternativa) Existe al menos un $\rho_i \neq 0$.

La hipótesis nula es rechaza si:

$$|LJ| > (\chi^2_m)_{\alpha=0,05} \tag{6.47}$$

En otras palabras, se rechaza Ho si el valor absoluto del estadístico, presenta un valor mas grande que el valor crítico estimado para la distribución Chi cuadrado con m grados de libertad dado un nivel de significancia de $\alpha = 0{,}05$. Equivalentemente se rechaza si el p-valor de la prueba se encuentra por debajo del nivel α, dado que el p-valor representa la probabilidad de cometer un error rechazando la hipotesis Ho.

A continuación se muestran resultados de dichos test sobre la serie de altura 79 km. Éstos comparan el residuo con el llamado ruido blanco, esto es, con una distribución normal de media cero y varianza constante.

Ljung-Box (L1) (Q):	0.72	**Jarque-Bera (JB):**	2158.69
Prob(Q):	0.39	**Prob(JB):**	0.00
Heteroskedasticity (H):	1.57	**Skew:**	0.22
Prob(H) (two-sided):	0.00	**Kurtosis:**	5.62

Figura 6.8: Residuo del diagnóstico para la serie correspondiente a la altura 79 km. Análisis Estadístico. Se muestra la salida de las pruebas estadísticas sobre el residuo.

Para el caso del test LJ no se rechaza la hipótesis nula de la independencia de los residuos, debido a que el p-valor (Prob(Q) = 0.39) no es menor al valor de significancia $\alpha = 0{,}05$. Es decir, no se puede negar que los residuos sean independientes (no correlacionados o ruido blanco). Luego en el caso del test JB la hipótesis nula Ho se rechaza por ser el p-valor (Prob(JB) = 0.00) menor al valor de significancia 0.05. En este caso, sí es posible decir que los residuos no siguen una distribución normal. Se concluye que por el resultado de ambos test, existe la sospecha de que el residuo no es de tipo normal.

Pronóstico

Este paso tiene como objetivo realizar una predicción de la serie sobre el intervalo no incluido en el entrenamiento a partir del modelo estimado

en el diagnóstico. Si el modelo elegido para una dada altura ajusta bien la serie y la evaluación de residuos del diagnóstico cumple con lo esperado, se procede a este paso.

En particular, cuando la serie presenta estacionalidad y el modelo elegido es ARIMA, bien puede resolverse la estacionariedad para el diagnóstico pero la etapa de pronóstico suele presentar problemas.

En principio, ARIMA no modela las componentes estacionales. En el proceso de transformar la serie en estacionaria es común remover estos efectos estacionales de la muestra de entrenamiento, luego el modelo realiza un diagnóstico óptimo de la misma. Luego en la etapa de pronóstico, se espera que el modelo elegido pueda aproximar los valores sobre la muestra de testeo, pero estos no cuentan con remoción de estacionalidad. De manera que ARIMA usualmente es capaz de generar un pronóstico acertado si la serie no presenta estacionalidad.

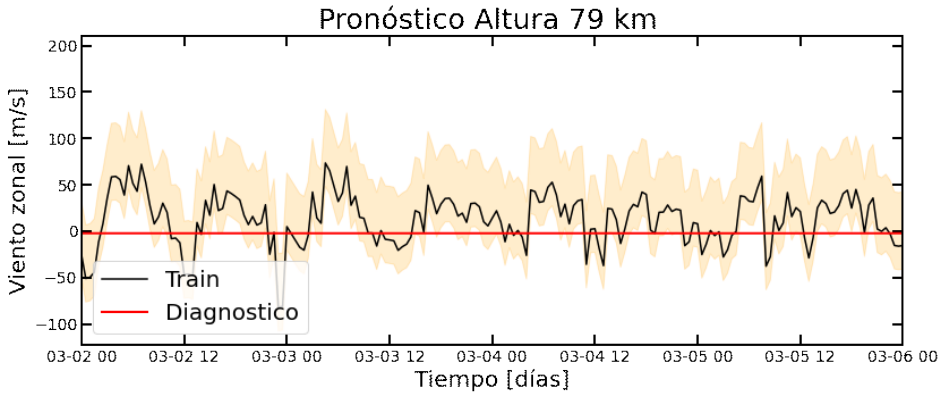

Figura 6.9: Pronóstico ARIMA para la serie correspondiente a la altura 79 km. Se observa la curva de testeo y la curva de valores pronosticados por el modelo ARIMA(0,1,4) simple.

ARIMA utiliza muestras hacia atrás para evaluar una determinada muestra actual. Por esto, si no se abordado la estacionalidad en el modelo es posible que éste no pueda realizar una predicción correcta. El modelo podrá pronosticar bien las primeras muestras, porque cuenta con la información del entrenamiento, pero conforme se avanza sobre el dominio de testeo, utilizará las propias muestras pronosticadas para generar las muestras siguientes y a éstas les faltará un componente de variación. El algoritmo tiende a establecerse en un valor medio y no puede reproducir el modelo pasadas un par de muestras como se ve en la figura 6.9.

El pronóstico se realiza en consideración de un intervalo de confianza. Este intervalo debe interpretarse como el rango dentro del cual debería encontrarse el apartamiento de la curva pronosticada para con la serie de testeo.

Enfoques alternativos para realizar un pronóstico acertado en caso de estacionalidad implicaría:

- Utilizar un modelo que explique la estacionalidad (modelo SARIMA). En cuyo caso, el principal problema consiste en el conocimiento a priori de los movimientos estacionales que contiene el dato y que el modelo pueda representarlos bien.
- Modelado de la dispersión. Los modelos del tipo ARCH y GARCH abordan este punto.

Como puede observarse, las anteriores opciones implican cambiar el modelo de representación. En esta tesis se opta por utilizar una metodología que permite mejorar este problema, sin cambiar estrictamente el modelo.

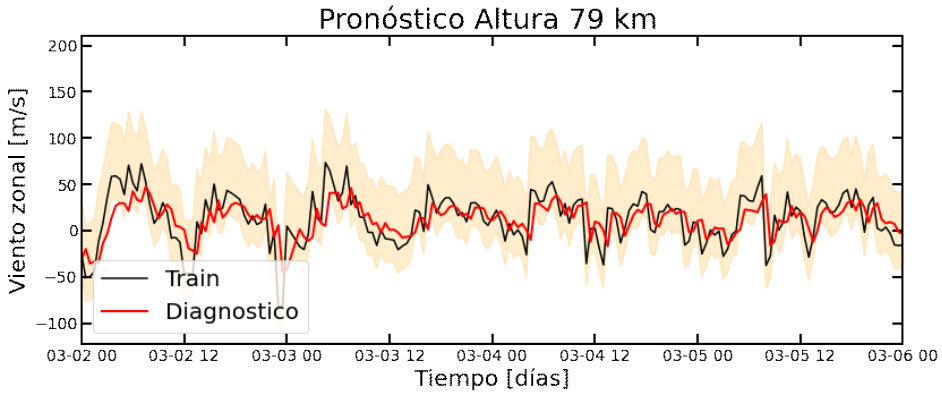

Figura 6.10: Pronóstico ARIMA con actualización punto a punto para la serie correspondiente a la altura 79 km. Se observa la curva de testeo y la curva de valores pronosticados por el modelo ARIMA(0,1,4).

La serie presenta estacionalidad y el pronóstico no resulta en un buen ajuste en todo el dominio del test. Se utiliza en el análisis la actualización del modelo con cada muestra pronosticada. Esta actualización representa una mejora para ARIMA comparado con la anterior estimación, como se observa en la figura 6.10.

Por otra parte, en un análisis de error se quiere cuantificar cuan bien se comporta el pronóstico para un determinado intervalo de testeo.

Los acontecimientos físicos que explican la serie y la elección de la partición de entrenamiento y testeo provocarán que el modelo ARIMA estimado en el entrenamiento sea más o menos acertado para uno u otro dominio de pronóstico. Este paso es significativo en el sentido de que el modelo será realmente representativo si puede responder bien en muestras secundarias del mismo proceso.

Se introducen a continuación criterios que permiten determinar si la estimación del modelo resulta acorde con la expectativa de ajuste. Estos criterios son medidas de error sobre las series temporales y sus pronósticos.

En primer lugar se considera a $\tilde{z}_i$ como el valor pronosticado para la muestra i y z_i es el valor real de la muestra.

Se define a continuación el error cuadrático medio, conocido como MSE, por sus siglas en inglés Mean Squared Error. Es una de las medidas de error mas conocidas y utilizadas.

El error cuadrático medio calculado sobre N muestras se expresa como:

$$MSE(z, \hat{z}) = \frac{1}{N} \sum_{i=0}^{N-1} (z_i - \hat{z}_i)^2 \tag{6.48}$$

El MSE es una medida de dispersión del error de pronóstico y es uno de los criterios de evaluación más usados para aplicaciones de aprendizaje supervisado de regresión. Por su forma matemática, donde los errores se elevan al cuadrado y se promedian, este criterio delata altos valores de error. Un valor alto de error mse podría indicar fallas en la estimación de muestras particulares y no serían representativas de la serie completa.

Por último, se considera el coeficiente de determinación, que es también llamado R^2. Se formula a partir de la siguiente expresión:

$$R^2(z, \hat{z}) = 1 - \frac{\sum_{i=1}^{N} (z_i - \hat{z}_i)^2}{\sum_{i=1}^{N} (z_i - \bar{z})^2} \tag{6.49}$$

Donde

$$\bar{z} = \frac{1}{N} \sum_{i=1}^{N} z_i \tag{6.50}$$

Y se considera que el numerador:

$$\sum_{i=1}^{N} (z_i - \hat{z}_i)^2 = \sum_{i=1}^{N} \epsilon_i^2 \tag{6.51}$$

El coeficiente de determinación R^2 es una medida de ajuste de regresión que se interpreta como sigue: un buen ajuste arrojaría valores cercanos a uno. Un ajuste deficiente presentaría en esta métrica valores cercanos a cero.

Para el caso de esta tesis, se utilizan ambos tipos de error. Como ejemplo, se presentan los valores para la serie correspondiente a la altura de 79 km. Notar que para cada una de estas medidas de error, se contará con un valor

de error por cada una de las series temporales. Los mismos, se calculan teniendo en cuenta el apartamiento de todas las muestras predichas respecto de los valores reales. Los resultados de los errores, basados en la figura 6.11 reflejan valores altos para el mse y muy cercanos a cero para r^2. Estos serían los errores resultantes para el mejor modelo encontrado para esta serie.

Altura	mse	r2
79	440.900003	0.353859

Figura 6.11: Medidas de Error en el Pronóstico para la serie correspondiente a la altura 79 km. Se observa el error cuadrático medio (mse) y el coeficiente de determinación (r2)

7 Resultados

Los procedimientos detallados en los capítulos anteriores para llevar a cabo ambas estrategias, la clásica y la de aprendizaje automático, se aplicaron a todas las series temporales a alturas fijas, en ambas componentes. A continuación, se presentan los resultados obtenidos para las series temporales correspondientes a alturas entre los 80 km y los 100 km, debido a que en este rango se encuentra la mayor densidad de datos.

En la sección 7.1 se describirán los resultados obtenidos con la primera estrategia, luego en la sección 7.2 se presentarán los modelos de ARIMA obtenidos para cada componente. En esta última sección, se presentarán resultados relacionados a la validación del modelo, esto es, análisis de coeficientes y normalidad de los residuos en el caso del diagnóstico y errores en el caso del pronóstico. Por último, se presentan brevemente los resultados de ambas estrategias de manera conjunta, con el propósito de contrastar las técnicas.

7.1. Resultados del método clásico

Para la primera estrategia, se tuvo como objetivo práctico hallar el valor medio y los coeficientes de la serie que explica las perturbaciones de marea según la ecuación 4.3.

En la figura 7.1 se muestran los vientos medios resultado del análisis de mínimos cuadrados sobre las componentes zonal y meridional. Los gráficos se muestran para el período transcurrido entre el 21 de septiembre y el 26 de marzo.

El mapa de viento medio para la componente zonal presenta en color rojo, vientos medios con dirección hacia el este y en color azul vientos medios dirigidos hacia el oeste. En él se puede observar que a fines de septiembre los vientos son predominantemente hacia el oeste con amplitudes que no superan los 14 m/s. Pero comienza a visualizarse en octubre, y se acentúa en los meses siguientes, una inversión del viento en altura que corresponde a los 85 km aproximadamente. Se observa también para comienzos de noviembre, un máximo de viento hacia el oeste de amplitud de entre 22 y 24 m/s que se sostiene por aproximadamente dos semanas, a los 80 km de altura. Durante diciembre se intensifican los vientos hacia el este alcanzando

amplitudes de 36 m/s para alturas mayores a 85 km a fines de este mes. En enero, parece decrecer la intensidad, pero luego se recupera. En febrero se alcanza el máximo en este período con vientos hacia el este de 40 m/s entre los 90 y los 95 km de altura. Finalmente, donde culmina la temporada de verano, febrero y principios de marzo, los vientos en el rango de alturas de 80 a 100 km son predominantemente hacia el este.

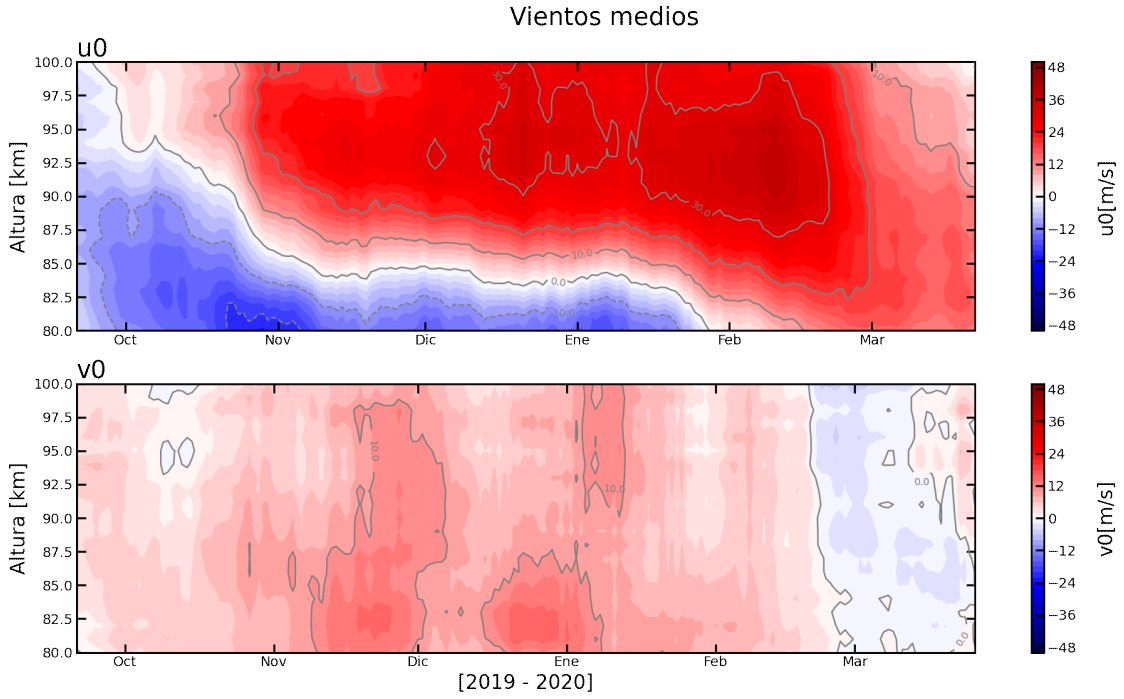

Figura 7.1: Vientos Medios de la componente zonal (arriba) y meridional (abajo).

Para la componente meridional, los vientos se presentan en colores rojos cuando la dirección es hacia el norte y azules cuando la dirección es hacia el sur. Para esta componente la intensidad en general del viento para todos los meses es menor. En septiembre, el viento en todo el rango de alturas, es predominantemente hacia el norte, pero de muy baja intensidad (de 2 a 4 m/s). Este comportamiento se sostiene en general hasta febrero. Se observa un pequeño período en octubre donde el viento meridional tiene dirección sur pero la amplitud es muy baja. Esta zona de dirección opuesta a los vientos dominantes coincide en altura y época, con el comienzo de la inversión en la componente zonal. En el período que abarca desde fines de septiembre a fines de febrero, el viento en esta componente tiene dirección norte. Los máximos hacia el norte se encuentran entre los 80 y los 85 km de altura, de aproximadamente 14 m/s, y se dan a fines de noviembre y a fines de diciembre. Estos máximos coinciden en altura con los máximos zonales

hacia el oeste. Cambia abruptamente la dirección con la llegada de marzo, para tornarse hacia el sur en ese mes con baja amplitud, de apenas 4 m/s.

En la figura 7.2 pueden observarse las amplitudes y fases de la componente semidiurna de mareas, de origen térmico, sobre ambas componentes, zonal y meridional.

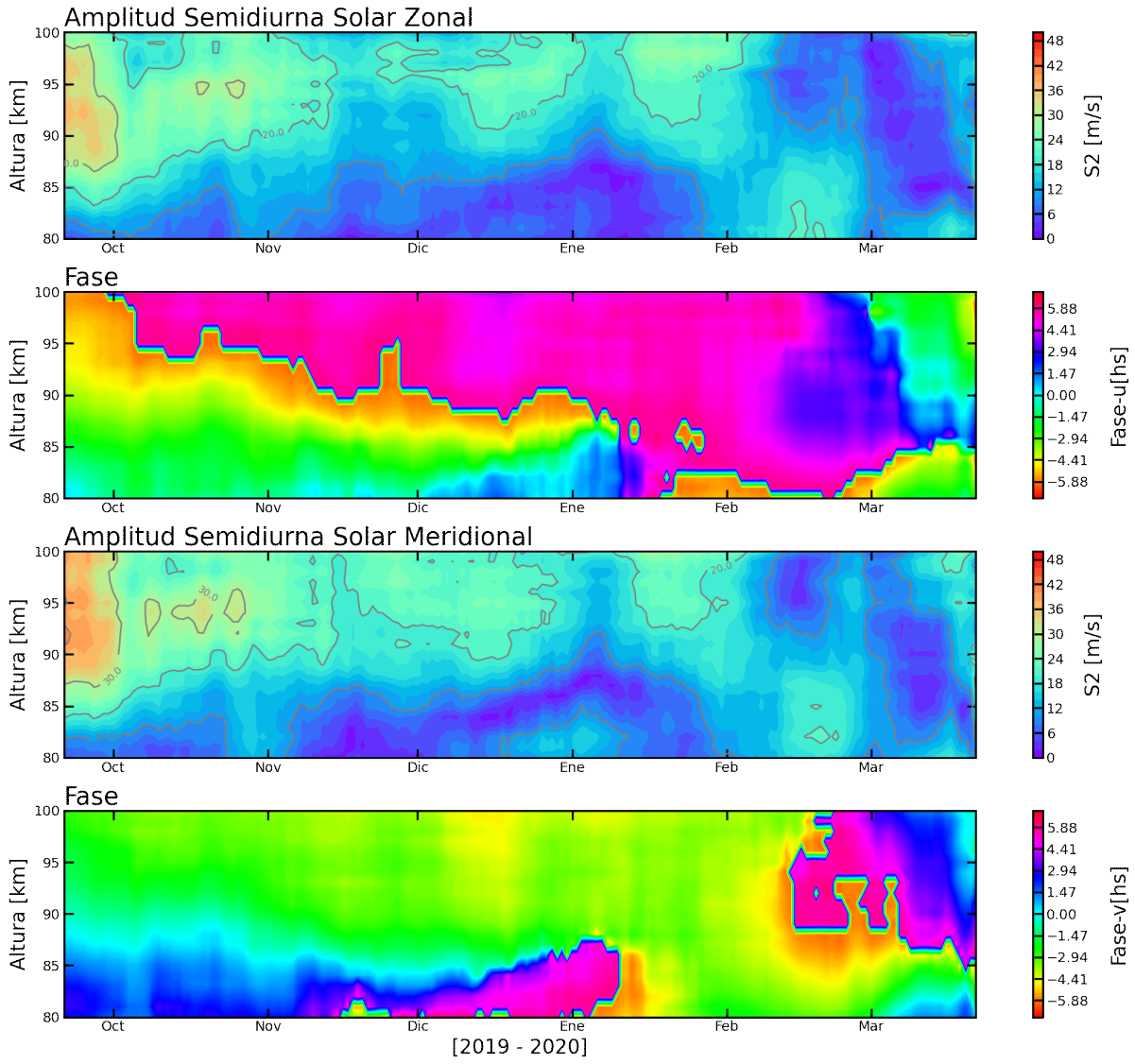

Figura 7.2: Marea Semidiurna Solar, amplitud y fase para la componente zonal (primer y segundo panel) y las correspondientes para la componente meridional (tercer y cuarto panel).

En la componente zonal se puede observar una zona de alta amplitud, alcanza los 36 m/s en el máximo, durante el intervalo del mes de septiembre

presente en el registro. En este mes se superan los 30 m/s en general para alturas entre 88 y 98 km. Para el mes de octubre y para alturas mayores a 90 km se mantiene una zona intensa, de amplitud mayor a 20 m/s, con dos máximos de amplitud mayor a 30 m/s a los 95 km de altura. Durante el mes de noviembre para alturas mayores a 90 km se ve que la marea se debilita y que a fines de diciembre se recupera, y pasa por un ciclo similar en el mes de enero. En estos casos las amplitudes máximas no superan los 30 m/s. Finalmente para el mes de febrero, la marea presenta un crecimiento de amplitud para alturas entre 80 y 90 km, que se debilita con la llegada de marzo.

En esta componente se analiza también la fase, limitando la observación a fases constantes y a identificar zonas de poca variación. En la componente zonal puede observarse que la fase se mantiene constante en gran parte del período muestreado (hasta enero) por debajo de los 90 km. Luego para otras alturas y en general a partir de enero, si bien se observan zonas de valor constante no se puede afirmar que la fase se mantenga relativamente constante en el tiempo a una altura determinada. Para los máximos de amplitud descriptos con anterioridad, solo podemos identificar para la zona de amplitud máxima de septiembre una fase constante por debajo de los 95 km, cuyo valor representa un retraso de unas 2 horas.

Para la marea semidiurna de la componente meridional, el comportamiento general observado es similar a la componente zonal. En septiembre se hace presente también el máximo de amplitud pero no tan concentrado en altura, se extiende por encima de los 88 km y alcanza valores mayores en amplitud superando los 40 m/s. También presenta los máximos de octubre con una amplitud similar, la cual supera los 30 m/s. En general, por encima de los 90 km se observan amplitudes mayores a 20 m/s de la marea en el período de septiembre hasta principios de enero. Se observan zonas de alta amplitud similares a la otra componente, pero abarcando mayor rango en altura y tiempo, durante todo diciembre y gran parte de enero. Y también se observa en febrero, un máximo para alturas entre 80 y 90 km.

La fase en cambio, se mantiene constante en un rango de alturas diferente, y en un período de tiempo mayor. Se observa fase constante por encima de los 90 km para el período que abarca desde septiembre hasta mitades de febrero, y para todas las alturas entre mediados de enero y mediados de febrero. Para alturas entre 82 y 85 km la fase se mantiene constante, desde septiembre hasta mediados de diciembre. En general, para todas las alturas a partir de mediados de febrero la fase varía en mayor medida.

Luego, la tendencia general de la marea semidiurna muestra que por encima de los 90 km suele tener amplitudes que superan los 20 m/s en el período de transición invierno-verano y en verano en ambas componentes, y en ambas el comportamiento cambia en febrero. Existe además un fuerte máximo en septiembre, en ambas componentes, cuyo valor supera la

tendencia general. Exceptuando este evento, los máximos de amplitud de esta componente se presentan ubicados en tiempo y espacio según varía el viento medio zonal hacia el este.

De la fase, cabe mencionar que cuando la marea es migratoria, puede que esta se mantenga relativamente constante en el tiempo. De las fases observadas, y en esta tesis, no es posible separar mareas migratorias de no migratorias. Aún así, observar la variabilidad de la fase, es un primer análisis de interés, que puede orientar posteriores estudios en este sentido.

En lo que respecta al resto de las componentes de mareas, se presentan sus amplitudes zonales en la figura 7.3.

La marea diurna (D1), muestra amplitudes que no superan los 20 m/s en septiembre, para alturas superiores a 92 km, donde la semidiurna también presenta el máximo. Luego se observa que en general esta componente se mantiene entre los 4 y los 6 m/s y sólo presenta una amplitud un poco mayor a 10 m/s en enero y a comienzos de febrero, por encima de los 92 km. También en febrero y por debajo de los 85 km se observa un aumento de amplitud similar.

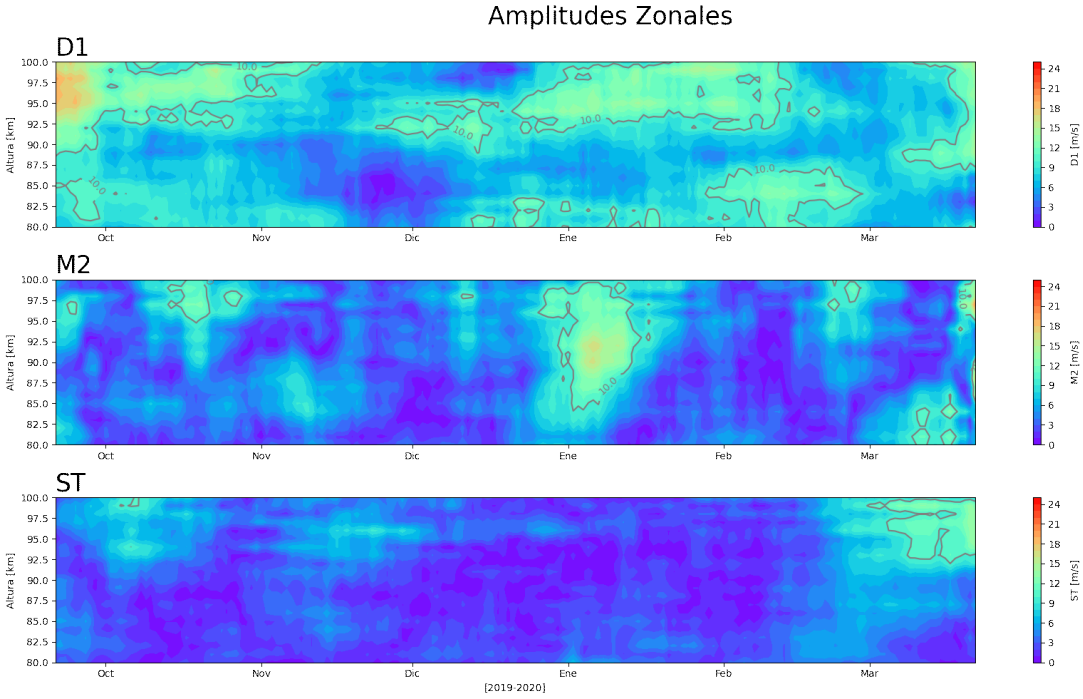

Figura 7.3: Amplitud de marea diurna (D1), semidiurna lunar (M2) y terdiurna (ST), entre los 80 y 100 km de altura, para la componente zonal.

La marea lunar semidiurna (M2) presenta cada dos meses aproximadamente, dos aumentos de amplitud por encima de los 5 m/s, en algunos

superando los 10 m/s. Se observa que el primero de estos aumentos se da por encima de los 90 km (en los meses de octubre y fines de febrero) y el otro, por debajo de los 90 km (en los meses de noviembre y marzo). El aumento de la primera mitad de enero es diferente al de noviembre y marzo, se extiende entre los 85 km y los 100 km y es el único que muestra amplitudes superiores a los 15 m/s. Se podría decir que el comportamiento es regular y que existe un aumento o evento especial en enero. Cabe destacar que sobre este aumento de enero se ubican dos gaps de datos faltantes que pudieron haber contaminado la amplitud para esta marea particular.

Por último, la componente terdiurna (ST) presenta amplitudes menores a 10 m/s, pero no presenta aumentos significativos. Se observa un aumento a comienzos de octubre cercano a los 100 km que supera levemente los 10 m/s, y un comportamiento similar en noviembre pero sin alcanzar el umbral anterior. Sobre marzo, por encima de los 95 km se ve también un aumento, pero se considera que se encuentra distorsionado por el efecto de los datos faltantes que siguen a este período de marzo.

En la figura 7.4 se presentan las restantes componentes de mareas meridionales.

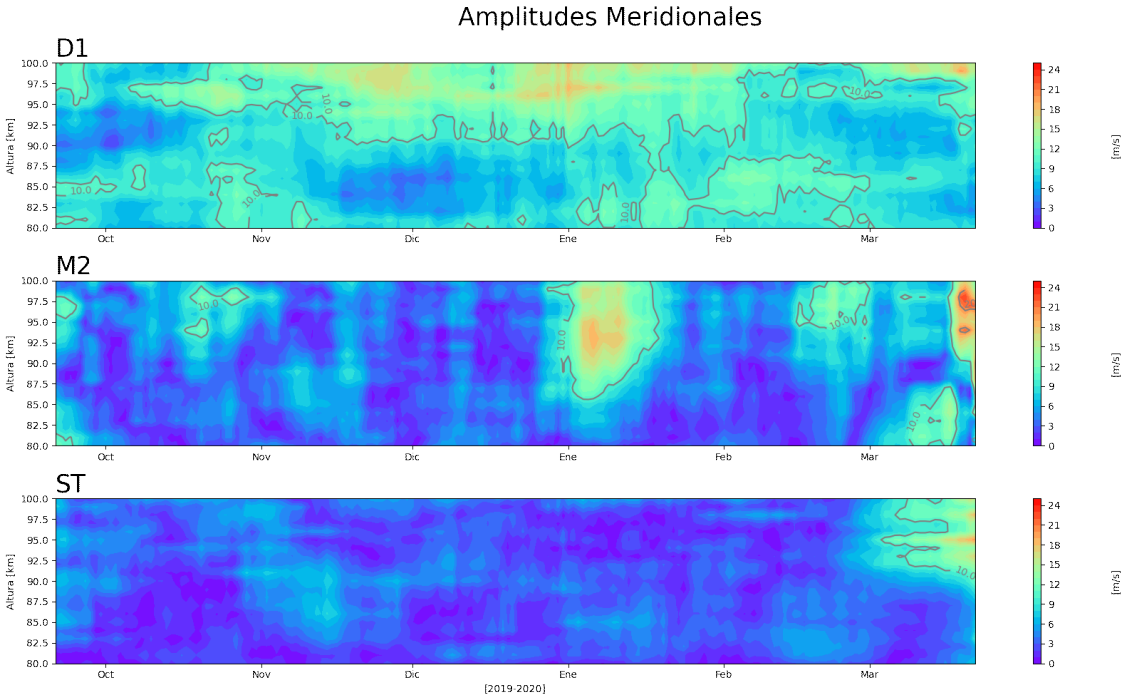

Figura 7.4: Amplitud de marea diurna (D1), semidiurna lunar (M2) y terdiurna (ST), entre los 80 y 100 km de altura, para la componente meridional.

En el primer recuadro se grafica la amplitud de la marea diurna meridional (D1) que en general es de amplitud mayor que la zonal en mayores

rangos. En septiembre, se observan amplitudes apenas por encima de los 10 m/s. Pero más tarde, entre noviembre e inicios de febrero por encima de los 90 km, la amplitud crece por encima de los 10 m/s hasta alcanzar los 18 m/s a inicios de diciembre y desciende para recuperarse nuevamente a principios de enero. Por debajo de los 90 km también se observan aumentos de menor magnitud a inicios de noviembre, y en enero y febrero.

La marea lunar meridional (M2) presenta un comportamiento similar a la componente zonal, con la diferencia que alcanza valores mas altos de amplitud en el aumento de enero.

La componente terdiurna (ST) muestra amplitudes menores a 10 m/s en general en todo el rango de alturas y el período observado. Presenta amplitudes un poco mayores a la media en septiembre y noviembre. Se observan aumentos pequeños también a inicios de enero y febrero, por debajo de los 90 km. De igual forma, se consideran datos de marzo muy cercanos al período de datos faltantes como poco confiables.

En principio, observando conjuntamente las componentes zonal y meridional, se observa que la meridional muestra mayor amplitud que la zonal y en rangos mas extensos para las componentes semidiurna solar y diurna. Es posible observar que en las componentes D1, M2 y ST se alcanzan amplitudes menores, en general, a las presentadas por la marea semidiurna solar. De las tres, la lunar parece la más regular y la terdiurna presenta menores amplitudes en general en todo el rango.

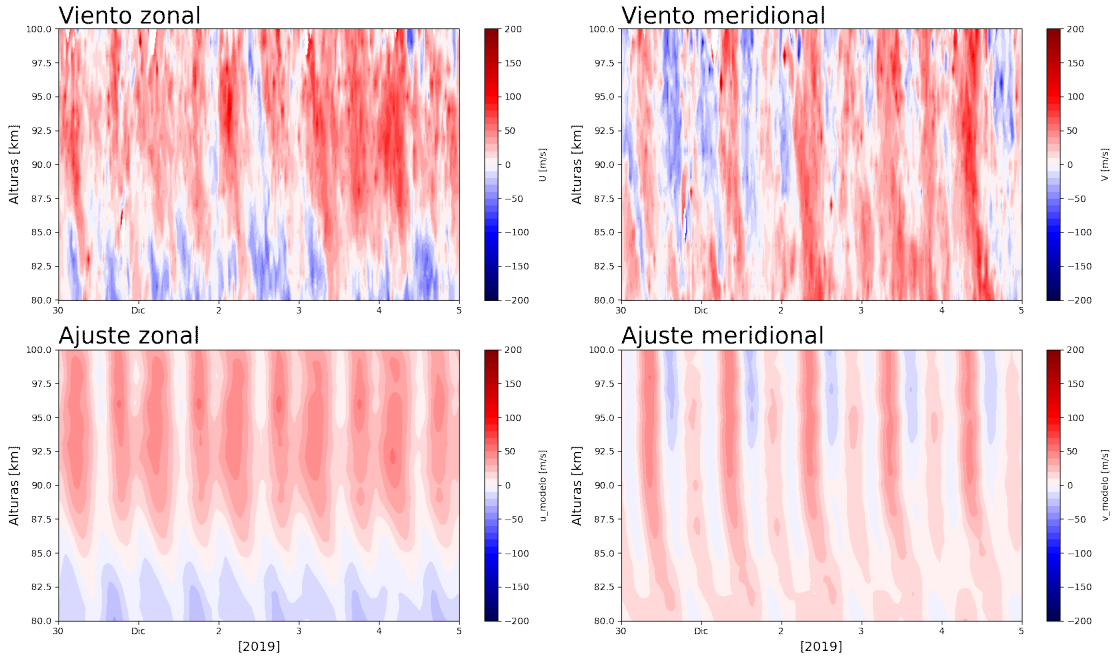

Figura 7.5: Ajuste de vientos por mínimos cuadrados, en la componente zonal y meridional.

En la figura 7.5 se presenta un gráfico de algunos días, del dato de viento y una estimación de este último obtenido a partir del ajuste de mínimos cuadrados, para la componente zonal y meridional.

Para cada punto de tiempo y altura, se han utilizado las correspondientes amplitudes y fases para constituir la expresión de viento y se ha evaluado el modelo en tiempo.

Visualmente, el ajuste de los coeficientes arroja valores de viento que se aproximan al comportamiento de los datos, aunque el ajuste resulta en amplitudes ligeramente menores y menos variables en el tiempo. Se observa el dominio de la marea semidiurna y la tendencia general del viento medio en ambas componentes.

7.2. Resultados del Modelo ARIMA

En este enfoque, se propuso como objetivo práctico hallar los coeficientes de las series autorregresiva y de media móvil que permitían aproximar la serie de vientos a cada altura fija, según la ecuación 6.26. Esta estimación determinaría el orden de cada componente y de un término medio si fuera necesario. Una vez realizadas estas estimaciones se validó el modelo realizando el diagnóstico In-Sample y el pronóstico Out-Sample.

7.2.1. Diagnóstico

Una vez identificados los modelos es necesario validar los mismos dentro y fuera de la muestra de entrenamiento.

Análisis de coeficientes

En la figura 7.6 se observa un esquema de coeficientes para el ajuste del diagnóstico ARIMA sobre la componente zonal. Cada coeficiente está caracterizado por su significancia, determinado de manera confiable en verde cuando su valor de p es menor que el nivel de significancia de $\alpha = 0{,}05$. Por separado se determina si dicha serie cuenta con un término medio. En cada altura y orden se especifica el valor determinado para tal coeficiente de forma comparativa, los valores son multiplicados por potencias negativas que no se expresan en el esquema por fines prácticos, pero se debe mencionar que todos los coeficientes resultaron menores a 1.

La componente zonal muestra que, en 23 de las 31 series temporales, un modelo de media móvil de segundo orden, sin un valor medio distinto de cero, es suficiente para representar la serie. Sin embargo, algunas series

precisan los componentes autorregresivos, estas series predominan por debajo de los 90 km. En determinadas alturas por debajo de los 90 km es necesario dos términos en el componente autorregresivo para poder representar la serie de vientos. Esto ocurre en 9 de las 31 alturas, como en los 78, 86, 87, 89 y 90 km.

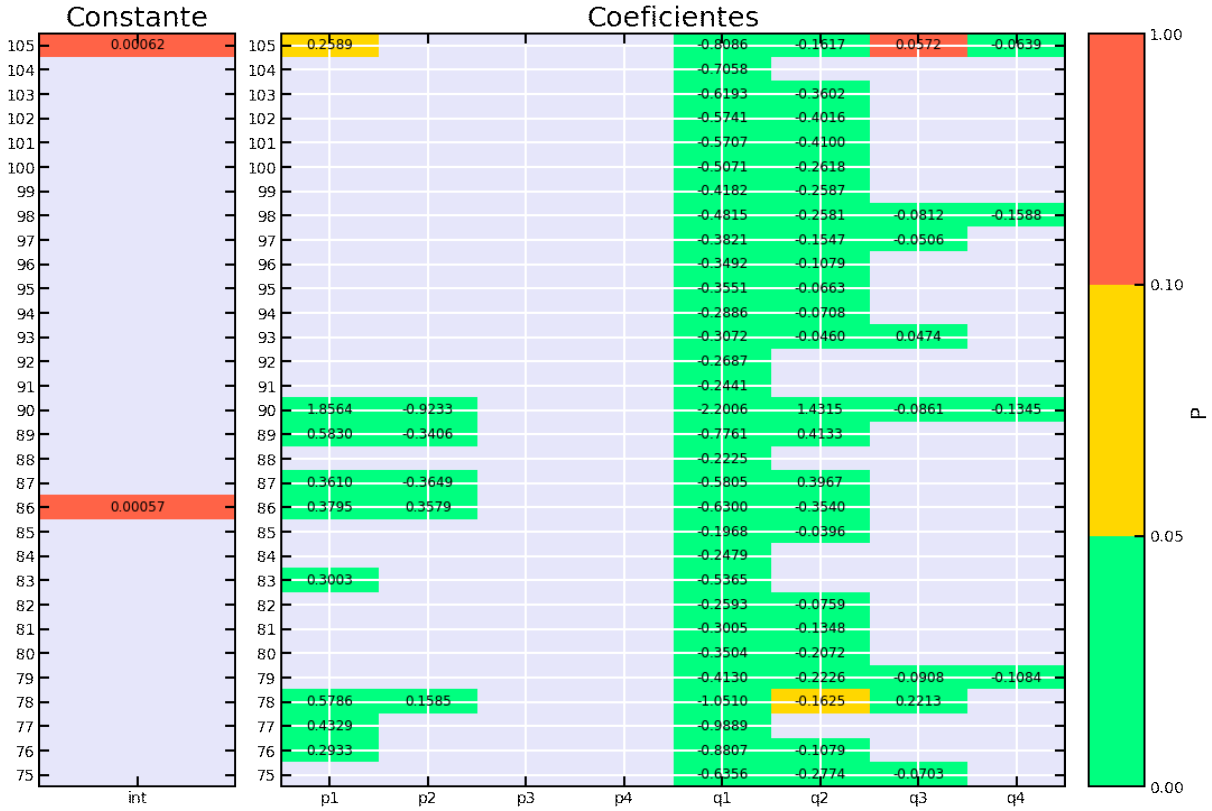

Figura 7.6: Ajuste de coeficientes para el modelo ARIMA sobre las series de vientos zonales a altura fija, con $p_i(i = 1, \ldots, 4)$ representando los ordenes de la serie autorregresiva y $q_i(i = 1, \ldots, 4)$ representando los ordenes de la serie de media móvil. Cada coeficiente está caracterizado con su valor de significancia, representado en colores por el valor de P.

Además, dos series presentan un valor constante, aunque este posee significancia por encima de 0.05 que es aceptado como valor de confianza. En general, se observa que las series no presentan un valor medio diferente de cero ya que han sido diferenciadas, en todos los casos una vez.

En pruebas anteriores de modelado con ARIMA, se observó que cuando se fija en cero el parámetro de diferenciación, las series presentan en su mayoría ordenes mayores a 2, puntualmente, cercanos a 4 o mayores. La diferenciación tiende a reducir los ordenes. Atendiendo a la cantidad de términos MA del modelo, ciertas alturas presentan ordenes altos incluso habiendo sido diferenciadas, esto es, con parámetro d=1. Alturas como la 78 km y 90 km requieren mas de dos componentes de media móvil y

componentes autorregresivos.

En la componente meridional, presentada en la figura 7.7, la tendencia general del modelo es diferente. En primer lugar, se observa que el valor medio está presente en la mitad de las alturas. Las series meridionales también han sido diferenciadas una vez. Aún así, los valores medios, se apartan en aproximadamente 20 m/s del cero. El orden de las componentes MA ahora es igual o mayor a 2 para 28 de las 31 series, incluso mayor o igual a 3 para 17 de las 31 series, estas últimas en general por encima de los 88 km. En general, las series en esta componente varían mas en torno al

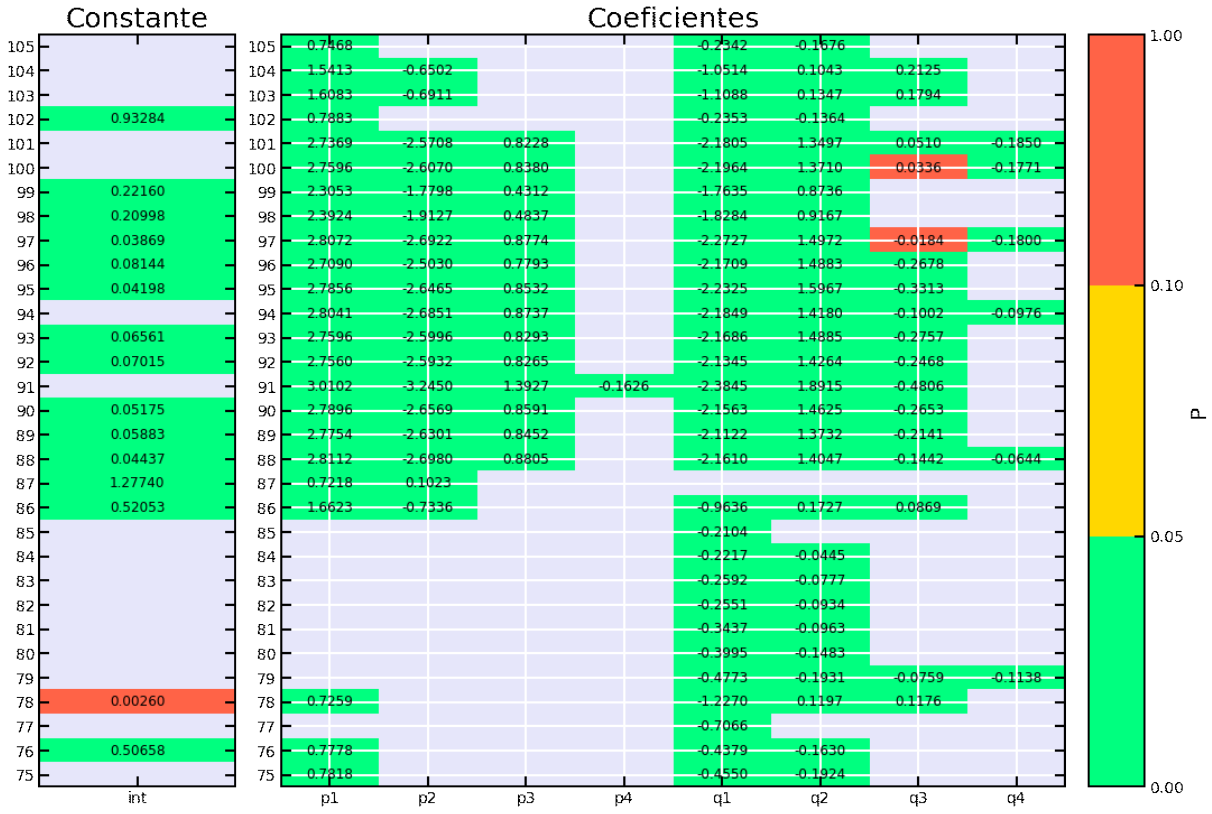

Figura 7.7: Ajuste de coeficientes para el modelo ARIMA sobre las series de vientos meridionales a altura fija, con $p_i(i = 1, \ldots, 4)$ representando los ordenes de la serie autorregresiva y $q_i(i = 1, \ldots, 4)$ representando los ordenes de la serie de media móvil. Cada coeficiente está caracterizado con su valor de significancia, representado en colores por el valor de P.

valor medio. En el caso de las series de la componente autorregresiva, se observan 18 series de las 31 totales que requieren 2 o más términos AR.

Se interpreta que la variación en la componente meridional es mas compleja, las series están más correlacionadas que en el caso zonal, probablemente esto se relacione a la estacionalidad de la serie que se presenta con mas intensidad en esta componente. Aun así, las series temporales que corres-

ponden a alturas por debajo de los 82 km presentan ordenes similares para ambas componentes, no así por encima de 87 km. Esta última observación podría indicar que, por encima de los 87 km, la variación se representa mejor en la serie diferenciada. Probablemente, esto se explique por la distribución de datos faltantes que se acentúa hacia los extremos.

Se observa una serie que presenta solo componente autorregresiva que se corresponde con los 87 km de altura.

Se concluye que la identificación del modelo ha presentado modelos ARIMA de mayor orden en la componente meridional que zonal. En la componente zonal predominó el modelo ARIMA(0,1,2) sin valor constante. Por el contrario en la componente meridional, el modelo mas representativo resultó ser el ARIMA(3,1,3) con término de valor medio. El aumento en los términos MA podría asociarse con mayor variabilidad en torno a la tendencia para la componente meridional, y la necesidad de los términos AR en ambas puede estar causado por cambios de variabilidad de la tendencia.

Ambas componentes presentan modelos generalmente similares por debajo de los 90 km y son mas bien diferentes por encima de esta altitud. Como se mostró en la observación de las componentes diurnas, semidiurnas y en el viento medio zonal, existe un comportamiento diferente sobre los 90 km y en general, mayores amplitudes.

Test sobre los residuos

Para la validación In-Sample se aplicaron pruebas de normalidad del residuo, con los test Ljung-Box(LJB) y Jarque- Bera(JB). .

Los resultados de la aplicación a cada serie de alturas fijas en la componente zonal, se muestran en la figura 7.8.

El test LJ se utiliza con el fin de que detectar una serie que pueda estar altamente autocorrelacionada, pero no asegura la independencia en la correlación. Recordando esto, el caso favorable para esta tesis consistiría en no rechazar la hipótesis nula Ho. El rechazo se asegura cuando el valor de p es menor o igual al valor de significancia $\alpha = 0{,}05$. De la figura, es claro que para la mayoría de series en el rango de alturas seleccionado, los valores de p se mantienen por encima del nivel de significancia. Con lo cual no se podría rechazar la hipótesis en general. Se revisa el valor de p para la altura 98 km, el cual parece cercano a $\alpha_{0{,}05}$. El p-valor para esta serie es de 0.07 por lo cual también se rechaza la hipótesis en este caso. Del test anterior no se pudo llegar a una conclusión sobre la independencia del residuo y se procedió a aplicar el test de JB. Los resultados también se encuentran representados en la figura 7.8.

El test JB se utiliza para confirmar que la distribución de los residuos no es de tipo normal. El caso favorable en esta tesis, es no rechazar la hipótesis

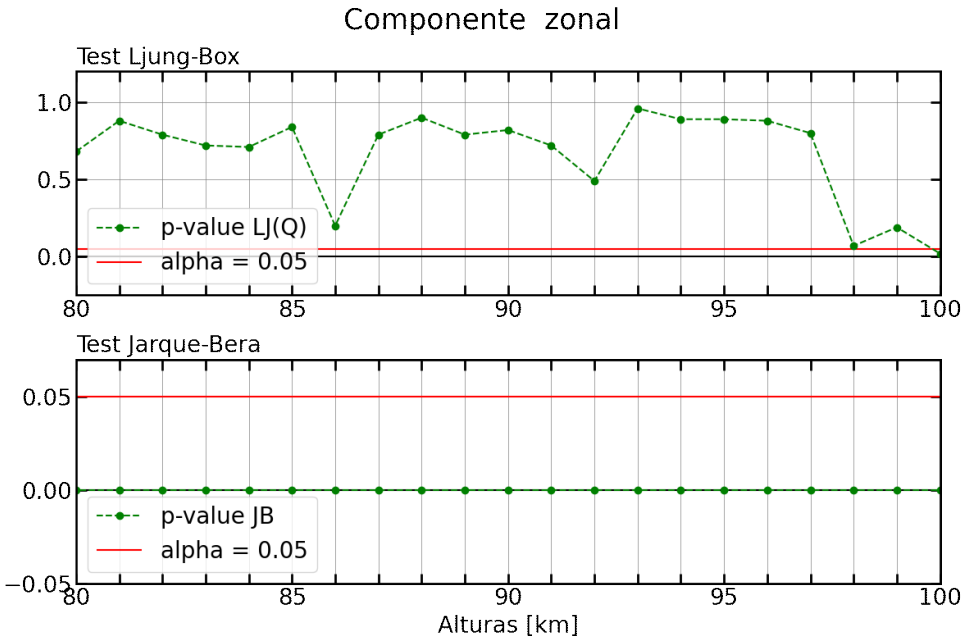

Figura 7.8: Diagnóstico: Test de normalidad de residuos sobre la componente zonal.

nula Ho. La misma se recusa si el p-valor de una serie es menor al nivel de significancia $\alpha = 0{,}05$. Este es el caso para las alturas consideradas, los valores de p son nulos para todas las alturas en el rango. El resultado del test da certeza de que el residuo no es normal en la componente zonal.

Para la componente meridional, se muestran también los resultados en la figura 7.9. El test LJ muestra resultados similares a los de la componente zonal. Los p-valores son, en general, mayores a α. De nuevo la prueba no da certeza sobre la hipótesis nula para la mayoría de las series. Sin embargo, en la serie correspondiente a los 95 km de altura se encuentra una excepción. El p-valor correspondiente es 0,01, por lo cual se rechaza la hipótesis y se podría considerar que el residuo estaría autocorrelacionado. Luego, al observar los resultados del test JB, para todas las series resulta en el rechazo de la hipótesis y en la confirmación de certeza de una distribución no normal.

En el caso de la altura 95 km, se observan los gráficos propios del análisis visual en la figura 7.10. El correlograma de los residuos de esta serie muestra, como otros, un patrón. En este caso, los valores de autocorrelación superan por poco el umbral de confianza, pero la existencia de este patrón denota una relación estacional en los datos. Se propone que la estacionalidad no modelada por ARIMA, evidente en el correlograma, puede ser la causa de no obtener residuos normales.

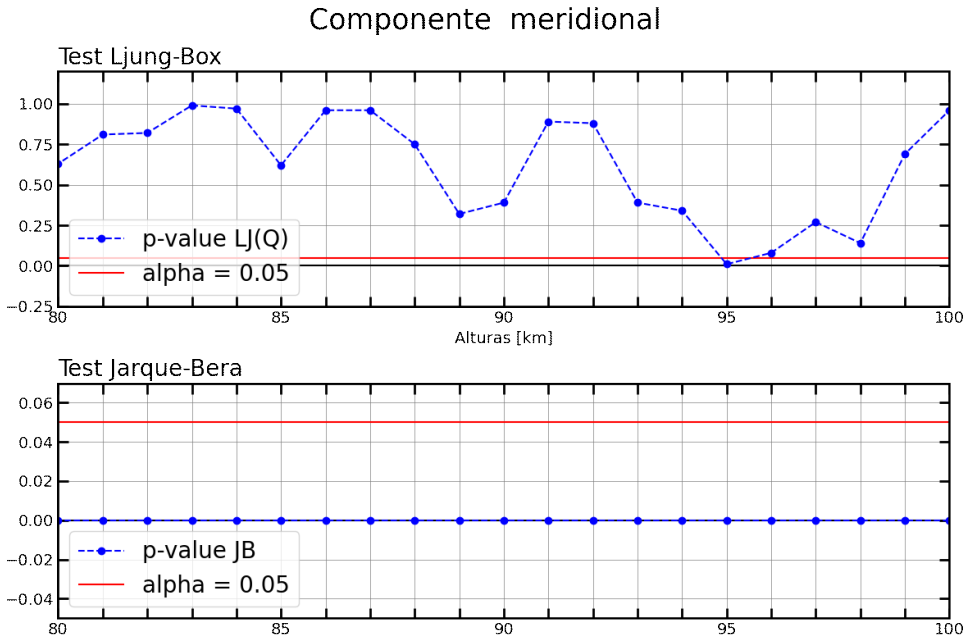

Figura 7.9: Diagnóstico: Test de normalidad de residuos sobre la componente meridional.

7.2.2. Pronóstico

La validación del modelo fuera de la muestra de entrenamiento consistía en realizar el pronóstico sobre algunos días de marzo dentro del rango reservado para testeo, entre alturas de 80 y 100 km.

El resultado, junto con los datos de vientos, se muestran en mapas en la figura 7.11 para la componente zonal.

En la inspección visual, el pronóstico realizó una buena aproximación de los datos en el rango seleccionado. La variabilidad presente en el pronóstico es similar a la del dato a excepción de algunos máximos que se presentan con menor intensidad, como el máximo al este el día 6 y el máximo hacia el oeste el día 7. Este resultado no hubiese podido alcanzarse si no se realizaba una actualización del modelo muestra a muestra, por no haberse modelado la componente estacional, como se explicó anteriormente.

Con énfasis en este último punto, se procede a analizar los resultados de las medidas de error para el pronóstico. Cabe destacar que los resultados a continuación sólo consideran un error por serie de altura, que corresponde al modelo ARIMA(p,d,q) elegido, es decir, aquel que minimiza el criterio BIC en la fase de entrenamiento. Se presenta en la figura 7.12 los valores resultantes para el error cuadrático medio (mse).

Relativamente al valor de los vientos en el dato, el mse presenta valores altos, incluso mayores de tres veces el valor máximo (198.4m/s). También se puede observar que el error aumenta conforme aumenta la altura, superando

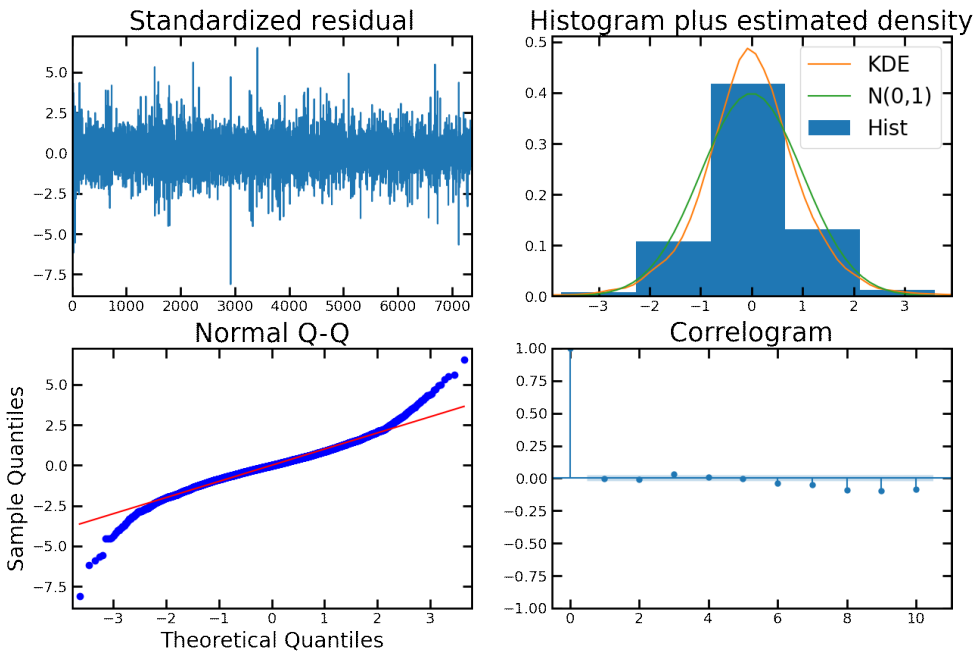

Figura 7.10: Resultados de la evaluación del residuo para la serie de la altura 95 km.

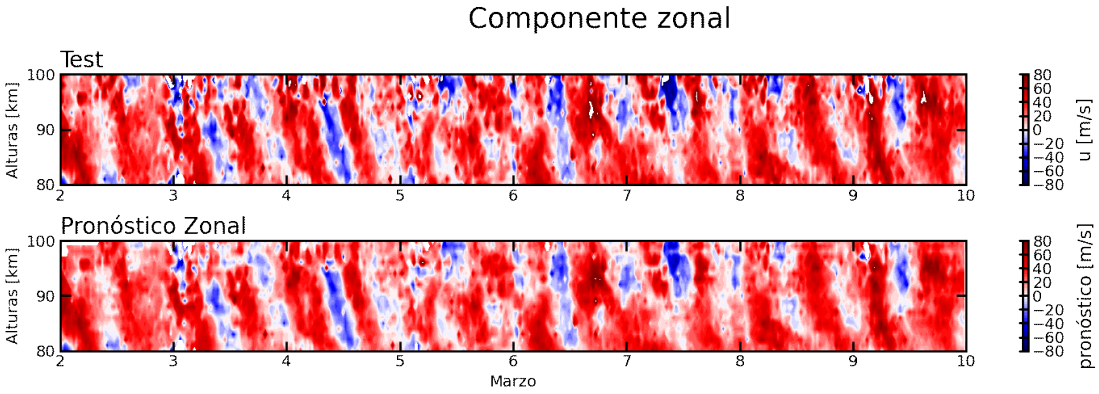

Figura 7.11: Pronóstico en la componente zonal: datos de vientos en el rango reservado para testeo (arriba), pronóstico obtenido sobre el mismo rango (abajo).

en la serie correspondiente a los 100 km el doble del error en la serie de 80 km. El error mse se define a partir de las distancias de cada muestra pronosticada al valor del viento como dato. En el pronóstico realizado cada muestra se genera a partir de las muestras anteriores.

Se presenta en la figura 7.13 el pronóstico a detalle para la altura 97

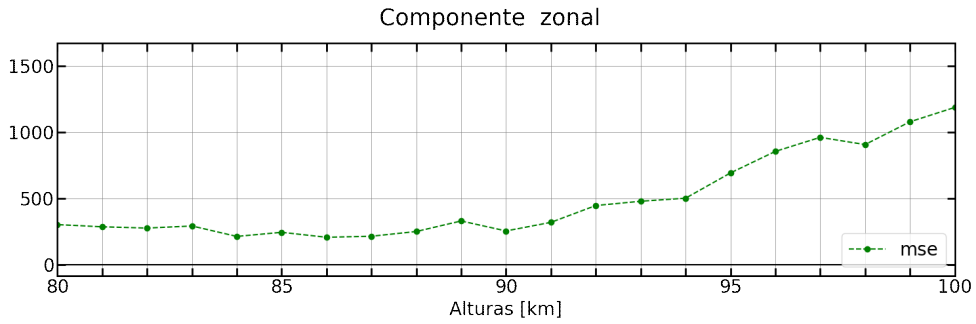

Figura 7.12: Error cuadrático medio (mse) evaluado sobre el pronóstico.

km, la cual presenta uno de los mse mas altos del rango (961.30 m/s). El

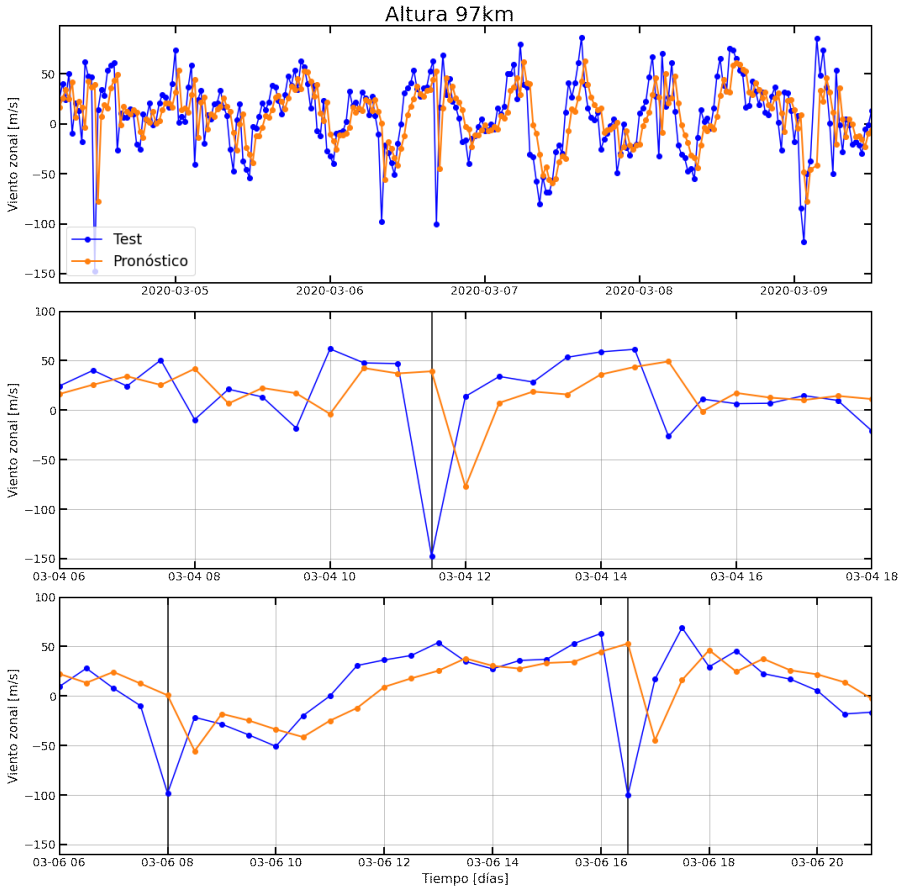

Figura 7.13: Detalle del pronóstico para la serie de la altura 97 km para el día 4 de marzo de 2020.

modelo ajustado para esta serie fue ARIMA(0,1,3). Puede verse que en una

determinada marca de tiempo el valor pronosticado es predominantemente similar al valor de la muestra anterior que al de la muestra correspondiente. En los recuadros siguientes de la misma figura se presentan algunas de las muestras con mayor diferencia entre el valor pronosticado y el dato, señaladas con marcadores negros. Aquí las diferencias alcanzan los 150 y 200 m/s. Se considera que el pronóstico, por la metodología del modelo se halla trasladado una muestra hacia adelante.

Se repite el análisis realizando la traslación de la serie pronosticada una muestra hacia adelante en tiempo como se muestra en la figura 7.14.

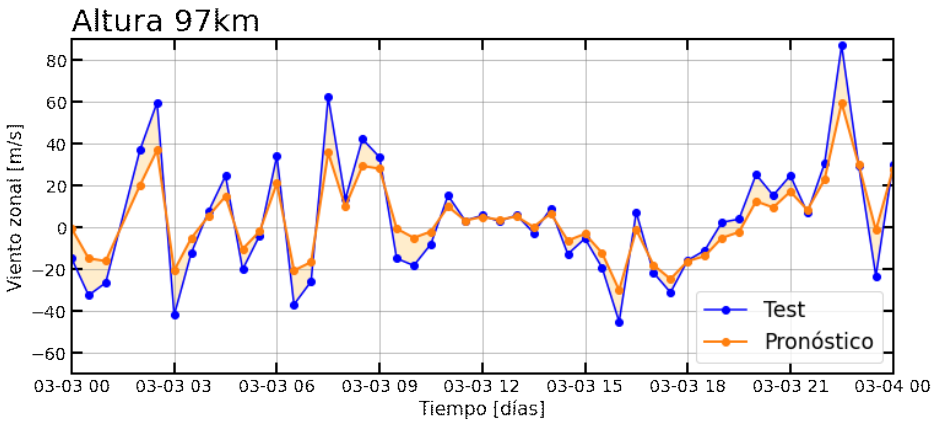

Figura 7.14: Serie pronosticada, adelantada una muestra.

El mse resultante en este caso desciende a 157.94 m/s. Esta mejoría del error se observa en general para el resto de las series como se observa en la figura 7.15.

De lo analizado anteriormente se concluye que cuanto mayor variabilidad y amplitud presente el dato de viento, mayor error presentará la serie. Esta será una de las razones que provoca un aumento del error con la altura.

El mse del pronóstico adelantado o corrido una muestra hacia adelante, aún parece alto en comparación con el valor de los vientos, siendo prácticamente de igual orden que los mismos. Esto se explica, en parte, por la fórmula matemática del mse que eleva al cuadrado las distancias entre puntos de manera que maximiza aquellos errores de mayor valor.

En la figura 7.16 se presentan los resultados para el coeficiente de determinación (r^2).

Este error toma valores entre 0 y 1, para el pronóstico en el rango de alturas de 80 km a 94 km el r^2 se mantiene entre 0.4 y 0.6, lo cual representa un término medio en la aptitud del modelo para esta métrica. Pero por encima de los 95 km los valores descienden. Teniendo en cuenta que estos modelos representan la mejor elección para cada altura, este descenso representa que a mayores alturas los modelos se alejan de la mejor predicción posible.

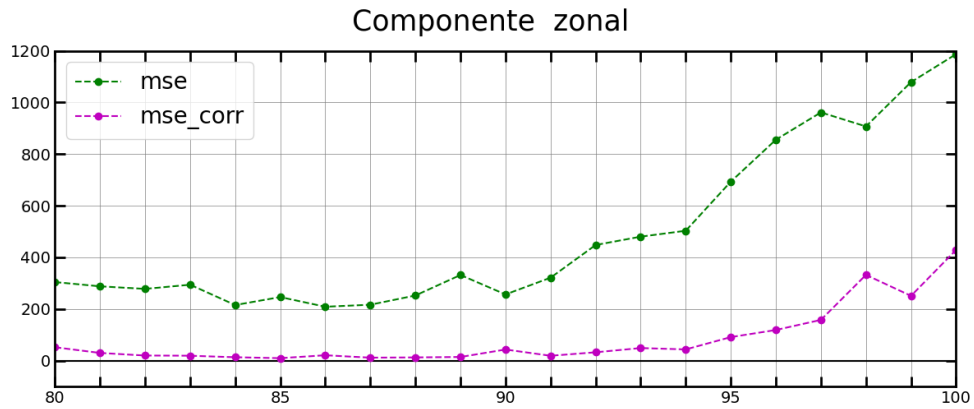

Figura 7.15: Mejoría en los errores para series pronosticadas adelantadas una muestra (mse_{corr}) en comparación con los errores de las series pronosticadas regulares (mse).

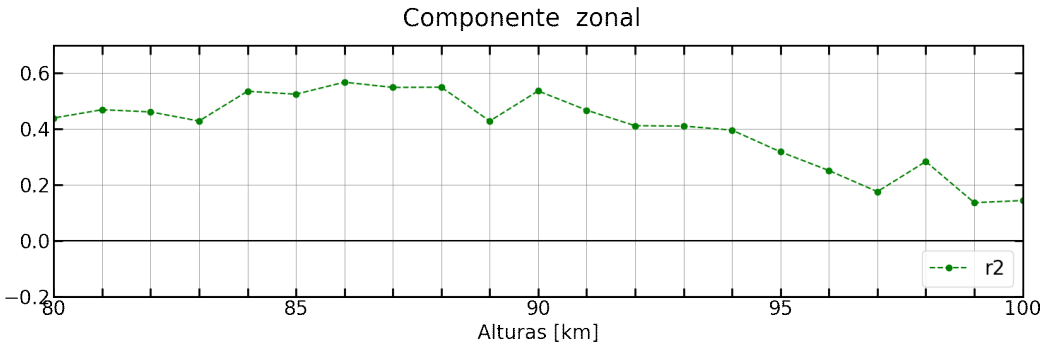

Figura 7.16: Coeficiente de determinación (r^2) evaluado sobre el pronóstico.

Por lo observado en el mse y en r^2, es claro que por encima de los 90 km y más precisamente por encima de los 95 km, la variabilidad del dato y la metodología del modelo resultan en predicciones de baja precisión.

Para la componente meridional la inspección visual del mapa del pronóstico, figura 7.17 indica que se presentan zonas en todo el período temporal donde la amplitud se ve disminuida, al igual que en la anterior componente. Esto supone apartamientos del pronóstico respecto del dato que deben reflejarse en los errores.

En la siguiente figura 7.18 se presentan los resultados obtenidos para los errores en la componente meridional, se exponen también los resultados para un pronóstico adelantado una muestra. En la anterior componente el comportamiento indicaba un crecimiento del error a mayores alturas. En la componente meridional, en cambio, el error primero decrece, siendo mínima para la serie correspondiente a los 86 km, y luego comienza a crecer

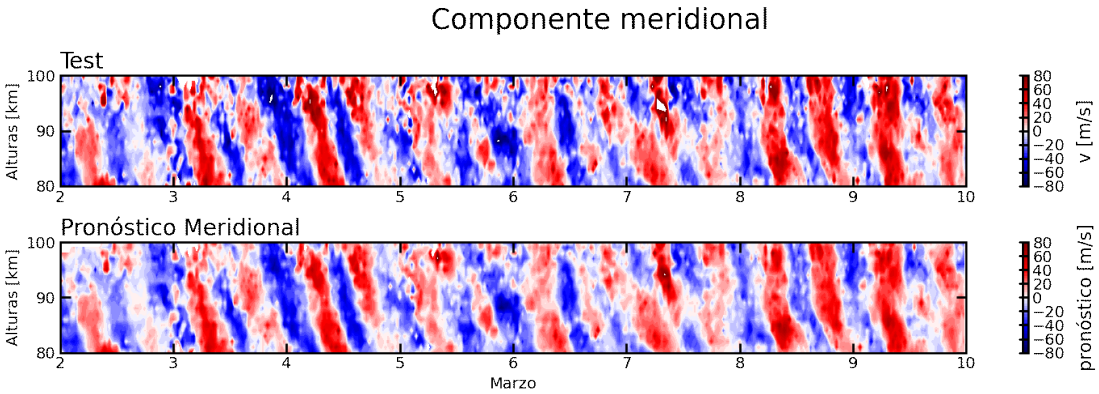

Figura 7.17: Pronóstico en la componente meridional: datos de vientos en el rango reservado para testeo (arriba), pronóstico obtenido sobre el mismo rango (abajo).

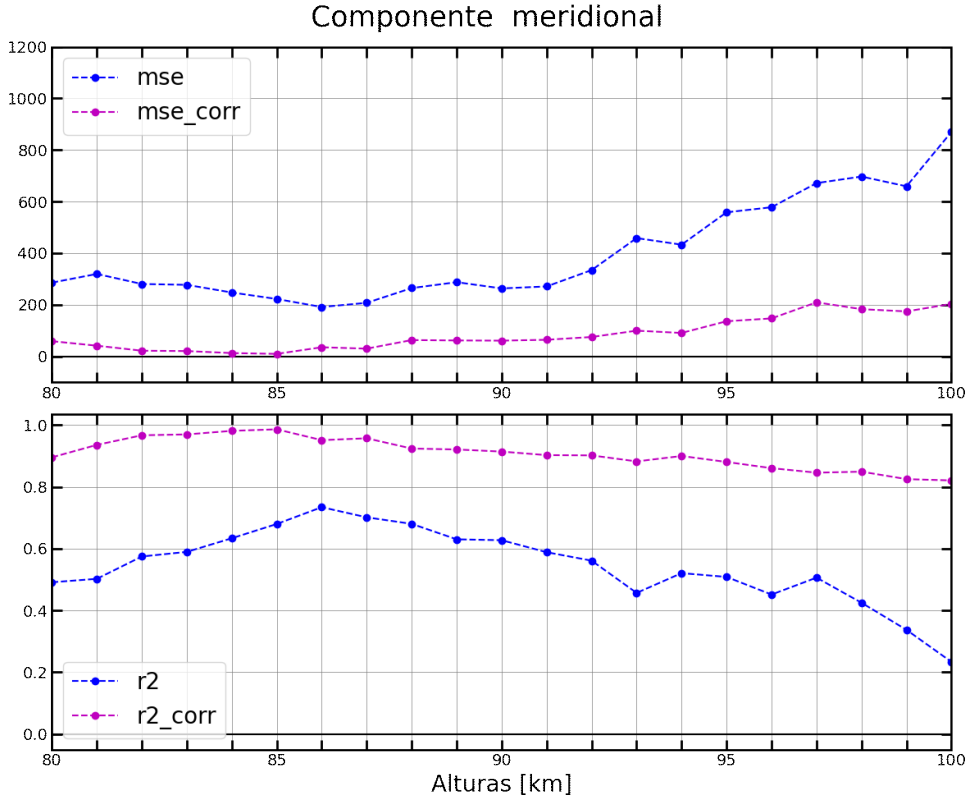

Figura 7.18: Error cuadrático medio y coeficiente de determinación (r^2) evaluado sobre el pronóstico (azul) y sobre el pronóstico adelantado (rosa).

nuevamente. Una conclusión posible, observando la distribución de los datos faltantes, es que sabiendo que esta componente posee mayor variabilidad y debido a los intervalos donde los datos cambian abruptamente de un valor

positivo a negativo, y viceversa (o simplemente cambian en valor absoluto de forma abrupta), y también explicado por la presencia de gaps, es posible que los errores arrojen valores más desalentadores. En especial el mse, que resalta la comparación de muestras particulares y están calculados a partir de las mismas. Esto se podría traducir como un error esperable de esta metodología en zonas con gran cantidad de datos faltantes, es decir, las series de alturas en los bordes del rango.

7.3. Breve contraste de estrategias

El ajuste clásico permite obtener aproximaciones diarias de los parámetros buscados. De esta forma, se analizan también promedios diarios simples de los resultados obtenidos en el modelo de aprendizaje automático y sobre los datos.

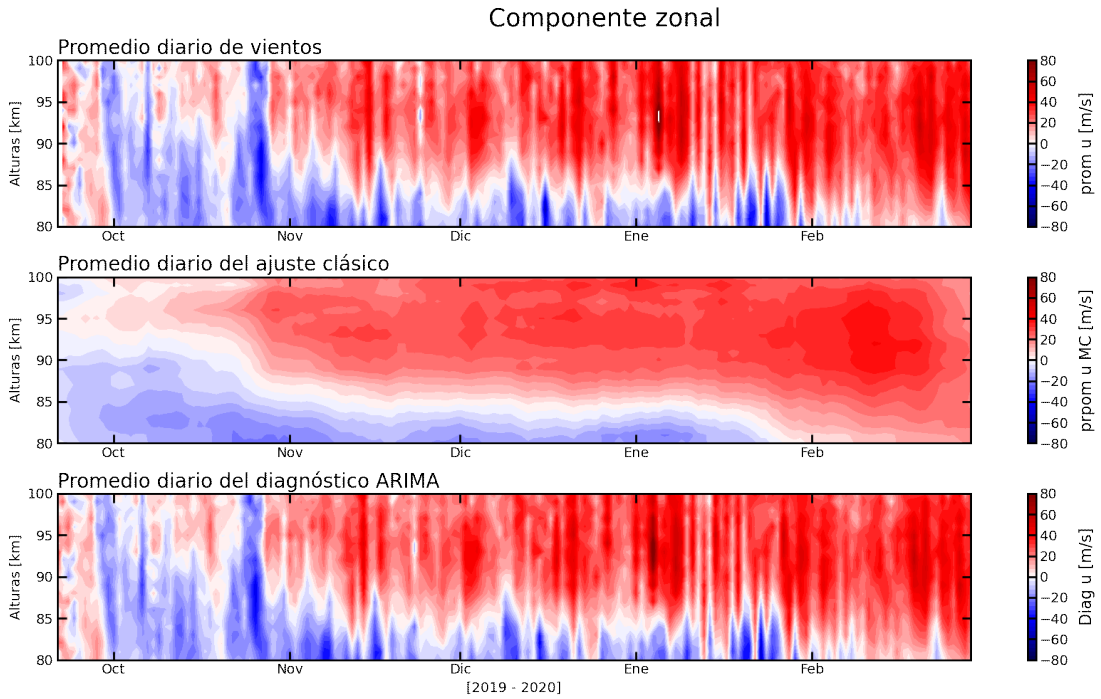

Figura 7.19: Promedios diarios de los datos de vientos (arriba), del viento medio resultante del método clásico de mínimos cuadrados (medio) y del viento ajustado por aprendizaje automático (abajo) para la componente zonal.

En la figura 7.19 se resumen los resultados de ambos enfoques para la componente zonal. En el primer recuadro se presentan los valores promediados de los datos de vientos, que son ajustados y presentados en el

segundo recuadro por el método clásico de mínimos cuadrados y en el tercer recuadro por aprendizaje automático.

Específicamente, el segundo recuadro muestra la evaluación de la serie modelada sobre los mismos puntos que fueron muestreados en tiempo y se obtiene un promedio diario por altura, habiéndose obtenido previamente parámetros, diarios también, de viento medio, amplitudes y fases para las diferentes componentes de mareas. En el tercer recuadro se muestran los promedios diarios de las series producto del diagnóstico ARIMA en cada altura fija. Es decir, se tomaron promedios diarios de los valores obtenidos por los modelos ARIMA sobre el intervalo de entrenamiento, la muestra utilizada para ajustar los coeficientes.

En la figura 7.20 se presenta los mismos resultados para la componente meridional.

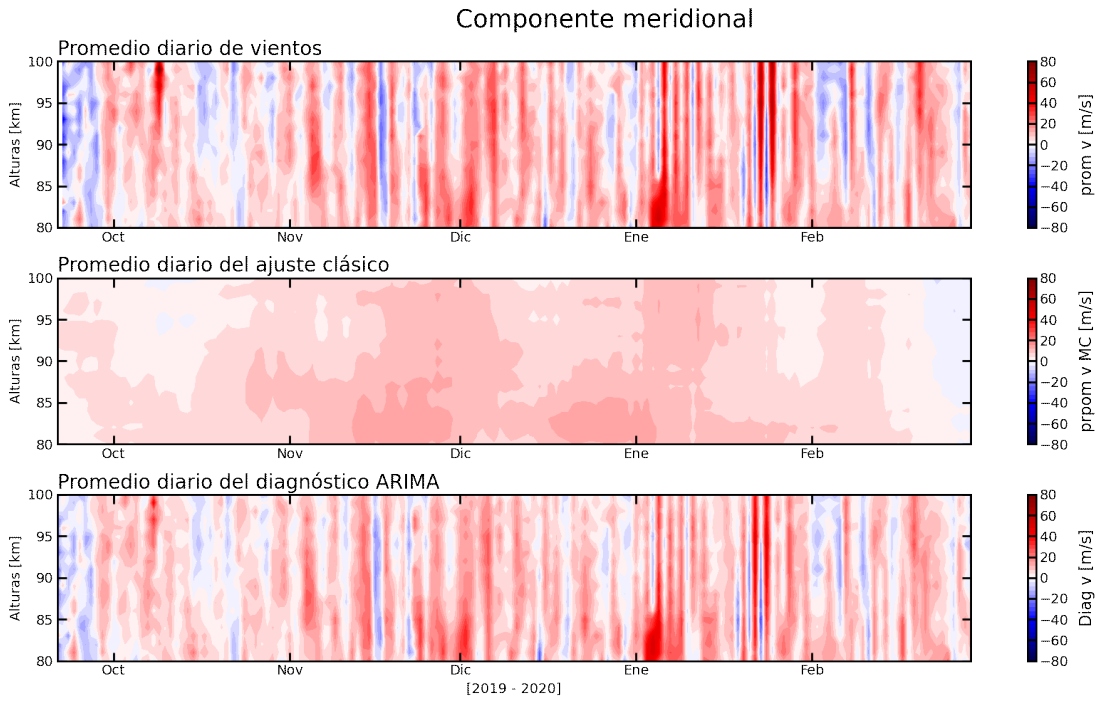

Figura 7.20: Promedios diarios de los datos de vientos (arriba), del viento medio resultante del método clásico de mínimos cuadrados (medio) y del viento ajustado por aprendizaje automático (abajo) para la componente meridional.

Mientras que el ajuste clásico se acerca al viento medio obtenido como parámetro de ajuste en torno al cual se presentan las variaciones de las componentes de marea, el modelo de aprendizaje automático reproduce casi exactamente el dato muestreado en contenido diario, en ambas componentes.

Con el objetivo de evidenciar las diferencias entre las dos técnicas utilizadas en este trabajo de tesis, se presentan los residuos de las series temporales

de forma visual, contrastado con las diferencias del modelo de ajuste clásico, en la figura 7.21 para la componente zonal y 7.22 para la meridional.

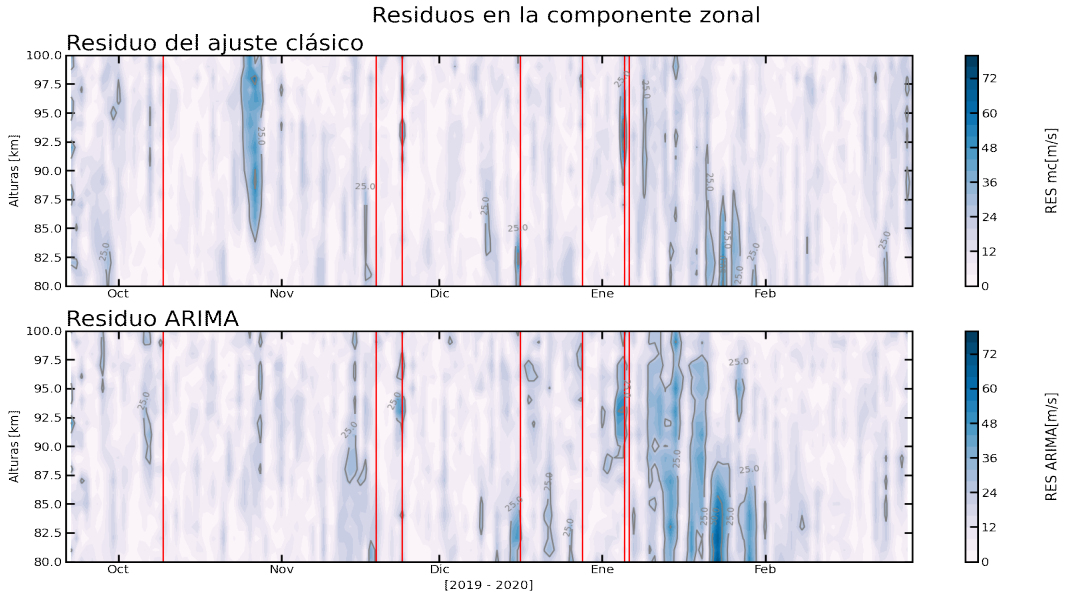

Figura 7.21: Diferencias de los modelos de ajuste clásico y ARIMA respecto a los datos de vientos, en la componente zonal. En rojo: marcadores de datos faltantes.

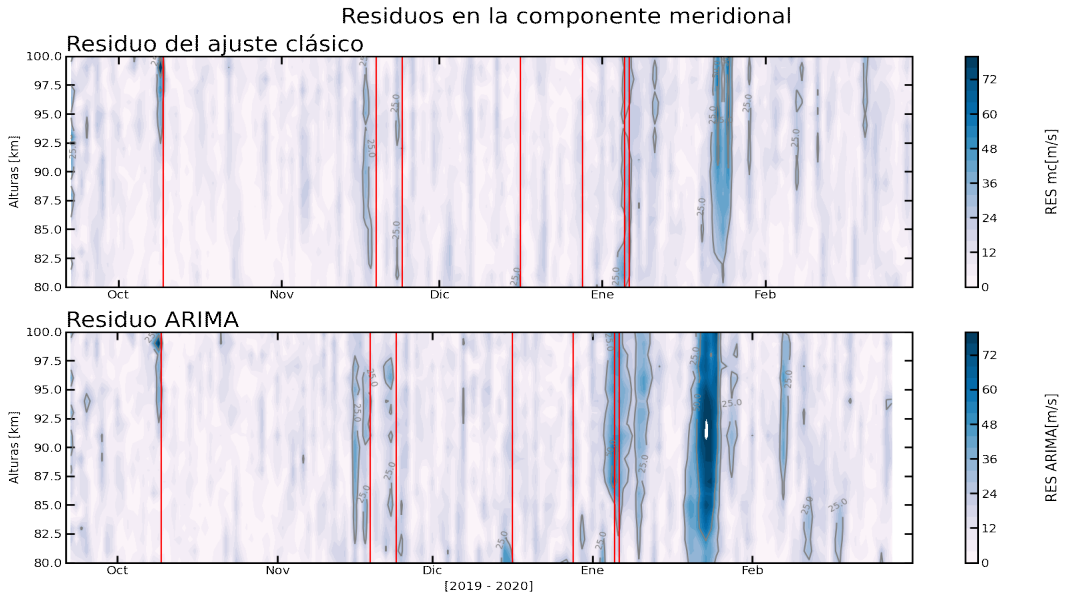

Figura 7.22: Diferencias de los modelos de ajuste clásico y ARIMA respecto a los datos de vientos, en la componente meridional. En rojo: marcadores de datos faltantes.

Se observa que los máximos de residuos en el caso de la componente zonal están más distribuidos y son de menor amplitud que en la componente meridional, donde se presentan localizados. Ambas componentes y ambos modelos presentan residuos notables en el mes de enero, esto podría indicar según el ajuste clásico un fenómeno no modelado por ondas de mareas, y de igual forma, en caso del aprendizaje automático, un evento diferente al comportamiento general de la serie.

En estas figuras se han incluido marcadores de períodos de datos faltantes. En ambas componentes y en ambos enfoques parte de los residuos muestran relación con los datos faltantes, pues se encuentran sobre y muy cercanos a los marcadores. Los datos faltantes se presume contaminarían las estimaciones dentro de la ventana de cálculo.

Se analizan a continuación aquellos máximos en el mapa de residuos que no están aparentemente relacionados con datos faltantes. Se seleccionaron entre éstos, los meses de octubre y enero para la componente zonal, representados en la figura 7.23 y 7.24 y sólo el mes de enero para la componente meridional, presentado en la figura 7.25.

En la figura 7.23 se agregan marcadores en color negro para los máximos sobre el mapa de residuos en octubre y enero (arriba), el de promedios diarios de dato (medio) y en el promedio diario del ajuste clásico (abajo). Se

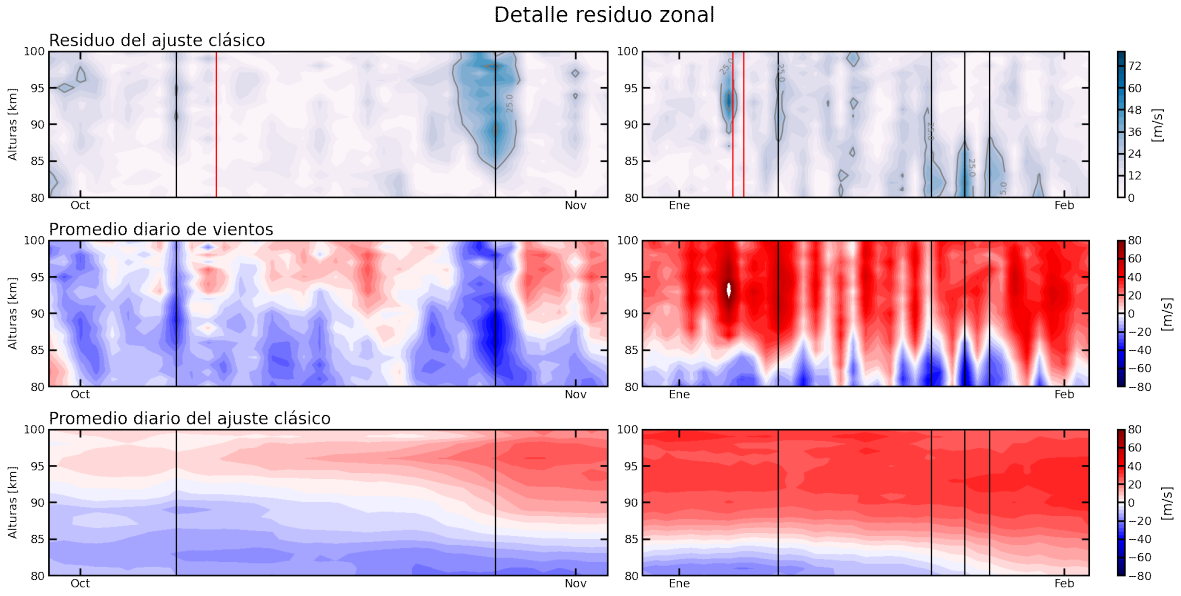

Figura 7.23: Promedios diarios de los datos de vientos (medio) y de los vientos medios del ajuste clásico (abajo) y la diferencia entre ambos (arriba), para los meses de octubre (izquierda) y enero (derecha), en la componente zonal.

identifican rápidamente en los mapas del dato muestreado aquellos vientos máximos que no son representados en la tendencia media del ajuste. Para

octubre se registran dos máximos de vientos hacia el oeste. En enero, se observa un fuerte máximo hacia el este por encima de los 90 km y tres hacia el oeste por debajo de los 90 km.

En la figura 7.24 se realiza un análisis similar, para los resultados del enfoque de aprendizaje automático. Los máximos sobre octubre persisten,

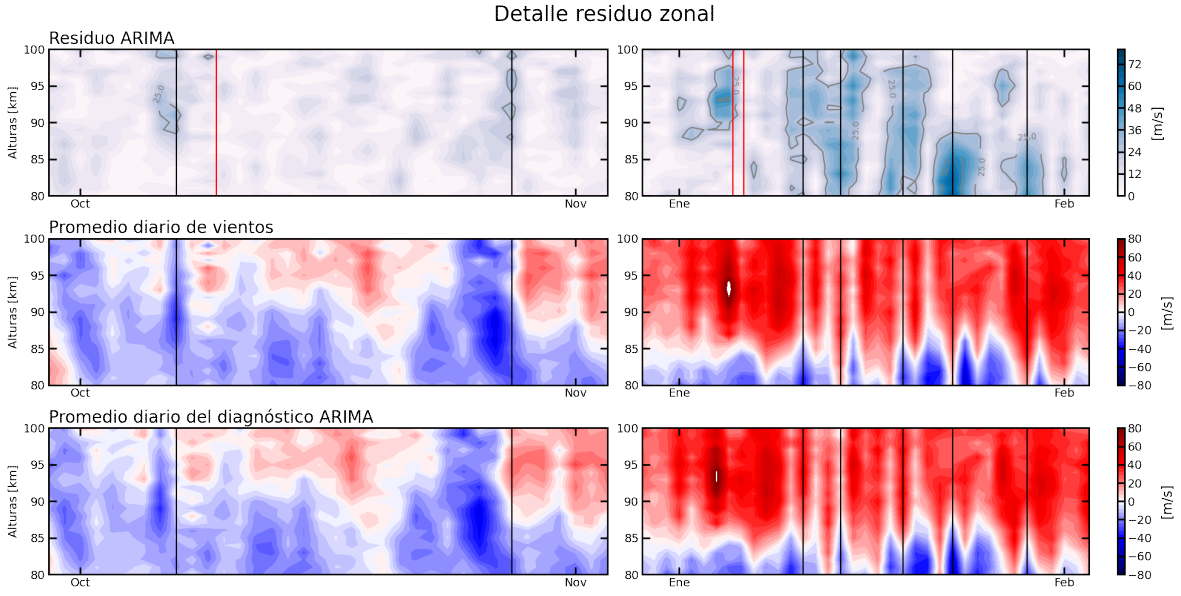

Figura 7.24: Promedios diarios de los datos de vientos (medio) y de los vientos ajustados por ARIMA (abajo) y la diferencia entre ambos (arriba), para los meses de octubre (izquierda) y enero (derecha), en la componente zonal.

mientras que en enero la misma estructura de máximos consecutivos hacia el oeste produce residuos, en este caso, corridos en tiempo a puntos previos respecto de los del ajuste clásico. Se observa en la comparación de la figura 7.24 que aunque similar, el mapa resultante del diagnóstico ARIMA se encuentra desfasado con respecto al dato. Este desfasaje y el valor destacado de esos máximos producirían el máximo en el residuo.

Como se mostró en la figura 7.22, no se identificaron máximos sobre octubre para la componente meridional. Se considera entonces que los residuos de octubre, suponen un efecto concentrado en aquella, la dirección zonal, y hacia el oeste. Para la componente meridional, se muestra en la figura 7.25 los residuos de enero para el enfoque clásico (primera columna) y para el diagnóstico ARIMA (en la segunda columna).

El evento presente en enero en la componente zonal y que se aparta del comportamiento regular, está presente también en la componente meridional, de manera que no esta arbitrariamente comportado sobre una dirección. La tendencia a valores medios del ajuste clásico evidencia el residuo del mismo, sobre la misma fecha. De nuevo, sobre el ajuste de aprendizaje automático,

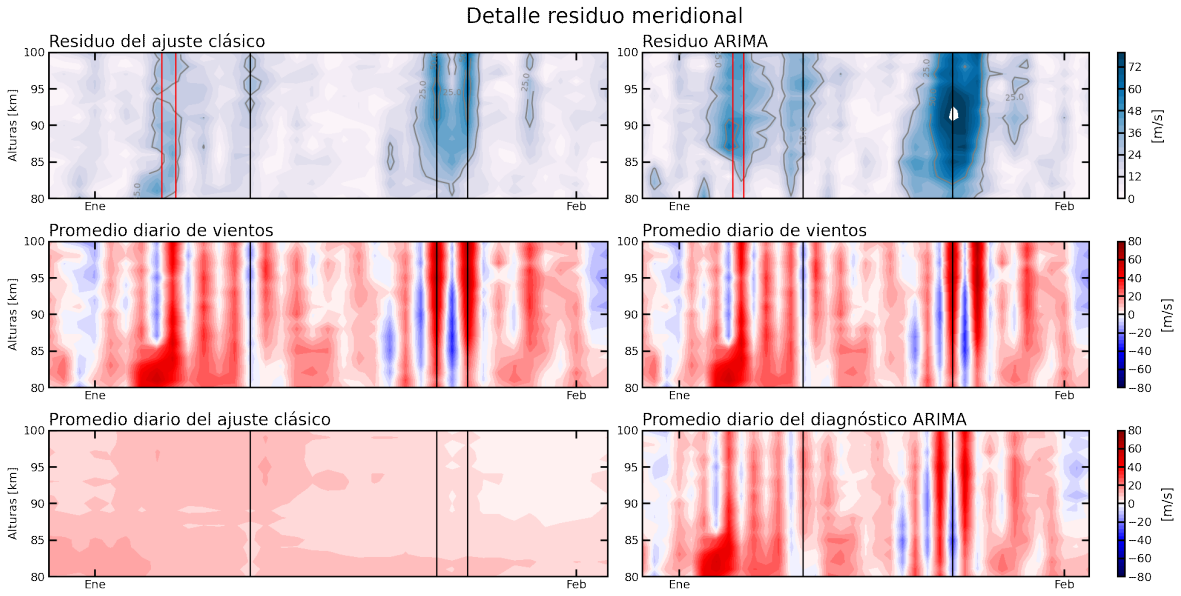

Figura 7.25: En la primera columna se muestran promedios diarios de diferencias del ajuste clásico para con el dato (arriba), del dato (medio) y del viento medio (abajo). En la segunda columna se muestran promedios diarios de diferencias del modelo ARIMA para con el dato (arriba), del dato (medio) y del viento ajustado por ARIMA (abajo. Todos los mapas se limitan al mes de enero en la componente meridional.

el responsable del máximo en el residuo es el aparente desfasaje en el ajuste, derivado de la misma metodología de ajuste.

Se concluye que en ambas componentes, el ajuste clásico representa la tendencia de los vientos exceptuando estos apartamientos particulares. El ajuste de aprendizaje automático permite una aproximación mas exacta incluyendo el evento de enero. Sin embargo, no es tan preciso como exacto, ya que es capaz de representarlo con un cierto desfasaje respecto de la muestra de entrenamiento.

8 Conclusiones y trabajos a futuro

En el presente trabajo, se logró obtener información de viento medio y perturbaciones de mareas en la mesosfera y baja termosfera ($\sim$ 80 a 100 km de altura) sobre la zona patagónica argentina. La variabilidad de los vientos medios en la componente zonal muestra una inversión en la dirección del viento de oeste a este que tiene lugar en la transición invierno-verano y los primeros meses de verano, y que se completa en el mes de febrero. La componente meridional muestra dos máximos de amplitud en los meses de verano y un cambio en el viento hacia el sur en febrero. En general, en esta componente ocurren variaciones más intensas que en la anterior componente zonal.

Además, de las componentes de la perturbación de marea, que pueden obtenerse a partir del ajuste clásico en amplitud y fase, la semidiurna de origen térmico es la que presenta amplitudes mayores. Las amplitudes diurnas y semidiurnas lunares apenas alcanzan la mitad de la amplitud de la semidiurna solar en ambas componentes, zonal y meridional. Particularmente, la marea semidiurna solar presenta un fuerte máximo durante fines de septiembre y principios de octubre, época posterior al SSW mayor que tuvo lugar en el hemisferio sur el 18 de septiembre de 2019. Se observó que la fase de esta componente permanecía constante en determinados rangos de alturas, por lo menos, desde septiembre hasta enero.

Por otro lado, la aplicación del modelo ARIMA constituyó un gran desafío en el tipo de series temporales que representan estos datos de vientos, por la marcada estacionalidad que repercutió en la condición de estacionariedad necesaria para el método de aprendizaje automático.

Un punto de especial atención para esta tesis ha sido que la aplicación de las técnicas de minería de datos y aprendizaje automático requieren un análisis previo cuidadoso de los movimientos o componentes de variación presentes en la serie temporal debido a que los modelos de tipo ARIMA son especialmente sensibles a los parámetros.

Otro punto que requirió un cuidadoso análisis han sido los datos faltantes, los cuales pueden afectar la evaluación de estacionariedad. En intervalos donde no se cuenta con suficiente información sobre la serie, y por lo tanto, sobre los movimientos que la componen, la evaluación de estacionariedad puede arrojar resultados poco útiles, especialmente, cuando esta circunstancia ocurre en el proceso de evaluación-diferenciación que se aplica con

el fin de obtener series estacionarias equivalentes a las originales. En este tratamiento que se le aplica a la serie, previamente a la identificación del modelo, se puede producir pérdida de información si no se cuenta con un test de raíces unitarias completo, o si la serie presenta marcada estacionalidad. En el primer caso, una evaluación deficiente de la estacionariedad lleva a incurrir en un sobre-procesamiento de la serie que genera gran cantidad de datos faltantes y puede acabar en la determinación de una serie estacionaria cuando en realidad se alcanza una serie con poca información sobre la variabilidad. En el segundo caso, la estacionalidad puede producir una clasificación de no estacionariedad. De la misma forma, se procesa innecesariamente la serie cuando solo se requeriría modelar, o remover dependiendo del caso, la estacionalidad. En resumen, con el propósito de alcanzar la adecuada estacionariedad se puede incurrir en la introducción de gran cantidad de datos faltantes, y por consiguiente, arribar a estacionariedad por la pérdida de información. Los modelos ARIMA no podrán predecir correctamente la serie fuera del rango de entrenamiento si no se modelan bien las componentes de variación.

Se pudo comprobar que las series temporales analizadas no se presentaban en primera instancia como series estacionarias y requirieron ser diferenciadas para alcanzar estacionariedad de segundo orden, la cual resulta suficiente para la utilización de esta técnica en general. Aceptar esta hipótesis implica asumir entre otras cosas, que las series temporales de altura fija, poseen varianza constante.

La identificación del modelo ha presentado modelos ARIMA de mayor orden en la componente meridional que zonal. En la componente zonal predominó el modelo ARIMA(0,1,2) sin valor contante, por el contrario en la componente meridional, el modelo mas representativo resultó ser el ARIMA(3,1,3) con término de valor medio. Ambas componentes presentan modelos generalmente similares por debajo de los 90 km y son mas bien diferentes por encima de esta altitud.

La validación del modelo mostró que la estacionalidad de los datos influye en mayor medida en el pronóstico. Dentro de la muestra de entrenamiento, la no contemplación de la componente estacional resulta en patrones periódicos en el correlograma de los residuos, y en un residuo no normal. Incluso con esta componente no modelada las estimaciones dentro de la muestra son exactas para la gran mayoría de las alturas. De manera que, los modelos ARIMA fueron capaces de representar gran parte de la información y de igual forma pudo realizarse una aproximación de la serie. En lo que respecta al pronóstico sobre el intervalo de testeo, se comprobó que, la selección de los rangos de entrenamiento y testeo son importantes para el éxito de este paso, pero resulta aún más importante y un punto clave que el modelo abarque las componentes estacionales. En este caso, no fue posible representar la muestra de testeo solo con el pronóstico del modelo, pero

mediante la actualización muestra a muestra, pudo aproximarse la serie en este intervalo.

Además, en este paso se pudo observar que esencialmente la metodología regresiva puede producir altos valores de error absoluto. Esto se observó principalmente en muestras que cambian de signo abruptamente y en muestras cercanas a puntos de tiempo con datos faltantes.

De la validación del modelo en conjunto, ha resultado claro que un modelo sencillo como ARIMA, para este caso de series temporales, puede resultar útil para la descripción de la muestra, o el llamado análisis In-Sample, pero para que pueda utilizarse para predecir la serie debe complementarse con otras prácticas como el modelado de la estacionalidad, y sobre todo, el modelado de la dispersión que presenta señales de no ser constante.

Cuando se contrastaron ambas estrategias se concluyó que, el modelo de aprendizaje automático de carácter regresivo, ARIMA, aproximó de forma más exacta la serie de vientos, sobre el rango de entrenamiento, pero sin distinguir la información de mareas. En cambio, el método clásico de análisis permitió obtener estimaciones de viento medio y amplitud de mareas por separado, teniendo en cuenta que se conocía el modelo de vientos. La limitación de este último modelo ha sido la longitud necesaria de la ventana de ajuste cuando se quieren obtener estimaciones de mareas semidiurnas de distinto origen y de períodos aproximados. La principal ventaja de este método reside en que los datos faltantes no son un gran problema en esta aplicación. Por el contrario, las principales limitaciones del modelo de aprendizaje automático han sido la condición de estacionariedad y los datos faltantes sobre la serie temporal. Asimismo, si bien se requiere un estudio de estacionariedad, una vez realizado este paso, es posible aproximar la serie temporal prescindiendo del modelo físico que la explica, siendo esta una de las ventajas de esta estrategia.

Por último, si se desea aplicar una variante del modelo ARIMA para este tipo de dato, es oportuno comentar que series temporales con más de una componente estacional son difíciles de modelar con esta técnica. Los modelos ARIMA-estacionales (SARIMA) requieren una adecuada elección de la multiplicidad que marca los períodos, siendo uno de los parámetros más sensibles de esta variante del modelo. Las series temporales de este análisis presentan más de una componente estacional, por lo cual, la elección de una única multiplicidad presentará dificultades en la representación de la información. Aún así, es claro que analizar la estacionalidad es el siguiente paso de interés en la técnica de aprendizaje automático.

En el modelo ARIMA se ha supuesto que la varianza es constante en cada serie temporal de altura fija. Luego, otra aplicación que sería interesante explorar sobre estos datos son aquellas técnicas que consideran la heterocedasticidad condicional, es decir, una variabilidad condicional no constante

a lo largo de la serie. Modelos como el de heterocedasticidad condicional autoregresiva (ARCH), y ARCH-generalizado (GARCH) permiten modelar la dispersión en este sentido.

Técnicas más complejas y sofisticadas de minería de datos y aprendizaje automático podrían utilizarse para modelar las diferentes componentes presentes en el dato. Un ejemplo de esto sería la técnica Prophet, que además permite modelar eventos inusuales, denominados componentes holidays en esta técnica.

Por otra parte, el modelo ARIMA de carácter regresivo podría ser una buena opción para el relleno o modelado de datos faltantes. Esto podría resultar una ventaja para aplicar posteriormente otra técnica de aprendizaje automático sensible a un muestreo regular.

Bibliografía

Adhikari R., Agrawal R. K. (2013). An Introductory Study on Time Series Modeling and Forecasting.

Alturi G., Karpatne A., Vipin K. (2018). Spatio-Temporal Data Mining: A Survey of Problems and Methods. *ACM Comput. Surv. 51, 4*, Art. 83. https://doi.org/10.1145/3161602

Andrienko N. y Andrienko G. (2006). Exploratory Analysis of Spatial and Temporal Data. A systematic Approach. *Springer-Verlag Berlin Heidelberg*, Germany.

Box G. E. P., Jenkins G. M., Reinsel G. C., Ljung G. M. (2015). Time Series Analysis. Forecasting and Control. 5 Edición. *Published by John Wiley Sons, Inc.*, Hoboken, New Jersey.

Box G. E. P. y Pierce D.A. (1970). Distribution of Residual Autocorrelations in Autoregressive Integrated Moving Average Time Series Models. *Journal of the American Statistical Association, 65*, 1509-1526.

Conte, J. F., Chau J. L., Stober G., Pedatella N., Maute A., Hoffmann P., Janches D., Fritts D., y Murphy D. J. (2017). Climatology of semidiurnal lunar and solar tides at middle and high latitudes: Interhemispheric comparison. *J. Geophys. Res. Space Physics, 122*. https://doi.org/10.1002/2017JA024396.

Conte, J. F., Chau, J. L., Peters, D. H. W. (2019). Middle- and high-latitude mesosphere and lower thermosphere mean winds and tides in response to strong polar-night jet oscillations. *J. Geophys. Res., 124*, 9262-9276. https://doi.org/10.1029/2019JD030828.

Conte J. F., Chau J. L., Urco J. M., Latteck R., Vierinen J. y Salvador J. O. First studies of mesosphere and lower thermosphere dynamics using a multistatic specular meteor radar network over southern Patagonia. *Earth and Space Science, 8*. doi:10.1029/2020EA001356, 2021.

Chau, J. L., Urco, J. M., Vierinen, J. P., Volz, R. A., Clahsen, M., Pfeffer, N., Trautner, J. (2019). Novel specular meteor radar systems using coherent MIMO techniques to study the mesosphere and lower thermosphere. *Atmospheric Measurement Techniques, 12(4)*, 2113–2127. https://doi.org/10.5194/amt-12-2113-2019

Chau J. L., Urco J. M., Vierinen J., Harding B. J., Clahsen M., Pfeffer N., Kuyeng K., Milla M. y Erickson P. J. (2021). Multistatic specular meteor radar network in Peru: System description and initial results. *Earth and Space Science, 8*, https://doi.org/10.1029/2020EA001293.

Clemesha, B. R., Batista, P. P., Buriti da Costa, R. A., Schuch, N. (2009). Seasonal variations in gravity wave activity at three locations in Brazil. *Annales Geophysicae, 27*, 1059–1065. https://doi.org/10.5194/angeo-27-1059-2009

Dickey D. A. y Fuller W. A. (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root, *Journal of the American Statistical Association*, 74, 427-431.

Gujarati D., Porter D. C., (2010). Econometría. 5 Edicion. McGraw-Hill Interamericana Editores, S. A. C.V.

Hagan, M. E., y J. M. Forbes (2002). Migrating and nonmigrating diurnal tides in the middle and upper atmosphere excited by tropospheric latent heat release, *J. Geophys. Res.*, 107 (D24), 4754, doi.org/10.1029/2001JD001236.

Hagan, M. E. y Forbes J. M. (2003). Migrating and nonmigrating semidiurnal tides in the upper atmosphere excited by tropospheric latent heat release, *J. Geophys. Res.*, 108(A2), 1062, https://doi.org/10.1029/2002JA009466

Han J., Kamber M. y Pei J. (2012). Data Mining. Concepts and techniques. 3 Edición. Morgan Kaufmann Publishers.

Hocking, W. K. (2005). A new approach to momentum flux determinations using SKiYMET meteor radars. *Annales Geophysicae*, 23(7), 2433–2439. https://doi.org/10.5194/angeo-23-2433-2005

Hoffmann, P., Becker, E., Singer, W. y Placke, M. (2010). Seasonal variation of mesospheric waves at northern middle and high latitudes. *Journal of Atmospheric and Solar - Terrestrial Physics*, 72 (14-15), 1068–1079. https://doi.org/10.1016/j.jastp.2010.07.002

Holton, J. R. (2004). An introduction to dynamic meteorology. 4 Edición. *El-sevier Academic Press*, San Diego, California.

Jarque C.M. y Bera A.K. (1987). A Test for Normality of Observations and Regression Residuals, *International Statistical Review*, 55, 163-172. Jia, M., Xue, X., Gu, S., Chen, T., Ning, B., Wu, J. (2018). Multiyear observations of gravity wave momentum fluxes in the midlatitude mesosphere and lower thermosphere region by meteor radar. *J. Geophys. Res.: Space Physics*, 123, 5684–5703. https://doi.org/10.1029/2018JA025285

Jones, J., Webster, A. R. y Hocking, W. K. (1998). An improved interferometer design for use with meteor radars. *Radio Science*, 33, 55–65. https://doi.org/10.1029/97RS03050

Kwiatkowski, D., P.C.B. Phillips, P. Schmidt, Y. Shin (1992). Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root, *Journal of Econometrics*, 54, 159-178.

Laskar, F. I., Chau, J. L., Stober, G., Hoffmann, P., Hall, C. M., Tsutsumi, M. (2016). Quasi-biennial oscillation modulation of the middle- and high-latitude mesospheric semidiurnal tides during August–September. *J. Geophys. Res.: Space Physics*, 121(5), 4869–4879. https://doi.org/10.1002/2015JA022065

Liu, A. Z., Lu, X. y Franke, S. J. (2013). Diurnal variation of gravity

wave momentum flux and its forcing on the diurnal tide. *J. Geophys. Res.: Atmospheres*, 118, 1668–1678.

Ljung G.M.y Box G.P.E. (1978). On a Measure of Lack of Fit in Time Series Models. *Biometrika*, 66, 66-72.

Lindzen Richard A. (1990). Dynamics in Atmospheric Physics *Cambridge University Press*. https://doi.org/10.1017/CBO9780511608285

Orallo Hernandez J. (2004). Introducción a la minería de datos. 1 Edición. *Alhambra, S. A.*, Madrid, España.

Phillips P. C. B. y Perron P. (1988). Testing for a Unit Root in Time Series Regression. *Biometrika*, 75, 335-346.

Said, S. E., Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. Biometrica, 71(3), 599-607. (ADF)

Sandford, D. J., H. G. Muller, and N. J. Mitchell (2006), Observations of lunar tides in the mesosphere and lower thermosphere at Arctic and middle latitudes, Atmos. Chem. Phys., 6(12), 4117–4127.

Schneider H., Chau J. L. y Stober J. (2016). Seasonal variation and short-term variability during SSWs of the gravity wave momentum flux. Rostock University, Rostock, Germany.

Stening, R. J. y Jacobi, C. (2001). Lunar tidal winds in the upper atmosphere over Collm, *Ann. Geophys: Atmos. Hydrospheres Space Sci.*, 18, 1645–1650. https://doi.org/10.1007/s005850000310

Stening, R. J. y Vincent, R. A. (1989). A Measurement Of Lunar Tides In The Mesosphere At Adelaide, South-Australia. *J. Geophys. Res: Space Physics.*, 94, 10121–10129. https://doi.org/10.1029/JA094iA08p10121

Trinh, Q. T., Ern, M., Doornbos, E., Preusse, P. y Riese, M. (2018). Characteristics of the quiet-time hot spot gravity waves observed by GOCE over the Southern Andes on 5 July 2010. *Ann. Geophys.*, 36, 425-444. https://doi.org/10.5194/angeo-36-425-2018.

Urco, J. M., Chau, J. L., Weber, T. y Latteck, R. (2019). Enhancing the spatio-temporal features of polar mesosphere summer echoes using coherent MIMO and radar imaging at MAARSY. *Atmospheric Measurement Techniques*, 12, 955–969. https://doi.org/10.5194/ amt-12-955-2019

Vierinen, J., Chau, J. L., Pfeffer, N., Clahsen, M. y Stober, G. (2016). Coded continuous wave meteor radar. *Atmospheric Measurement Techniques*, 9(2), 829–839. https://doi.org/10.5194/amt-9-829-2016

Vierinen, J., Chau, J. L., Charuvil, H., Urco, J. M., Clahsen, M., Avsarkisov, V. (2019). Observing mesospheric turbulence with specular meteor radars: A novel method for estimating second-order statistics of wind velocity. *Earth and Space Science*, 6, 1171–1195. https://doi.org/10.1029/2019EA000570

Vincent, R.A. (2015). The dynamics of the mesosphere and lower thermosphere: a brief review. *Prog. in Earth and Planet. Sci.* 2, 4 https://doi.org/10.1186/s40645-015-0035-8

Wald A. (1943). Tests of Statistical Hypotheses Concerning Several Parame-

ters When the Number of Observations is Large. *Transactions of the American Mathematical Society* 54, 3, 426-482. https://doi.org/10.2307/1990256

Winch, D. E. y Cunningham, R. A. (1972). Lunar Magnetic Tides at Watheroo: Seasonal, Elliptic, Evectional, Variational and Nodal, *J. Geomagnetism and Geoelectricity*, 24, 381–414.