Universidad Autónoma de Madrid
Escuela politécnica superior

Doble Grado en Ingeniería Informática y Matemáticas

Trabajo Fin de Grado

USO DE ENSEMBLES DE METEOROLOGÍA PARA PREDICCIÓN DE ENERGÍA EÓLICA

Autor: Ricardo García García
Tutor: José Ramón Dorronsoro Íbero

MAYO 2017

USO DE ENSEMBLES DE METEOROLOGÍA PARA PREDICCIÓN DE ENERGÍA EÓLICA

Autor: Ricardo García García
Tutor: José Ramón Dorronsoro Íbero

Dpto. de Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
MAYO 2017

Resumen

Resumen

Desde comienzos de siglo ha tenido lugar un enorme desarrollo de la energía renovable, la cual supuso en 2016 el 40,8 % del total de la energía producida en España, y todo hace indicar que seguirá siendo así en el futuro. No obstante, las energías de este tipo, precisamente por su carácter renovable, se basan en fuentes poco fiables e irregulares, como el viento o el sol. Es por ello que se ha vuelto de vital importancia para el sector eléctrico conocer de antemano la producción de energía proveniente de estas fuentes, a fin de poder ajustar lo mejor posible la producción total a la demanda.

En este trabajo nos hemos fijado exclusivamente en la energía eólica, partiendo de lo que ya se había realizado previamente en el IIC - Instituto de Ingeniería del Conocimiento. El proceso de predicción de la producción de este tipo de energía requiere de datos meteorológicos sobre los que aplicar los métodos de aprendizaje automático. No obstante estos datos son ya de por sí predicciones meteorológicas, lo cual añade una dificultad extra al problema al no ser totalmente exactas. Una posible solución a este problema ha sido la utilización de varios pronósticos meteorológicos diferentes, en lugar de uno solo, lo cual se conoce como Ensemble Forecasting. Para ello nos hemos basado en los datos del European Centre for Medium-Range Weather Forecasts (ECMWF) el cual, además de la predicción meteorológica de siempre, nos ofrece un servicio conocido como Ensemble Prediction System (EPS) formado por un total de 51 predicciones con las que trabajar.

El objetivo de este TFG ha sido pues el desarrollo de un sistema de predicción de la energía eólica basado en dicho EPS, en lugar de en una única predicción meteorológica como hasta ahora. Este nuevo enfoque nos ha permitido mejorar los resultados respecto a grandes errores, si bien no son tan buenos como los que obteníamos con una sola predicción meteorológica. Como algoritmos de aprendizaje automático hemos seleccionado Support Vector Machines y Multilayer Perceptron, lo cual no quita que quizá pudiéramos haber obtenido mejores resultados mediante la aplicación de otros modelos como Random Forest o Gradient Boosting.

Palabras Clave

Energía eólica, Datos de meteorología, Ensemble Forecasting, Aprendizaje automático.

Abstract

Since the beginnings of the century there has been a huge growth of the renewable energies, which in 2016 reached a 40.8 % of the total energy produced in Spain, and everything indicates that it will keep being like that in the future. Nevertheless, the energies of this kind, precisely because of their renewable character, are based on unreliable and irregular sources such as the wind or sun. So that, it has become vitally important for the electricity sector to know in advance what the production of energy from these sources will be, in order to adjust the total production to the demand in the best way.

In this work we have focused on the production of wind energy, starting from what had been previously done in the IIC - Institute of Knowledge Engineering. The process for predicting the generation of this type of energy requires meteorological data in order to apply machine learning methods. However, these data are predictions by themselves, which adds an extra difficulty to the problem since they are not totally accurate. A possible solution to this problem has been the use of several different forecasts instead of only one, what it is known as Ensemble Forecasting. For that we have used the data from the European Center for Medium-Range Weather Forecasts (ECMWF) which offers us, in addition to the usual weather forecast, a service known as Ensemble Prediction System (EPS), consisting of a total of 51 predictions with which to work.

The objective of this TFG has been the development of a wind power prediction system based on this EPS, instead of on a single meteorological forecast as before. This new approach has allowed us to obtain more robust results against large deviations, although these are not as good as the ones we already had. The machine learning algorithms chosen for this work have been Support Vector Machines and Multilayer Perceptron, even though we could have achieved better results by applying other models such as Random Forest or Gradient Boosting.

Key words

Wind power, Meteorological data, Ensemble Forecasting, Machine learning.

Agradecimientos

En primer lugar me gustaría dar las gracias al IIC y en particular a José y Julia por permitirme desarrollar este trabajo y haberme ayudado en todo momento.

Así mismo, quisiera agradecer a mis compañeros de carrera el haber estado ahí durante toda la carrera y haber creado un grupo tan bueno porque sin ellos (y sin sus apuntes) seguramente no estaría escribiendo este trabajo. Y en especial a Cris, gracias a quien estos últimos años han sido mucho mejores.

Por supuesto también a mis padres y a mi familia en general, que siempre me ha animado a estudiar y mirar adelante, y que de no ser por ellos nunca hubiera llegado tan lejos.

Por último me gustaría agradecérselo a todos aquellos que no he nombrado pero que de una forma u otra me han apoyado durante estos años.

Índice general

Índice de Figuras

Índice de Tablas

Glosario de acrónimos

- **IIC**: Instituto de Ingeniería del Conocimiento
- **REE**: Red Eléctrica de España
- **EPS**: Ensemble Prediction System
- **MARS**: Meteorological Archival and Retrieval System
- **ECMWF**: European Centre for Medium-Range Weather Forecasts
- **GEFS**: Global Ensemble Forecast System
- **NCEP**: National Centers for Environmental Prediction
- **HRES**: High Resolution Forecast
- **NMAE**: Normalized Mean Absolute Error
- **NMBE**: Normalized Mean Bias Error
- **NRMSE**: Normalized Root Mean Squared Error
- **ML**: Machine Learning
- **SVM**: Support Vector Machines
- **RBF**: Radial Basis Function
- **MLP**: Multilayer Perceptron
- **RF**: Random Forest
- **GB**: Gradient Boosting

1 Introducción

1.1. Motivación del proyecto

El incremento del consumo eléctrico, junto con la disminución de las fuentes de energía tradicionales, ha vuelto imprescindible la búsqueda de nuevas fuentes de energía alternativas y sostenibles. Así pues, desde comienzos de siglo se ha producido una gran expansión de las energías renovables, las cuales llegaron a suponer en 2016 un 40,8 % del total de la energía generada en España. Además, a pesar de que en los últimos años se ha producido un parón en su implantación, en gran parte provocado por los cambios en las políticas de ayudas al respecto, se espera que este porcentaje siga aumentando a fin de cumplir con las directrices marcadas por la Unión Europea. No obstante, las energías de este tipo suelen tener su origen en fuentes poco fiables e irregulares, como el viento o el sol. Si a esta irregularidad le sumamos el alto coste que supone su almacenamiento vemos que es necesario mantener fuentes tradicionales como soporte en caso de escasez de producción.

Debido a que las compañías eléctricas deben cubrir siempre la demanda, en general se produce una cantidad de energía superior a la que se demanda. Este excedente supone importantes pérdidas para el sector, para el que es de vital relevancia poder ajustar al máximo la producción a la demanda. Es por ello que ha aumentado notablemente el interés en conocer de antemano la producción de energía proveniente de estas fuentes irregulares, a fin de poder cubrir la demanda restante sin llegar a caer en la sobreproducción.

Por otro lado, un aumento en la producción de energía proveniente de estas fuentes implica una bajada en el precio de la energía y viceversa. Esto es importantísimo de cara a predecir y fijar los precios en el mercado eléctrico español, en el cual muchas empresas, principalmente las comercializadoras de energía, buscan conseguir la mejor oferta.

Este trabajo se ha llevado a cabo dentro del grupo de Salud y Energía del IIC - Instituto de Ingeniería del Conocimiento - y en él nos hemos fijado únicamente en la generación de energía eólica, la cual representa en la actualidad aproximadamente el 19 % de la energía producida en España. Es decir, en torno a la mitad de la energía proveniente de fuentes renovables[1].

[1] Estos porcentajes están sujetos a fuertes variaciones según el mes o la época del año.

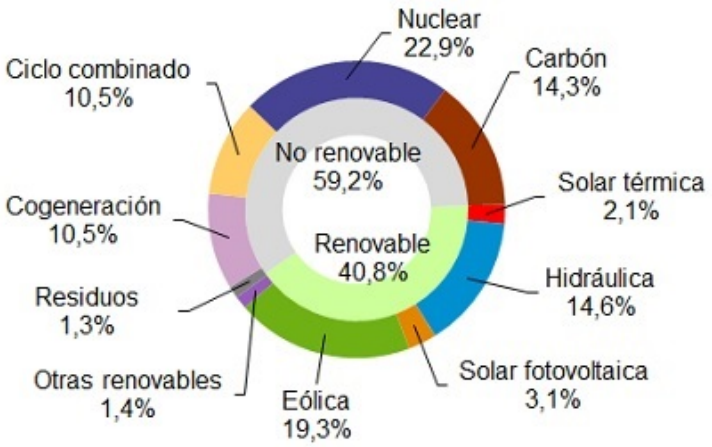

Figura 1.1: Producción de energía según su origen, en España y 2016. [1]

.

La importancia de la energía eólica queda reflejada también en la siguiente gráfica, donde se aprecia el efecto de la Ley de la Oferta y la Demanda en el mercado eléctrico español. En ella se observa cómo al aumentar la producción de energía eólica (con respecto a la demanda) hacen falta menos fuentes tradicionales, por lo que el precio de la electricidad disminuye.

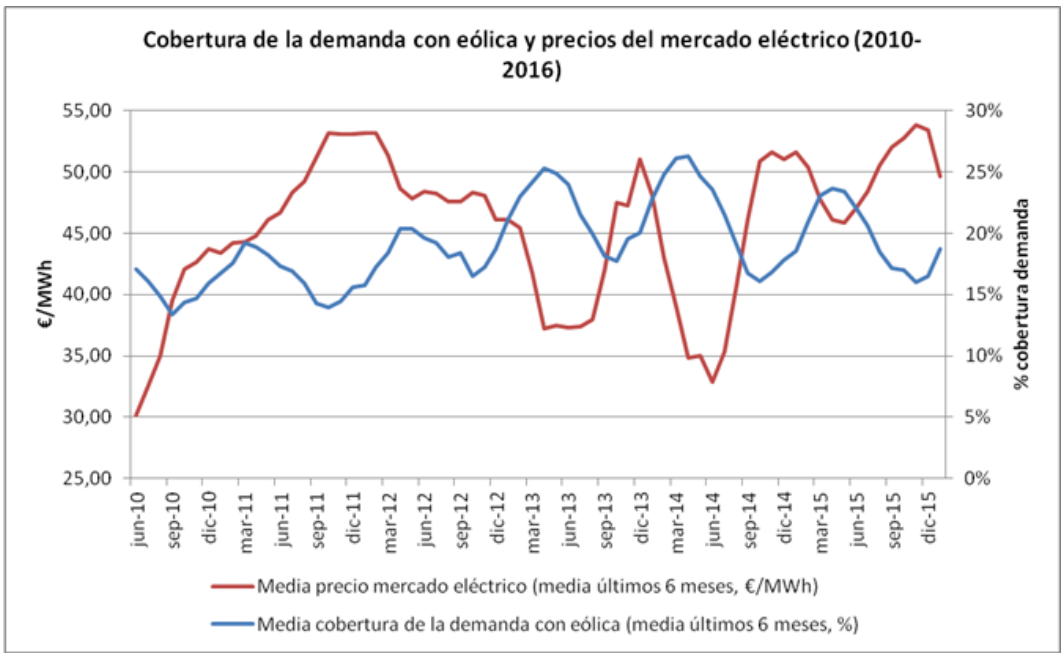

Figura 1.2: Relación entre la producción de energía eólica y el precio de la electricidad. [2]

Así pues, predecir la generación de energía eólica de la forma más ajustada posible es un tema crucial para los diversos agentes del sector eléctrico, por lo que nuestro objetivo ha sido mejorar dichas predicciones.

1.2. Objetivos y enfoque

Como hemos visto, el objetivo de este TFG ha sido desde un primer momento la mejora de las predicciones de producción de energía eólica. Al igual que en cualquier otro problema de aprendizaje automático teníamos dos posibles vías de mejora: perfeccionamiento de los modelos o mejora de los datos de partida.

Puesto que los modelos ya habían sido estudiados previamente, nos hemos centrado en la mejora de los datos, los cuales son ya de por sí predicciones meteorológicas. Más específicamente, en las pruebas de un nuevo sistema de predicción meteorológica, conocido como Ensemble Prediction System (EPS) y formado por un conjunto de predicciones más o menos fiables, en lugar de una sola.

Así pues, se han fijado como objetivos del trabajo:

- La adaptación del sistema de predicción de energía eólica actual para su posible ejecución con la nueva fuente de datos.
- La ejecución y las pruebas de dicho sistema con las nuevas predicciones de meteorología como datos de partida.
- Desarrollo e implantación de un método de evaluación de las predicciones obtenidas.

1.3. Estructura del documento

Esta memoria consta de seis capítulos:

- Capítulo 1: Motivación y objetivos del proyecto.
- Capítulo 2: Introducción al proceso de predicción de la energía eólica, los datos de partida, los algoritmos de aprendizaje automático y las medidas de error más utilizadas.
- Capítulo 3: Breve exposición sobre el Ensemble Forecasting, seguido del análisis estadístico de los datos de meteorología del EPS.
- Capítulo 4: Exposición detallada de los algoritmos seleccionados para la regresión, Support Vector Machines y Multilayer Perceptron.
- Capítulo 5: Descripción de los experimentos realizados y de los resultados obtenidos para cada uno de ellos.
- Capítulo 6: Visión global del trabajo llevado a cabo, conclusiones y posibles vías de mejora aún por explorar.

Y dos anexos:

- Anexo A: Información relativa a la descarga de la meteorología y su conversión a un formato legible.
- Anexo B: Tablas y gráficas con los resultados de REE, omitidos en el capítulo 5 por su similitud con los de Sotavento.

2

Predicción de energía eólica. Estado del arte

2.1. Introducción

El proceso para predecir la producción de energía eólica consiste principalmente en:

1. Descarga de los datos de meteorología y de las producciones.
2. Conversión de los datos a un formato legible.
3. Tratamiento y selección de los datos.
4. Aplicación de métodos de ML para la predicción de la producción.

Figura 2.1: Proceso predicción energía eólica.

Es importante notar que los datos de partida son ya de por sí predicciones meteorológicas, lo cual añade una dificultad extra al problema. Es decir, el proceso es en realidad el siguiente:

1. Observación de las condiciones iniciales de la atmósfera.
2. Aplicación de un modelo numérico para elaborar la predicción meteorológica.
3. Descarga de las predicciones de meteorología y de las producciones.
4. Conversión de los datos a un formato legible.
5. Tratamiento y selección de los datos.
6. Aplicación de métodos de ML para la predicción de la producción.

Figura 2.2: Proceso predicción energía eólica.

2.2. Historia, nacimiento y evolución

Conocer de antemano cuál va a ser la producción de energía eléctrica permite estudiar y analizar la evolución de los precios en el mercado eléctrico. Es por ello que existe un gran interés en conocer dicha producción por parte de todas las empresas que acuden al mercado en busca de la mejor oferta (principalmente productoras y comercializadoras de energía). Por lo tanto, es de entender que con el crecimiento de las energías renovables haya aumentado también el interés por predecir su producción, principalmente en el caso de la eólica y la solar.

Las energías renovables han introducido un factor aleatorio o desconocido en la generación de energía eléctrica, el cual antes no existía. Este factor es la meteorología, y si bien antes se podía intuir directamente cuál iba a ser el precio de la energía en base al precio de cada fuente utilizada, ahora es necesario conocer también cuál va a ser la producción obtenida a partir de dichas fuentes.

Así pues, la evolución de la predicción de energía eólica ha estado íntimamente ligada a la evolución de la propia energía eólica, de manera que conforme más relevancia adquiere con respecto al resto de fuentes mayor es el interés en la predicción de su producción.

2.3. Datos de meteorología

Existen diversas agencias y fuentes de las que obtener la meteorología, todas ellas con un distinto nivel de precisión en sus predicciones. La razón de ello se encuentra tanto en el origen de sus datos (diferentes satélites y estaciones de medida) como en los modelos matemáticos utilizados (distintos modelos con distintos parámetros).

Existen diversos parámetros que afectan a la bondad de las predicciones meteorológicas, entre los que se encuentran:

- Variables utilizadas: Cada variable meteorológica se obtiene de manera diferente, siendo más complicadas de obtener unas que otras. Así pues, las distintas componentes del viento o la temperatura pueden tener niveles de error diferentes.
- Resolución: Puesto que no se puede trabajar con el estado de la atmósfera en todos sus puntos, en su lugar se ha optado por elegir varios con una distancia determinada entre ellos. Esta distancia es lo que se conoce como resolución del modelo y en general se mide en grados, ya que se utilizan la latitud y la longitud (y no es lo mismo un grado en el Ecuador que en el Polo Norte al no ser el planeta completamente esférico). Por lo tanto, se debe definir una resolución vertical para la latitud, y otra horizontal para la longitud, que no obstante puede ser la misma.
- Distancia temporal entre predicciones: No es lo mismo predecir para cada hora, que para cada 3 o 6, siendo el nivel de detalle mayor cuanto menor es el salto entre las predicciones. No obstante, disminuir la longitud de estos saltos ralentiza notablemente la ejecución de

los modelos, por lo que no suele ser viable. Es por ello que se suelen tomar en torno a 3 horas como base para el futuro inmediato, que se van aumentando conforme nos alejamos en el tiempo.

- Horizonte: Se refiere a la distancia temporal entre el momento "en" el que se predice y el momento "para" el que se predice. Por ejemplo, si se realiza la predicción de forma trihoraria, el primer horizonte será la predicción para 3 horas después al momento en el que se ejecuta el modelo, el segundo para 6 horas después, y así sucesivamente. No obstante hay que tener cuidado ya que la distancia temporal entre los horizontes puede variar, siendo menor entre los primeros que entre los últimos. Esto se hace así debido a que requiere menos tiempo computacional predecir para el futuro cercano que para el lejano.

En cualquier caso, si seguimos el estudio de Lorenz [5, 6, 7] sobre meteorología sabemos que es imposible obtener una predicción meteorológica totalmente exacta, ya que existen diversos factores y limitaciones al respecto. A saber:

- Las ecuaciones utilizadas por cualquier modelo matemático no capturan todos los procesos de la atmósfera (véanse por ejemplo los efectos de la fricción o la propagación del calor).
- Los modelos trabajan con puntos a cierta distancia entre sí, de manera que no es posible capturar la información entre dichos puntos.
- Es imposible disponer de observaciones iniciales en todos los puntos del planeta.
- Los datos observados no pueden ser medidos con precisión infinita, siendo este error de precisión a la larga importante en los resultados del modelo.

Así pues existe una cierta incertidumbre en las condiciones iniciales de la atmósfera, la cual repercute directamente sobre la bondad de las predicciones meteorológicas. Este hecho es especialmente relevante en un sistema caótico como es la atmósfera, en el que cualquier ligera perturbación sobre las condiciones de partida puede derivar en importantes errores en poco tiempo. Este fenómeno, conocido popularmente como Efecto Mariposa, implica que se debe tener mucho cuidado con las conclusiones extraídas de los resultados de los modelos.

En la siguiente figura (2.3) podemos ver una representación de cómo afectan pequeños cambios en las condiciones iniciales a las predicciones obtenidas, aumentando la dispersión conforme nos alejamos del momento inicial. Además, atendiendo a las desviaciones finales de la gráfica podemos deducir que algunas de las predicciones son más probables que otras.

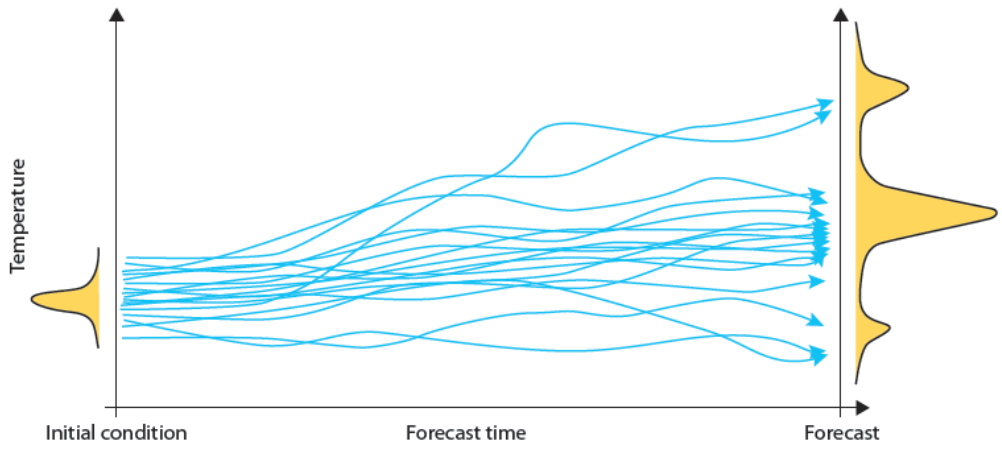

Figura 2.3: Dispersión de la predicción de la temperatura a lo largo del tiempo [3].

Debido a esto, durante los últimos años se ha buscado medir y cuantificar la fiabilidad de las predicciones meteorológicas. Una posible solución a este problema sería la aplicación de la Ecuación de Liouville [8, 9], que partiendo de una función de densidad de probabilidad del estado inicial tiene en cuenta su incertidumbre y nos da una función de densidad para la predicción o estado final (en lugar de una predicción sin más). No obstante, dicha ecuación es excesivamente compleja y requiere demasiado tiempo computacional como para poder ser utilizada en la práctica. En su lugar se ha optado por una aproximación más sencilla, conocida como Ensemble Forecasting y que se detallará en su propio capítulo.

2.4. Datos de producciones

Además de los datos de meteorología necesitamos otros datos propios de las plantas eólicas. Entre ellos se encuentran la localización, imprescindible para poder seleccionar los puntos de meteorología cercanos; la potencia máxima instalada, necesaria para poder trabajar directamente con las producciones normalizadas; y los datos históricos de producción, a fin de poder entrenar los modelos y comparar los resultados. Estos datos son en general de carácter privado y nos los entregan exclusivamente los dueños de las plantas para las que se predice. No obstante, existen algunas excepciones como el parque eólico Sotavento [10], cuyos datos se encuentran disponibles en su página web pudiéndose obtener mediante scraping.

La producción de los parques se mide en general en kW/h y se obtiene directamente del contador del propio parque. No obstante, puesto que normalmente se predice de forma horaria se toman como datos de producciones directamente las potencias medias de cada hora, medidas en kW. Por otro lado, las mediciones de los parques no son totalmente exactas sino que pueden verse afectadas por consignas. En tal caso los errores de medición se van acumulando hasta que se comprueban las diferencias, las cuales se introducen en la hora siguiente para compensar. Deberá comprobarse por tanto que la diferencia entre dos horas seguidas no sobrepase un cierto límite, ya que de haber consignas las mediciones serían engañosas y podrían conducir a error al modelo. Además, puesto que las producciones se obtienen por vías externas es factible que en caso de fallo del sistema o del contador haya datos no disponibles en un determinado momento. Estas indisponibilidades se deben tener en cuenta al entrenar el modelo.

2.5. Métodos y sistema de predicción

De cara a predecir la producción de energía eólica se utilizan todo tipo de algoritmos de aprendizaje supervisado. Puesto que las variables, los puntos de meteorología utilizados y los metaparámetros escogidos en los modelos afectan de manera importante a la predicción se utilizan en general diversas versiones de cada algoritmo, cuya eficacia puede variar a lo largo del tiempo. De esta manera se puede seleccionar cada cierto tiempo aquel que mejor se aproxime a la realidad. Es importante notar que el proceso de buscar y seleccionar los mejores parámetros es muy costoso computacionalmente, por lo que no puede llevarse a cabo constantemente.

Los algoritmos más utilizados en la industria son principalmente aquellos que llevan más tiempo implementados, como las Máquinas de Vectores de Soporte (SVM) o el Perceptrón Multicapa (MLP). No obstante nuevos algoritmos, como aquellos basados en Deep Learning o Aprendizaje Profundo, van haciéndose hueco poco a poco también en este sector.

2.6. Evaluación de los resultados

De cara a comprobar la calidad de las predicciones es necesario observar no solo el error cometido sino también su origen. Así pues, no es lo mismo sobreestimar la producción que subestimarla, pudiendo ser más peligroso lo primero (habría que corregir dicho error mediante otras fuentes o podría producirse un apagón). De la misma manera, no es lo mismo equivocarse para unas horas que para otras, siendo de mayor relevancia las horas del mercado eléctrico intradiario, en las cuales es imprescindible para las compañías poder disponer de la información lo más ajustada y actualizada posible [11]. Debemos tener en cuenta también si se trata de un error absoluto o relativo, ya que en períodos de alta producción un error relativo puede ser mayor que en períodos de baja producción.

Además, en el caso particular de la energía eólica o la solar nos basamos en predicciones meteorológicas como datos de partida. Esto implica que en períodos de gran variabilidad en la atmósfera el error en los datos de partida pueda ser muy alto y propagarse, obteniéndose así también un error importante en nuestras predicciones. Este hecho debe tenerse en cuenta también al observar el error cometido en un cierto parque, ya que éste podría ubicarse en lugares con gran variabilidad, por lo que los errores no son comparables con los del resto de parques.

2.6.1. Medidas del error cometido

En general se utilizan varias métricas diferentes para medir la bondad de los resultados, siendo la más utilizada seguramente el NMAE:

- Power deviation:
$$PD = \frac{\sum |Pred - Prod|}{\sum Prod} \tag{2.1}$$
- Mean Bias Error:
$$MBE = \frac{1}{N} \sum (Pred - Prod) \tag{2.2}$$
- Normalized Mean Bias Error:
$$NMBE = \frac{\frac{1}{N} \sum (Pred - Prod)}{Potencia} \tag{2.3}$$
- Mean Absolute Error:
$$MAE = \frac{1}{N} \sum |Pred - Prod| \tag{2.4}$$
- Normalized Mean Absolute Error:
$$NMAE = \frac{\frac{1}{N} \sum |Pred - Prod|}{Potencia} \tag{2.5}$$
- Root Mean Squared Error:
$$RMSE = \sqrt{\frac{\sum (Pred - Prod)^2}{N}} \tag{2.6}$$
- Normalizad Root Mean Squared Error:
$$NRMSE = \frac{\sqrt{\frac{\sum (Pred - Prod)^2}{N}}}{Potencia} \tag{2.7}$$

En las fórmulas anteriores Pred hace referencia a la predicción, Prod a la producción y Potencia a la potencia máxima instalada en cada uno de los parques.

3

Ensemble Forecasting

3.1. Idea general

Como ya hemos visto existe una cierta incertidumbre en los datos de partida de los modelos atmosféricos. Es por ello que se ha desarrollado un sistema de predicción alternativo, similar a los métodos de Montecarlo y conocido como Ensemble Forecasting. Éste consiste en generar un número discreto de posibles estados iniciales y predecir la meteorología para cada uno de ellos, con lo que se tienen varios pronósticos diferentes. De esta manera podemos estimar la función de probabilidad de nuestro estado final, obteniendo así un criterio de la fiabilidad de cada una de las predicciones individuales y pudiendo extraer conclusiones globales [12, 13, 14].

A continuación podemos apreciar con ejemplos reales la relevancia de utilizar varias predicciones en lugar de una sola.

En la primera gráfica (3.1) se muestra la predicción de la temperatura en Hamburgo para los 10 días siguientes al 17 de Octubre de 2008. Se aprecia como durante los primeros 3-4 días apenas hay cambios, de forma que todas las posibles predicciones se asemejan bastante. Sin embargo el quinto día tienen lugar cambios importantes en la atmósfera, con lo que las predicciones decaen considerablemente. En caso de tener en cuenta una única predicción como hasta ahora, utilizaríamos la supuesta más fiable, señalada en negro. No obstante, ésta tiene un error muy importante con respecto a la realidad, en verde.

En la siguiente imagen (figura 3.2) aparecen reflejados varios posibles recorridos del huracán Katrina para las 120 horas siguientes al momento de la ejecución, así como el que a posteriori se hizo realidad. Atendiendo a todos los posibles caminos se podía estimar la probabilidad de que el huracán pasara por un cierto punto (por ejemplo una ciudad), mientras que basándonos en uno solo hubiéramos perdido muchísima información de gran valor.

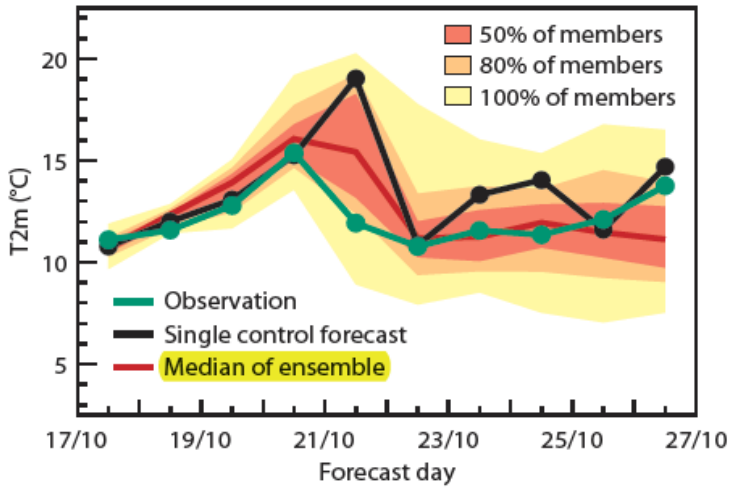

Figura 3.1: Comparación predicciones y medición real de la temperatura [3].

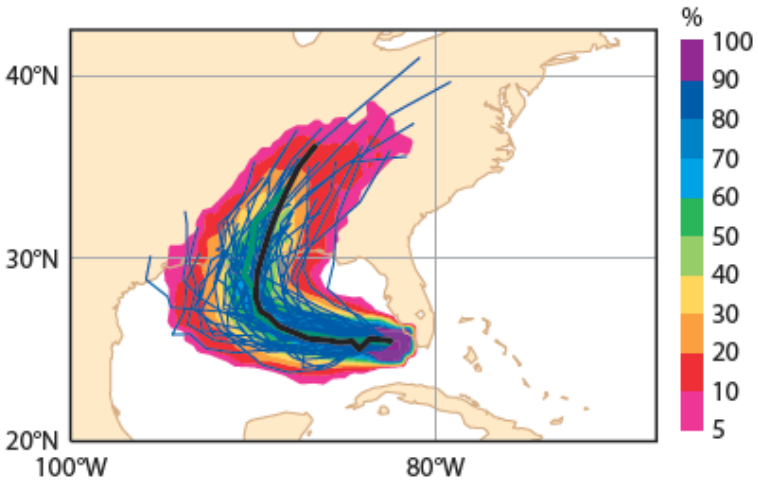

Figura 3.2: Evolución del Katrina atendiendo a varias predicciones de meteorología [3].

Algunas de los principales agencias de meteorología que ofrecen este tipo de servicio son:

- European Centre for Medium-Range Weather Forecasts (ECMWF): Consiste en una agencia intergubernamental ubicada en Reino Unido y formada por varios países europeos, incluido España. Lleva bastantes años trabajando tanto con predicciones numéricas como con este tipo de sistemas basados en conjuntos (en su caso conocido como Ensemble Prediction System - EPS) [3].
- National Centers for Environmental Prediction (NCEP): Se trata de una subdivisión del National Weather Service (NWS), el servicio estatal de meteorología estadounidense. Pese a ello ofrecen servicios globales como la predicción meteorológica numérica (Global Forecast System - GFS) y el sistema basado en conjuntos de predicciones (Global Ensemble Forecast System - GEFS).
- Otras: Además de las dos anteriores, existen muchas otras agencias que llevan a cabo este tipo de análisis en la actualidad. Entre ellas se encuentran la United Kingdom Met Office, Meteo France, Environment Canada, Japanese Meteorological Agency, Bureau of Meteorology (Australia) y China Meteorological Administration, así como universidades

dedicadas a la investigación y algunas otras agencias no relacionadas directamente con la meteorología (como la US Navy).

Para este TFG hemos utilizado únicamente los datos provenientes del ECMWF, del cuál hemos tomado, además de la predicción de meteorología determinista (con la que comparar), el Ensemble Prediction System (cuya utilidad queremos comprobar).

3.2. Ensemble Prediction System

Así pues, el Ensemble Prediction System es el sistema de predicción meteorológica del ECMWF basado en la generación de condiciones o estados iniciales de la atmósfera alternativos. Está formado por un total de 51 predicciones meteorológicas de las cuales una, conocida como predicción de control, se obtiene a partir de las condiciones iniciales medidas, y el resto, de modificar o perturbar dichas condiciones.

Es importante notar que existen diferencias entre las distintas predicciones arriba mencionadas, de forma que la fiabilidad de éstas varía. Por ejemplo, la predicción que utilizábamos hasta ahora, y que de aquí en adelante llamaremos determinista o HRES (High Resolution), se obtiene con resolución 0.1° (lo que equivale a unos 9 kilómetros entre puntos) mientras que las del EPS se obtienen con resolución 0.2° (18 kilómetros). Además, para la generación de las predicciones se utilizan modelos matemáticos distintos. Incluso dentro del propio EPS existen diferencias entre la predicción de control y el resto, ya que si bien se obtienen con la misma resolución, se rigen por parámetros y modelos diferentes. Todo esto influye en la calidad final de las predicciones, siendo en general más ajustada la determinista, luego la de control y finalmente el resto de predicciones del ensemble. Para más información al respecto ver [15].

Meteorología	Modelo	Resolución	Alcance	Horizontes
Determinista	Mejor	0.1°	10 días (240h)	3h hasta 144, 6h hasta 240
Control	Mejor	0.2°	15 días (360h)	3h hasta 144, 6h hasta 360
Resto Ensemble	Modificado	0.2°	15 días (360h)	3h hasta 144, 6h hasta 360

Tabla 3.1: Comparación entre las distintas predicciones de meteorología.

Para la obtención de los datos de meteorología se han seguido los pasos especificados en el apéndice A.

3.2.1. Análisis de la meteorología

Una vez extraídos y convertidos dichos datos hemos comparado las distintas predicciones del ensemble con respecto a la determinista. Para ello hemos estudiado cada una de las variables de meteorología por separado, a fin de hacer más visibles las diferencias. Pese a que en cada fichero tenemos la predicción para al menos los dos días siguientes, hemos seleccionado únicamente la meteorología para el octavo horizonte, es decir, 24 horas después. Esto tiene sentido ya que al predecir para un cierto día lo hacemos con los datos más recientes, que suelen ser los obtenidos el día anterior.

En primer lugar hemos estudiado las distintas predicciones alrededor del parque eólico Sotavento, para lo cual hemos comenzado seleccionando el punto con datos de meteorología más cercano. En las siguientes gráficas podemos observar la evolución de las predicciones para dicho

punto a lo largo de un mes[1]:

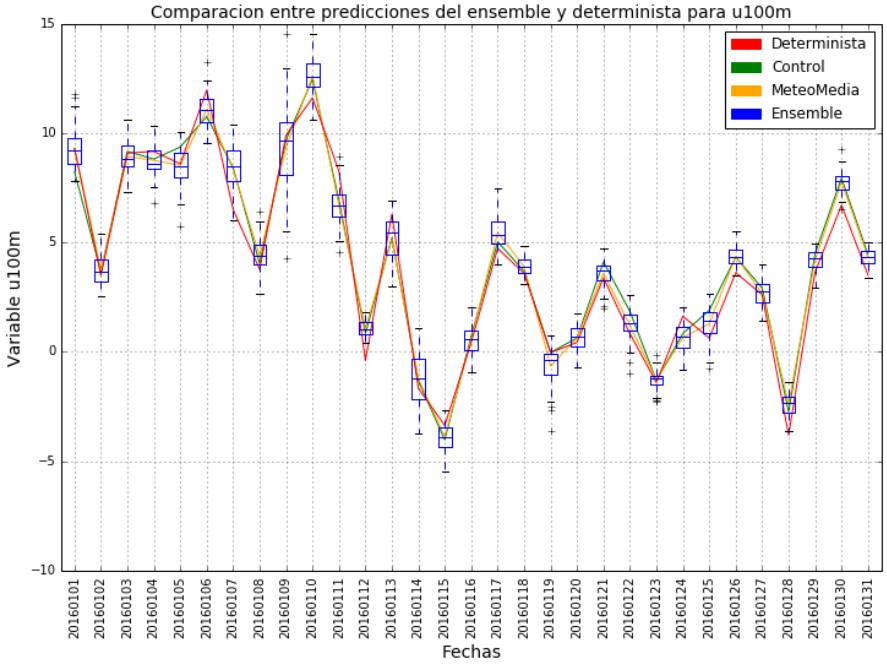

Figura 3.3: Evolución de la coordenada u a 100 metros en Sotavento para Enero de 2016.

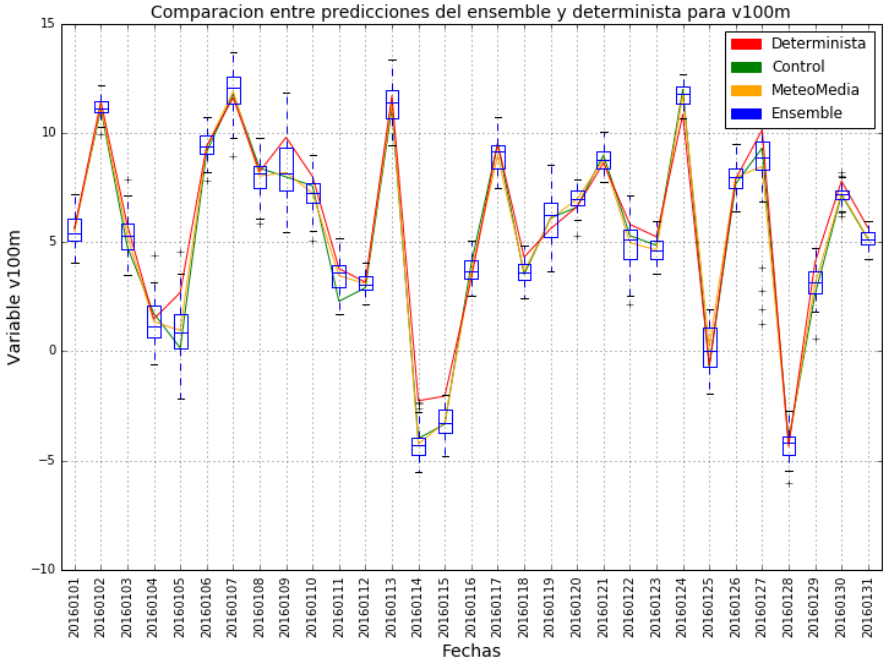

Figura 3.4: Evolución de la coordenada v a 100 metros en Sotavento para Enero de 2016.

[1] Aunque aquí solo hayamos incluido un mes, Enero de 2016, este mismo análisis lo hemos realizado para otros períodos y épocas del año. A pesar de que la meteorología es notablemente diferente según la estación en la que nos encontremos, los resultados y las conclusiones siguen siendo las mismas.

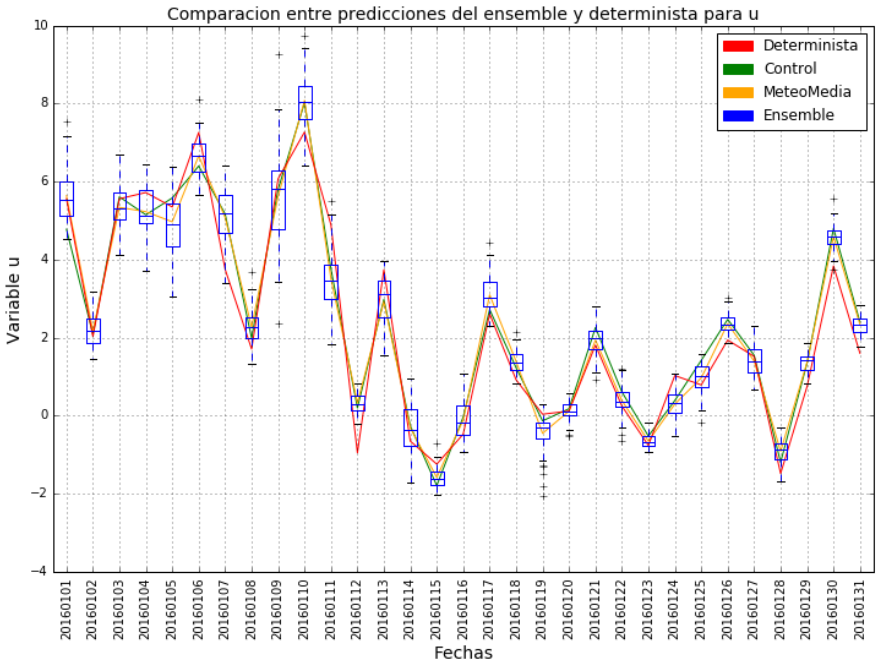

Figura 3.5: Evolución de la coordenada u a 10 metros en Sotavento para Enero de 2016.

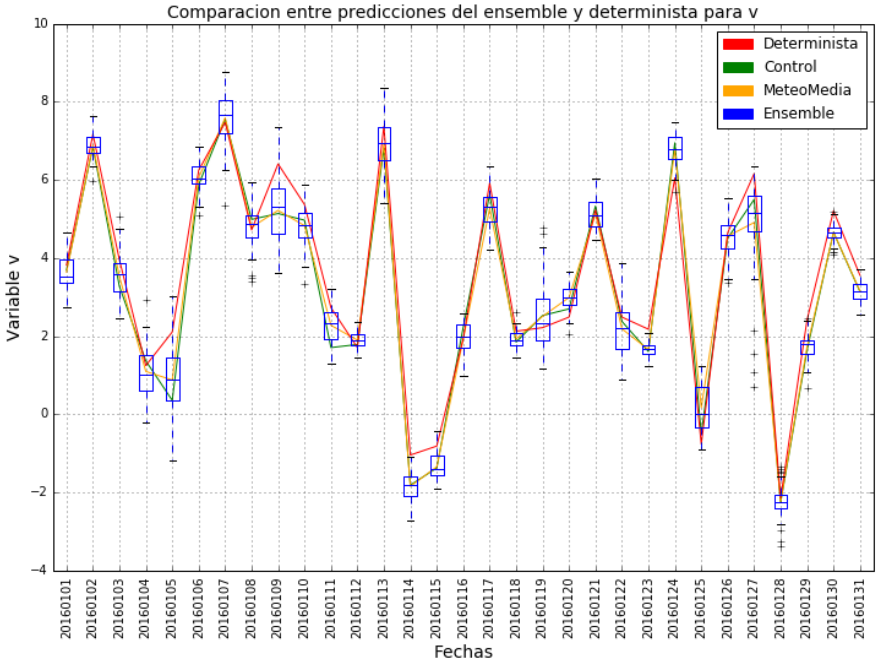

Figura 3.6: Evolución de la coordenada v a 10 metros en Sotavento para Enero de 2016.

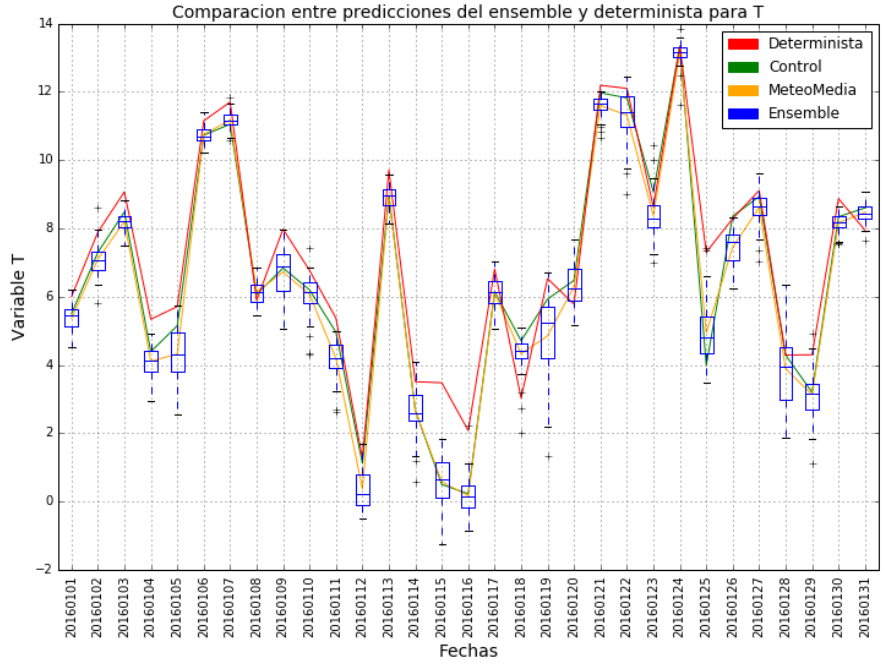

Figura 3.7: Evolución de la temperatura en Sotavento para Enero de 2016.

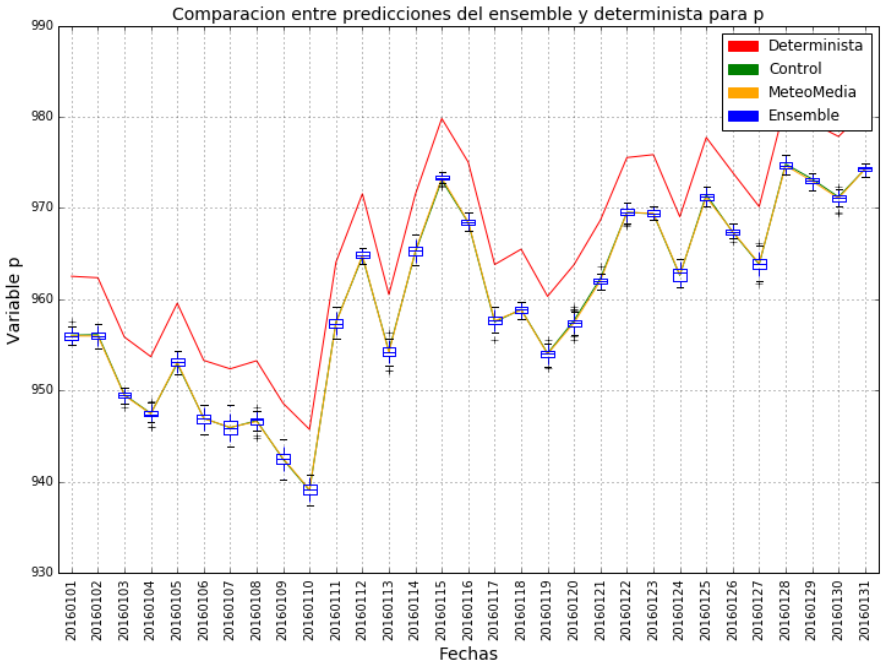

Figura 3.8: Evolución de la presión en Sotavento para Enero de 2016.

Se advierte como en general todas las predicciones son bastante similares, excepto en el caso de la presión. Esta variable es especialmente interesante ya que, por un lado, la meteorología determinista toma valores bastante más altos que los del ensemble, y por otro, la dispersión dentro del propio ensemble es mucho menor que para el resto de variables. Además, en cualquiera de las gráficas anteriores podemos ver que la predicción de control se asemeja notablemente a la media del ensemble, especialmente en épocas de escasa variación en las predicciones.

El hecho de que la meteorología determinista tome valores de presión más altos que los del ensemble podría ser un caso aislado, ya que tan solo hemos visto lo que ocurre en un punto. Es por ello que hemos decidido tomar la media para todo ese mes y comparar en varios puntos cercanos a Sotavento en lugar de en uno solo.

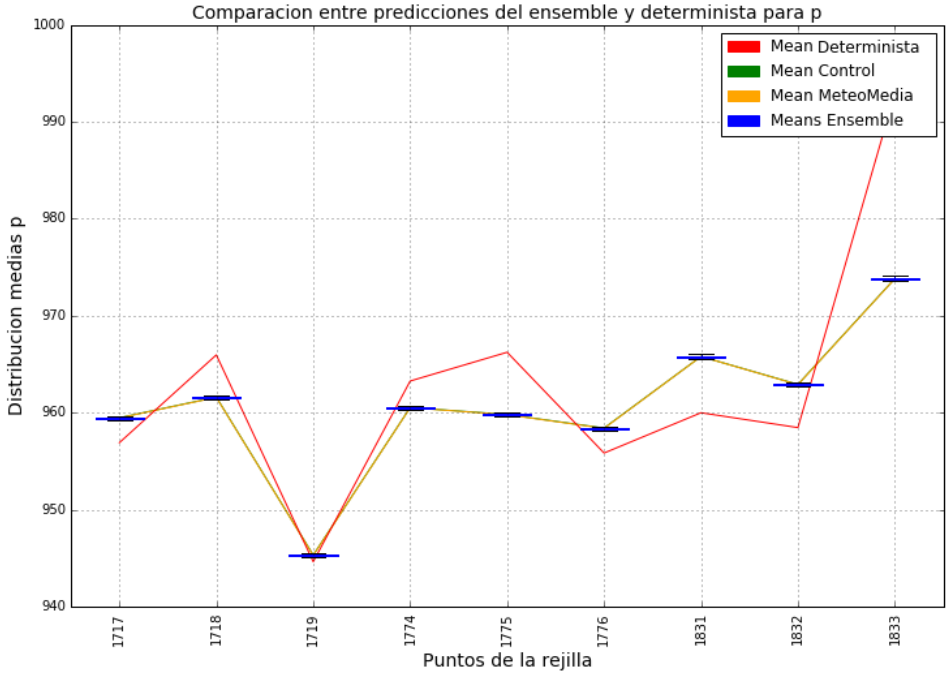

Figura 3.9: Medias de la presión en Sotavento para Enero de 2016.

Vemos que efectivamente la presión se escapa de los valores esperados, si bien no necesariamente por encima. Esto nos hace pensar que el cálculo de la presión no se lleva a cabo de la misma manera para la predicción determinista que para el ensemble.

Todos los resultados anteriores se referían exclusivamente al área de Sotavento, mientras que nosotros queremos comprobar la fiabilidad de las predicciones para toda la Península. En tal caso tendremos un total de 1965 puntos de meteorología, por lo que no es viable tratar cada uno de ellos por separado. Debido a esto únicamente hemos tomado la media de todos esos puntos y examinado su evolución a lo largo del tiempo (de nuevo un mes). Las gráficas relativas se encuentran al final del capítulo y van desde la 3.10 a la 3.15.

Atendiendo a estas gráficas asumiríamos pues que todas las predicciones se comportan igual, sin embargo ya hemos visto que esto no es así. La razón se encuentra en que al tomar la media de los puntos, aquéllos en los que la predicción toma valores superiores se compensan con aquéllos en los que los toma inferiores, por lo que las relaciones entre las distintas fuentes de meteorología no son representativas. A pesar de ello, sí podemos extraer algunas conclusiones globales, como el hecho de que no exista un sesgo positivo o negativo en las estimaciones de las variables, o que la dispersión para la presión sea efectivamente menor que la esperada.

Por otro lado, si nos fijamos exclusivamente en el ensemble, vemos que su dispersión tampoco es demasiado grande para el resto de variables. Esto se debe a que tan solo hemos estudiado la meteorología para 24 horas después al momento de ejecución de los modelos numéricos, por lo que apenas hay tiempo para que las predicciones se separen. Si nos alejáramos en el tiempo esta varianza aumentaría (sobre todo en el caso de las variables de viento), sin embargo para este trabajo no nos interesan las predicciones a más largo plazo.

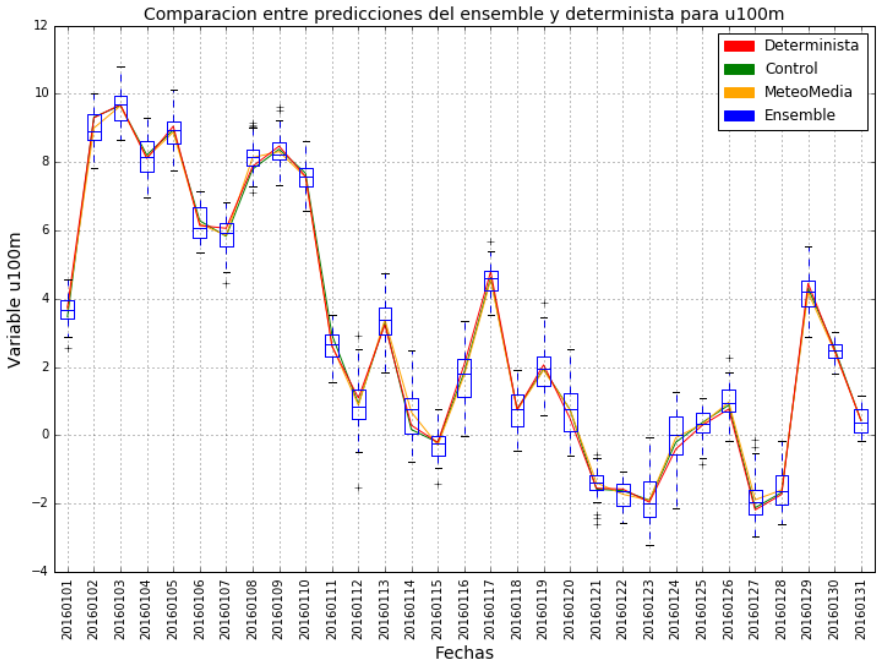

Figura 3.10: Evolución de la coordenada u a 100 metros en la Península para Enero de 2016.

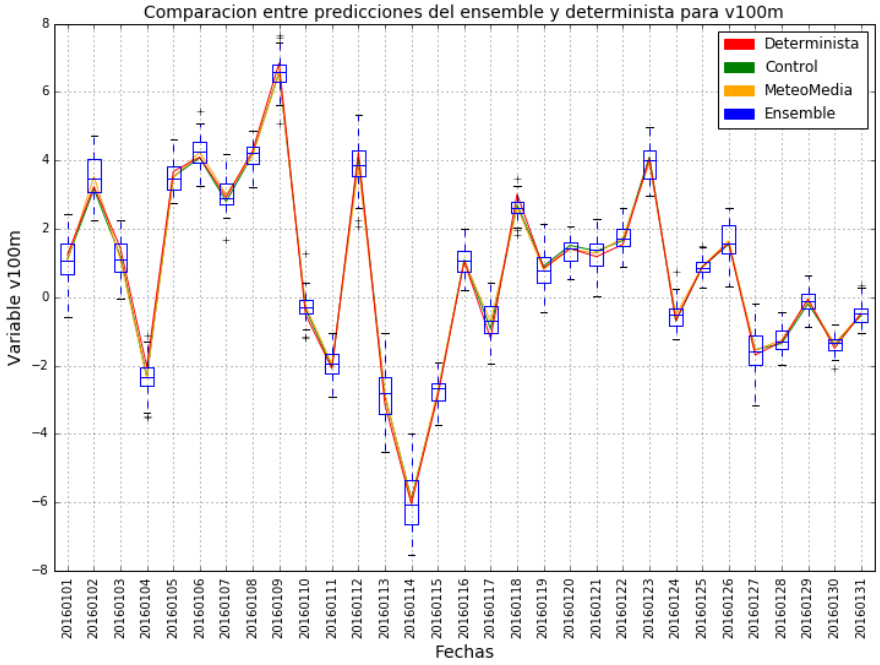

Figura 3.11: Evolución de la coordenada v a 100 metros en la Península para Enero de 2016.

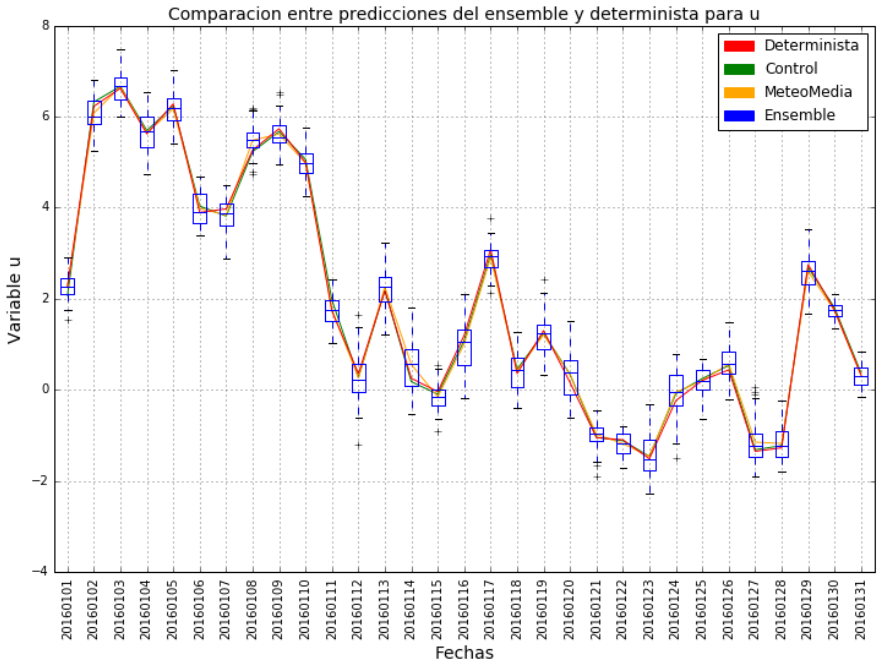

Figura 3.12: Evolución de la coordenada u a 10 metros en la Península para Enero de 2016.

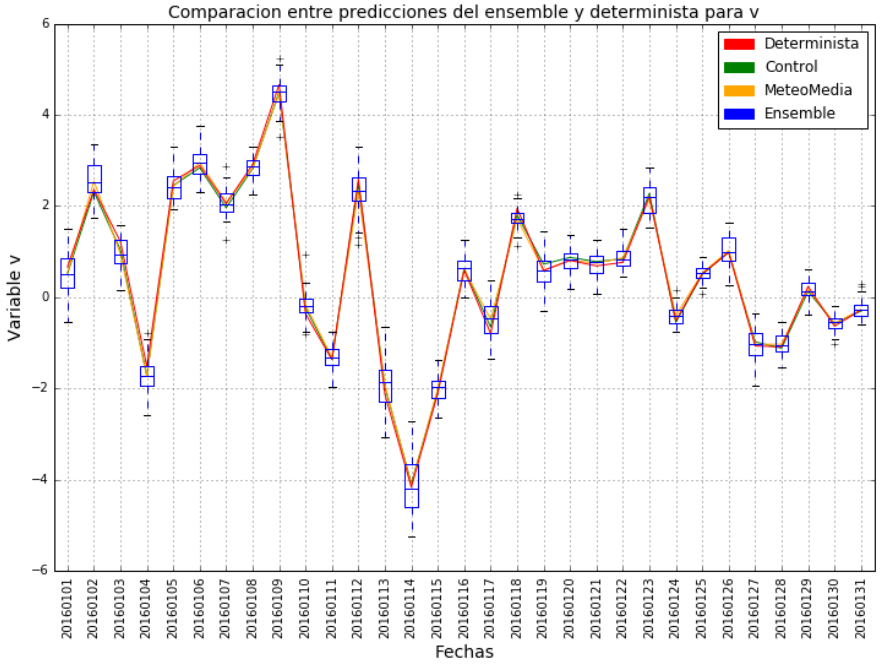

Figura 3.13: Evolución de la coordenada v a 10 metros en la Península para Enero de 2016.

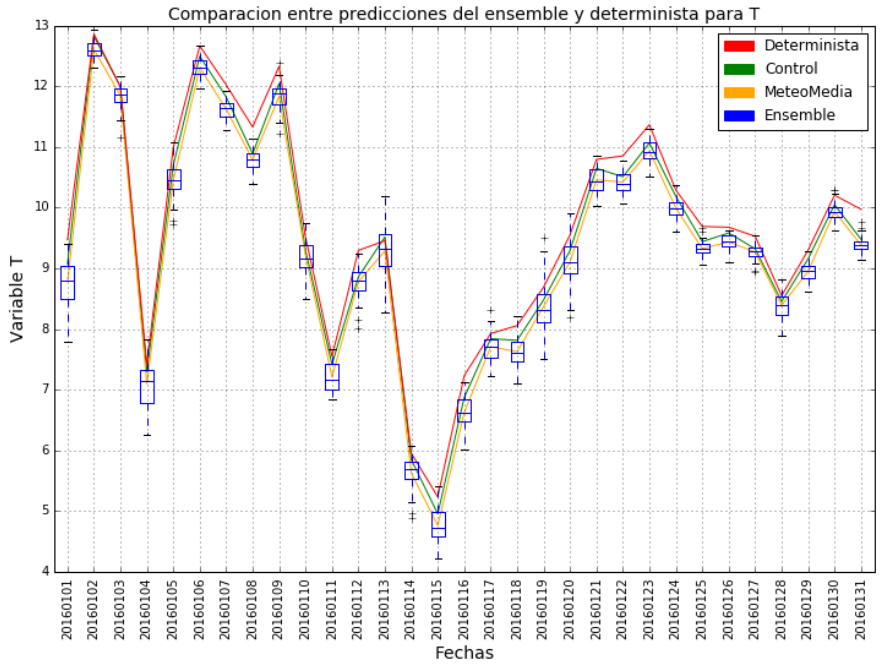

Figura 3.14: Evolución de la temperatura en la Península para Enero de 2016.

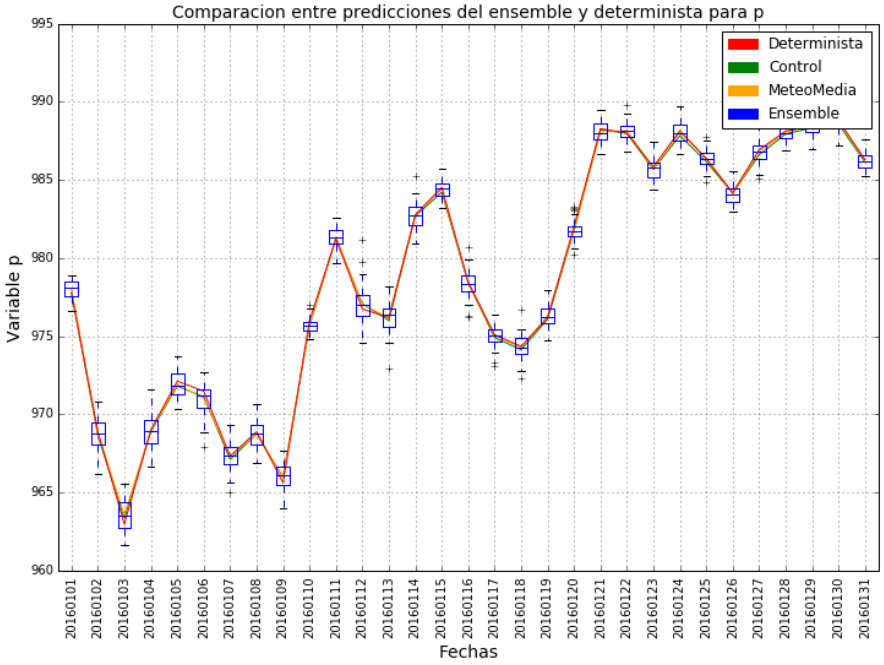

Figura 3.15: Evolución de la presión en la Península para Enero de 2016.

4

Algoritmos de Aprendizaje Automático

4.1. Introducción

El objetivo de este trabajo ha sido mejorar las predicciones de producción de energía eólica con respecto a las que ya se estaban realizando. Es por ello que para poder comparar se han seguido los mismos pasos que para la predicción HRES o determinista. Así pues, se han tomado el mismo período, modelos de predicción y variables de entrada como viento o temperatura que ya se utilizaban previamente. De esta manera, se han elegido como modelos de regresión aquellos que ya se habían probado y daban buenos resultados, en este caso Support Vector Machine (SVM) y Multilayer Perceptron (MLP). Como posible trabajo futuro se deberían probar otros algoritmos como Random Forest o Gradient Boosting.

4.2. Support Vector Machine

Las SVM tal como las conocemos en la actualidad fueron desarrolladas en 1992 por Vapnik y sus colaboradores en AT&T. No obstante, no fue hasta 1995 cuándo se definió su uso para problemas de regresión, obteniéndose un buen rendimiento.

En el caso de la regresión tenemos una entrada x $\in$ X $\subset \mathbb{R}^d$ (la meteorología) y una salida y $\in \mathbb{R}$ (la producción) y nuestro objetivo es encontrar la función f(x) que nos de la mejor salida posible para dicha x. Para ello al entrenar buscaremos dicha función de tal manera que todas nuestras muestras y_i se encuentren a una distancia como mucho ε de la función y al mismo tiempo ésta sea lo más "plana" u horizontal posible [16]. De aquí en adelante estudiaremos el caso particular de la regresión lineal, ya que cualquier otro tipo de función de regresión puede convertirse en lineal mediante una transformación a espacios superiores (via kernel functions). Así pues tendremos una función de la forma:

$$f(x) = \langle w, x \rangle + b \text{ con } w, x \in X, b \in \mathbb{R} \tag{4.1}$$

Puesto que queremos que f sea lo más "plana" posible buscaremos que w sea lo más cercano a 0. Pero además queremos que nuestra función cumpla que todos los puntos del conjunto de entrenamiento se encuentran a distancia como mucho ε, de modo que tendremos un problema

de optimización definido como sigue:

$$\begin{aligned} &\text{minimizar} && \frac{1}{2}\|w\|^2 \\ &\text{tal que} && |y_i - \langle w, x_i\rangle - b| \leq \varepsilon \end{aligned} \tag{4.2}$$

o lo que es lo mismo:

$$\begin{aligned} &\text{minimizar} && \frac{1}{2}\|w\|^2 \\ &\text{tal que} && \begin{cases} y_i - \langle w, x_i\rangle - b \leq \varepsilon \\ -y_i + \langle w, x_i\rangle + b \leq \varepsilon \end{cases} \end{aligned} \tag{4.3}$$

Puesto que lo anterior no siempre es factible en ocasiones es necesario permitir que algunos de los puntos puedan encontrarse a distancia mayor que ε. En tal caso, al elegir nuestra función se deben tener en cuenta, tanto los errores cometidos, como que ésta sea lo más "plana" posible. Para ello se define una constante $C > 0$, la cual determina los errores que se pueden cometer manteniendo que nuestra función sea lo más horizontal posible, y las variables ξ_i y ξ_i^*, que representan la distancia de nuestra muestra i hasta la función f. Nuestro problema queda entonces definido como:

$$\begin{aligned} &\text{minimizar} && \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(c(\xi_i) + c(\xi_i^*)) \\ &\text{tal que} && \begin{cases} y_i - \langle w, x_i\rangle - b \leq \xi_i \\ -y_i + \langle w, x_i\rangle + b \leq \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \end{aligned} \tag{4.4}$$

Normalmente se utiliza como función de pérdida c(ξ) una función conocida como ε-intensitive y que se define como:

$$|\xi|_\varepsilon = \begin{cases} 0 & \text{si } |\xi| \leq \varepsilon \\ |\xi| - \varepsilon & \text{en cualquier otro caso} \end{cases}$$

Atendiendo a esta función solo se tienen en cuenta los errores superiores a nuestro margen ε y se ignoran el resto, si bien se pueden utilizar otras opciones como la función Gaussiana o la de Huber.

Para entender más fácilmente lo anterior se pueden ver los siguientes ejemplos para una sola variable de entrada (d=1). En el primer caso se puede resolver el problema, mientras que en el segundo es necesario aceptar ciertos errores.

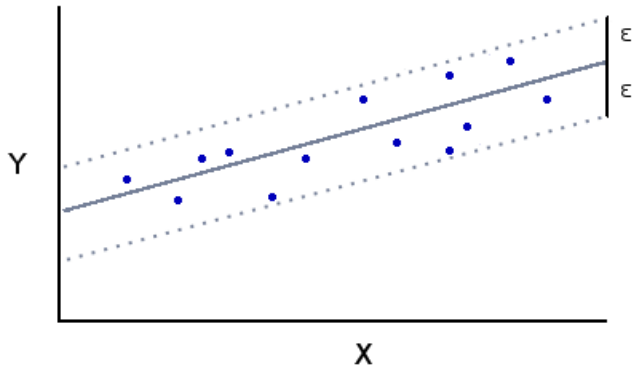

Figura 4.1: Ejemplo de SVR con solución.

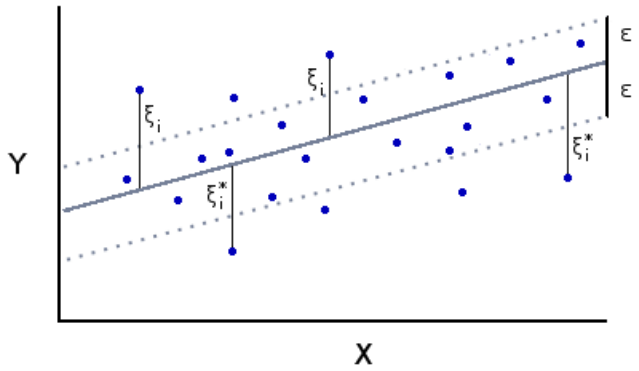

Figura 4.2: Ejemplo de SVR en que hay que asumir ciertos errores.

Una vez se tiene la formulación anterior podemos convertir el problema al dual, sobre el cual se puede trabajar más fácilmente para pasar a espacios superiores y, por lo tanto, utilizar funciones no lineales. En el caso lineal anterior obtendríamos como resultado:

$$f(x) = \sum_{i=1}^{N}((\alpha_i - \alpha_i^*)\langle x_i, x\rangle) + b \text{ , donde } \alpha_i \text{ y } \alpha_i^* \text{ son las variables duales.} \tag{4.5}$$

Mientras que en el caso no lineal tendríamos que pasar a espacios superiores quedando:

$$f(x) = \sum_{i=1}^{N}((\alpha_i - \alpha_i^*)\langle \phi(x_i), \phi(x)\rangle) + b \tag{4.6}$$

O lo que es lo mismo:

$$f(x) = \sum_{i=1}^{N}((\alpha_i - \alpha_i^*)K(x_i, x)) + b \tag{4.7}$$

donde K es la función específica del Kernel.

Uno de los métodos de kernel más utilizados es RBF, cuya función se define como:

$$K(u, v) = exp(-\frac{\|u - v\|^2}{2\sigma^2}) = exp(-\gamma\|u - v\|^2) \tag{4.8}$$

donde $\gamma = \frac{1}{2\sigma^2}$ es una constante que controla la forma de los "picos" al pasar a dimensiones superiores, de tal manera que un valor de γ bajo implica gaussianas suaves y uno alto gaussianas más afiladas. Se debe tener especial cuidado al determinar esta constante, ya que en caso de ser demasiado alta puede producir un sobreajuste del modelo.

4.3. Multilayer Perceptron

El origen del perceptrón es mucho anterior al de las SVM. De hecho ya en 1957 se había ideado una primera versión de éste, si bien se trataba de un perceptrón básico, de una sola capa. Puesto que este tipo de perceptrón tan solo permite separar y clasificar problemas linealmente separables su desarrollo fue dejado de lado hasta mucho tiempo después, en que se comenzó a ver el potencial de utilizar más de una capa.

El perceptrón es un tipo de red neuronal, compuesta por tanto por neuronas artificiales. Estas neuronas se comportan de forma similar a las neuronas biológicas, recibiendo una entrada y produciendo una salida en base a una función de activación que tengan definida [17, 18].

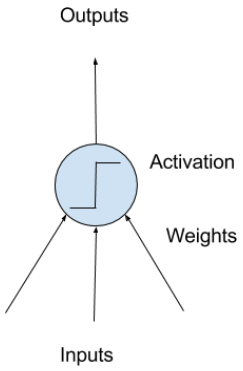

Figura 4.3: Representación de una neurona artificial [4].

Las dos funciones de activación más conocidas son las sigmoides:

- Tangente hiperbólica:

$$\tanh(z) = \frac{\sinh z}{\cosh z} = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{4.9}$$

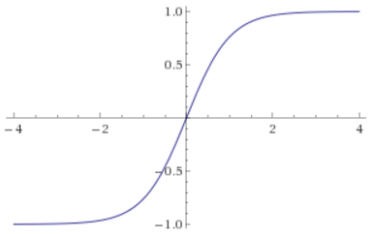

Figura 4.4: Tangente hiperbólica.

- Función logística

$$f_{log}(z) = \frac{1}{1 + e^{-z}} \tag{4.10}$$

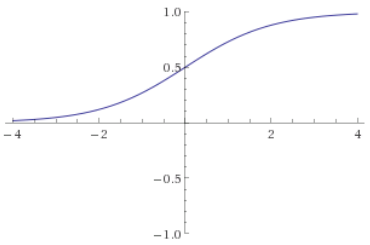

Figura 4.5: Función logística.

Estas funciones de activación reciben un único valor de entrada, mientras que las neuronas pueden recibir más de una. La razón se encuentra en que el valor de nuestra función de activación se corresponde en realidad con la suma de cada una de las variables de entrada, más un término

independiente que llamaremos bias. Para llevar a cabo esta suma las neuronas necesitan además dar un peso específico a cada una de las entradas, así como al término independiente. Estos pesos afectan de manera importante a nuestra salida, de forma que deben buscarse los mejores pesos al entrenar. La entrada de nuestra función de activación quedaría por tanto como:

$$z = \langle w, x \rangle + w_0 = \sum_{i=1}^{K} w_i * x_i + w_0 \ , \tag{4.11}$$

donde x_i representa cada entrada de la neurona y w_i su peso.

Ahora bien, las neuronas por separado tan solo pueden representar a una función lineal, por lo que no se utilizan de forma aislada sino formando redes. De esta manera la entrada de una neurona se utiliza como salida de otra y así sucesivamente. El perceptrón es un caso particular de una red neuronal, en el que las neuronas se organizan en capas, de tal manera que la salida de las neuronas de la capa i sirven como entrada únicamente para las de la capa $i+1$. Se distinguen aquí pues la capa de entrada, que recibe directamente nuestra entrada x; las capas ocultas, que son aquellas cuyas neuronas toman como entrada y salida las de otroas neuronas; y la de salida, que es la que produce nuestra salida y. Desde el punto de vista de un grafo el perceptrón se representaría como un grafo acíclico dirigido, como el que se puede ver en la imagen:

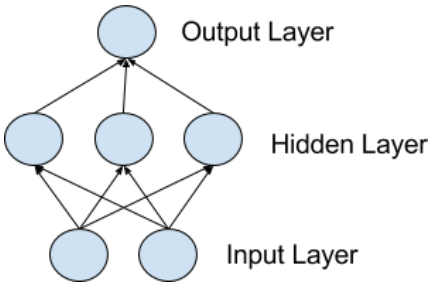

Figura 4.6: Representación de un perceptrón multicapa [4].

Como hemos visto, el funcionamiento del perceptrón es completamente diferente al de las SVM, permitiendo directamente su uso para problemas de regresión (la salida es un valor continuo). No obstante, al igual que en el caso de las SVM, existen ciertos parámetros que se deben especificar previamente a la ejecución del modelo. En este caso son aquellos que describen la forma de la red, como por ejemplo el número de capas ocultas o el número de neuronas en cada una de ellas.

Una vez determinados estos parámetros se deben eligir los pesos de cada neurona. Para ello se suele utilizar el algoritmo de propagación hacia atrás (backpropagation) junto con el método de descenso por gradiente. El descenso por gradiente nos sirve para encontrar un mínimo local para cualquier función diferenciable, en este caso la del error cometido. En general tomaremos como función de coste el error cuadrático medio, que para un conjunto de entrenamiento con I muestras se define como:

$$E = \sum_{i=1}^{I} (e_i - o_i)^2 \ \text{, donde } e_i \text{ representa la salida esperada y } o_i \text{ la obtenida de la red.} \tag{4.12}$$

Para encontrar dicho mínimo el método del descenso por gradiente se basa en el hecho de que dada una función $F(a)$ la dirección de mayor decrecimiento de ésta viene dada por su gradiente ∇F negado. De esta forma si nos encontramos en $a_0 \in A$ entonces para $a_1 = a_0 - \eta \nabla F(a_0)$, con η una constante que cumpla ciertos requisitos, se tiene que $F(a_1) \leq F(a_0)$. Por lo tanto si tomamos la sucesión $a_{k+1} = a_k - \eta \nabla F a_k$ es de esperar que acabemos llegando a un mínimo

local de la función. La constante anterior η, conocida como tasa de aprendizaje, nos sirve para determinar la longitud del paso que podemos dar en la dirección de decrecimiento. Puesto que esta dirección se calcula de forma local, la tasa de aprendizaje η (y por tanto la longitud del paso) no puede ser demasiado grande, ya que la condición de decrecimiento anterior podría cambiar al alejarnos. Por otro lado, η tampoco puede ser excesivamente pequeña, ya que en ese caso el método requeriría de demasiadas iteraciones para alcanzar el mínimo, con lo que la búsqueda se ralentizaría notablemente. Es por ello que la tasa de aprendizaje no suele ser fija sino que se elige en cada iteración: $a_{k+1} = a_k - \eta_k \nabla F a_k$, siendo menor cuanto más cerca nos encontremos del mínimo. En el caso que nos interesa de minimizar el error cuadrático obtendremos para cada peso una dirección, de manera que:

$$w_{ij} = w_{ij} + \Delta w_{ij} = w_{ij} - \eta \frac{\partial E}{\partial w_{ij}} \ , \tag{4.13}$$

donde w_{ij} representa el peso dado por la neurona j a la entrada que recibe de la neurona i de la capa anterior.

Quedaría por tanto calcular $\frac{\partial E}{\partial w_{ij}}$, para lo cuál utilizamos la regla de la cadena:

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial output_j} \frac{\partial output_j}{\partial input_j} \frac{\partial input_j}{\partial w_{ij}} \ , \tag{4.14}$$

donde $output_j$ representa la salida de la neurona j e $input_j$ su entrada.

Recordando lo anterior:

$$output_j = f_{activacion}(input_j) \ \text{, y}$$
$$input_j = \sum_{i=1}^{K} w_{ij} * output_i$$

Por lo tanto para realizar el cálculo de $\frac{\partial E}{\partial w_{ij}}$ tendremos en primer lugar $\frac{\partial input_j}{\partial w_{ij}} = output_i$.

Por otro lado tendremos $\frac{\partial output_j}{\partial input_j}$, que dependerá de la $f_{activacion}$ escogida.

Y finalmente $\frac{\partial E}{\partial output_j}$. Para el cálculo de este último término podemos tomar E como una función que en lugar de depender de $output_j$ dependa de las entradas de las neuronas que reciban esa salida. Es decir:

$$\frac{\partial E(output_j)}{\partial output_j} = \frac{\partial E(input_a, input_b, input_c ...)}{\partial output_j} \ , \tag{4.15}$$

con a,b,c... neuronas de la siguiente capa.

Obtendremos así una recursión, de forma que para calcular este término necesitamos conocer los de la capa siguiente más cercana a la de salida (excepto en el caso de la de salida, que podemos calcularlo directamente). Es por esto que el proceso para calcular los pesos se conoce también como propagación hacia atrás o backpropagation.

Puesto que los pesos van creciendo en cada iteración es posible que tras varias actualizaciones éstos sean demasiado altos, lo cual suele indicar un sobreajuste del modelo. Para evitar que esto ocurra se utilizan en general dos tipos de regularizaciones [19], que consisten en modificar la función de coste E de manera que se limite la magnitud de dichos pesos. Los dos tipos de regularización son:

- L1:

$$\tilde{E} = \sum_{i=1}^{I} (e_i - o_i)^2 + \sum_{i,j} |w_{ij}| \tag{4.16}$$

- L2:

$$\tilde{E} = \sum_{i=1}^{I}(e_i - o_i)^2 + \sum_{i,j} w_{ij}^2 \tag{4.17}$$

En nuestro caso necesitamos que la función de coste sea diferenciable por lo que únicamente nos vale la regularización L2, en cuyo caso obtenemos:

$$w_{ij} = w_{ij} + \Delta w_{ij} = w_{ij} - \eta \frac{\partial \tilde{E}}{\partial w_{ij}} = w_{ij} - \eta \frac{\partial E}{\partial w_{ij}} + 2w_{ij} \tag{4.18}$$

De cara a controlar dicha regularización se puede introducir además una constante λ, conocida como weight decay o factor de decaimiento, con lo que la función de coste nos quedaría:

$$\tilde{E} = \sum_{i=1}^{I}(e_i - o_i)^2 + \lambda \sum_{i,j} w_{ij}^2 \tag{4.19}$$

Y por tanto:

$$w_{ij} = w_{ij} - \eta \frac{\partial \tilde{E}}{\partial w_{ij}} = w_{ij} - \eta \left(\frac{\partial E}{\partial w_{ij}} + 2\lambda w_{ij} \right) \tag{4.20}$$

De esta manera cuanto mayor sea el factor de decaimiento λ, mayor será también la reducción y menores los pesos resultantes. Debemos controlar por tanto que el valor de este factor no sea demasiado alto, en cuyo caso los pesos se irían a 0, ni demasiado bajo, en cuyo caso no afectaría prácticamente nada.

5

Experimentos Realizados y Resultados

5.1. Experimentos realizados

Hasta ahora habíamos visto la utilidad del Ensemble Prediction System para evaluar y mejorar las predicciones de meteorología. En teoría esto nos debería permitir mejorar las predicciones para la generación de energía eólica, sin embargo como veremos en la práctica no tiene por qué ser así. Ejemplos de estudios anteriores al respecto se pueden encontrar en [20, 21].

Para nuestro trabajo hemos comenzado estudiando la posibilidad de utilizar únicamente la predicción de control, para a continuación explorar otras opciones como calcular la media de las predicciones obtenidas con cada uno de los miembros del ensemble o bien predecir directamente con la media de la meteorología. Puesto que cada parque eólico es diferente hemos optado por llevar a cabo las pruebas en varios de ellos. En este caso hemos seleccionado para las pruebas Sotavento y el conjunto de parques de toda la Península, por cuyos datos me gustaría dar las gracias a Red Eléctrica de España, que nos ha permitido disponer de ellos.

En general para las pruebas hemos utilizado los datos de aproximadamente los tres últimos años, desde el 01/01/2014 hasta el 31/10/2016. El primer año entero y parte del segundo (hasta el 31/10/2015) nos han servido para buscar los mejores parámetros (hiperparametrizar), mientras que hemos dejado el último para evaluar la calidad del modelo. Puesto que los datos de meteorología son trihorarios y las producciones horarias, al entrenar y predecir hemos descartado las producciones de las horas intermedias. De esta manera las predicciones se obtienen también de forma trihoraria, por lo que necesitamos hacer una interpolación lineal para poder comparar las horas intermedias. Para la ejecución de estas pruebas nos hemos servido del software de EA2 del IIC, el cual ya estaba implementado, si bien podíamos haber utilizado otras opciones como la librería Scikit-Learn de Python.

Para cada tipo de algoritmo se han metaparametrizado distintas variables:

- SVM: Hemos metaparametrizado todas las constantes del modelo, C, ε, y γ, las cuales han sido explicadas en el capítulo anterior.
- MLP: En este caso se han metaparametrizado el número de capas ocultas de la red y el factor de decaimiento λ, mientras que hemos fijado el número de neuronas de las capas ocultas en 10. Además, hemos utilizado un filtrado Hampel para eliminar outliers, de

manera que debíamos definir también el porcentaje de valores eliminados al entrenar. Este porcentaje se ha tratado por tanto como un metaparámetro más del modelo.

Por otro lado, puesto que el proceso de hiperparametrización es muy costoso computacionalmente, se ha realizado exclusivamente para la predicción de control, basándonos en sus resultados para el resto de miembros del ensemble. Así pues, el primer paso para cada parque ha sido crear un nuevo modelo (o varios) para la predicción de control y encontrar sus mejores parámetros. Para la selección de estos parámetros se ha utilizado validación cruzada, tomando en cada iteración un mes para test y el resto para train. Es importante notar que no se ha creado un modelo único para cada parque sino varias aproximaciones de dicho modelo, teniendo en cuenta diferentes variables y puntos de meteorología. Las especificaciones de cada aproximación se detallarán más adelante, si bien aquí vamos a dar únicamente una idea general de lo que se ha hecho para todas ellas.

Una vez seleccionados los mejores parámetros ya podemos entrenar y realizar las predicciones. De cara a predecir hemos entrenado los modelos para cada día con los datos de los 13 meses anteriores y, una vez hecho esto, predicho para todas las horas de ese día.

A continuación estudiamos para cada aproximación la opción de predecir con cada uno de los miembros restantes del EPS. De esta manera obtuvimos 50 predicciones distintas, las cuales no tenían apenas relación entre ellas. El siguiente paso fue por tanto estudiar todo el cómputo de predicciones. Como primera aproximación optamos por hacer la media de las predicciones y compararla con la producción real. A continuación modificamos dicha media para que fuera una media ponderada en base a la bondad de cada predicción, en lugar de una media uniforme. El sentido de hacer esto reside en el hecho de que algunos de los miembros del ensemble parten de perturbaciones prácticamente inverosímiles sobre las condiciones iniciales, de manera que sus resultados también lo son, mientras que otros son mucho más factibles. Para el cálculo de los coeficientes se ha seguido la siguiente fórmula, tomando como error el NMAE:

$$Coef_i = \frac{Acierto_i - Acierto_{min}}{\sum(Acierto_i - Acierto_{min})} \text{ , donde } Acierto_i = 1 - (\frac{Error_i}{\sum Error_i}) \tag{5.1}$$

Esta fórmula es claramente mejorable y se podía haber sustituido por otras opciones como una regresión lineal, tomando cada miembro del ensemble como una variable. No obstante, como veremos más adelante con los intervalos de confianza, la mejora mediante esta vía está bastante limitada por lo que no hemos seguido explorando.

Finalmente, hemos calculado la meteorología media del ensemble (sin incluir a la de control), la cual utilizamos como dato de partida para predecir. Esta media es muy útil, ya que si bien se asemeja bastante a la predicción de control, en períodos de gran variabilidad debiera ser más fiable. Pese a ello tampoco hemos ajustado el modelo para este tipo de meteorología, lo cual quedaría pendiente como trabajo futuro.

5.2. Aproximaciones de los modelos

Antes de continuar es necesario describir un nuevo parámetro de los modelos, que llamaremos dilatación y que hace referencia al número de puntos de meteorología que se toman alrededor de cada parque. Así pues, supongamos que un parque es un punto en el mapa, tomaremos para entrenar y predecir todos los puntos de meteorología "lo suficientemente cercanos", es decir, aquellos que se encuentren a una distancia menor que $dist_{max}$, que en nuestro caso definimos como:

$$dist_{max} = resolucion * dilatacion \tag{5.2}$$

En primer lugar hemos seleccionado las versiones de los modelos a ejecutar, que en el caso de Sotavento han sido Multilayer Perceptron y en el de REE Support Vector Machines. La razón se encuentra en que estos algoritmos ya habían sido probados y daban buenos resultados en el momento de comienzo de las pruebas. Pese a ello hubiera sido aconsejable probar ambas opciones en los dos casos, lo cual no ha sido posible debido a limitaciones de tiempo.

Para Sotavento hemos ejecutado una sola versión de MLP, con todas las variables de meteorología[1], para cada una de las posibles fuentes de datos de entrada. Tendremos pues unas predicciones diferentes para la meteorología determinista, de control, el ensemble y la media del ensemble (independientemente de que el modelo para estas dos últimas sea el mismo que para la de control). Por otro lado, en el caso del ensemble obtendremos 50 predicciones, pese a lo cual lo trataremos como un todo del que sacaremos más adelante tres posibles predicciones: la media, la media ponderada y la mediana. Los detalles de la aproximación para cada tipo de meteorología:

Identificador	Algoritmo	Meteorología	Variables	Resolución	Dilatación
MLP01	MLP	Determinista	V100, V10, P y T	0.125°	1.5
MLP02	MLP	Control	V100, V10, P y T	0.25°	1.5
MLP03	MLP	Ensemble (50 preds)	V100, V10, P y T	0.25°	1.5
MLP04	MLP	Media Ens.	V100, V10, P y T	0.25°	1.5

Tabla 5.1: Aproximaciones de los modelos para Sotavento.

En el caso de REE en cambio tendremos dos aproximaciones de SVR para cada meteorología, una con todas las variables y otra sin las de viento a 10 metros. Además, en esta ocasión hemos reducido la dilatación de ambas aproximaciones a 1. Las tablas con la descripción de cada aproximación:

Identificador	Algoritmo	Meteorología	Variables	Resolución	Dilatación
SVR01-1	SVM	Determinista	V100, V10, P y T	0.125°	1
SVR02-1	SVM	Control	V100, V10, P y T	0.25°	1
SVR03-1	SVM	Ensemble	V100, V10, P y T	0.25°	1
SVR04-1	SVM	Media Ens.	V100, V10, P y T	0.25°	1

Tabla 5.2: Aproximaciones para REE con todas las variables.

Identificador	Algoritmo	Meteorología	Variables	Resolución	Dilatación
SVR01-2	SVM	Determinista	V100, P y T	0.125°	1
SVR02-2	SVM	Control	V100, P y T	0.25°	1
SVR03-2	SVM	Ensemble	V100, P y T	0.25°	1
SVR04-2	SVM	Media Ens.	V100, P y T	0.25°	1

Tabla 5.3: Aproximaciones para REE sin viento a 10 metros.

[1]Se toman como variables de viento tanto sus coordenadas u y v por separado como su módulo $\|V\| = \sqrt{u^2 + v^2}$, el cual se calcula específicamente a partir de estas. En las tablas denotaremos como V100 (o V10) al conjunto de todas las variables de viento a esa altura, es decir, tanto las coordenadas como el módulo.

5.3. Resultados obtenidos

De cara a evaluar la calidad de las distintas aproximaciones hemos empezado comparando el error obtenido para todo el período de prueba (1 año). En las tablas siguientes podemos ver resumido dicho error, tomando como indicador principal el NMAE.

Identificador	NMAE(%)	NRMSE(%)	NMBE(%)
MLP01	6.198691	9.408009	-1.102091
MLP02	6.439199	9.633736	-0.448360
MLP03 (media)	6.317558	9.216153	-0.244265
MLP03 (media pond.)	6.300994	9.191209	-0.251715
MLP03 (mediana)	6.256347	9.265063	-0.475976
MLP04	6.224751	9.346695	-1.113905

Tabla 5.4: Error para cada una de las aproximaciones de Sotavento.

Identificador	NMAE(%)	NRMSE(%)	NMBE(%)
SVR01-1	2.553876	3.459167	0.102147
SVR02-1	2.644048	3.526296	0.296974
SVR03-1 (media)	2.594312	3.437740	0.420291
SVR03-1 (media pond.)	2.578830	3.420621	0.413832
SVR03-1 (mediana)	2.587551	3.439882	0.392235
SVR04-1	2.694860	3.606860	-0.791094
SVR01-2	2.511161	3.405695	0.501944
SVR02-2	2.741438	3.642274	0.973928
SVR03-2 (media)	2.716142	3.590327	1.092114
SVR03-2 (media pond.)	2.699424	3.572599	1.073783
SVR03-2 (mediana)	2.709596	3.588820	1.064230
SVR04-2	2.635040	3.546535	-0.1261311

Tabla 5.5: Error para cada una de las aproximaciones de REE.

Viendo los resultados obtenidos para el NMAE podemos empezar a extraer ciertas conclusiones, como el hecho de que ninguna de las predicciones del ensemble mejora realmente a la basada en la meteorología determinista. Sin embargo, si atendemos al NRMSE, los pronósticos del ensemble sí mejoran a los de la determinista (o al menos la diferencia es menor). Esto nos indica que aunque las predicciones sacadas del ensemble sean en general peores, éstas son más robustas ante grandes desviaciones. Por otro lado el NMBE nos sirve para ver si las predicciones subestiman o sobrestiman la producción (parece que en Sotavento son ligeramente inferiores y en REE superiores, pese a lo cual dicho sesgo no es especialmente significativo). Además, en todas las aproximaciones parece que tanto la predicción de la meteorología determinista como la de la media del ensemble toman valores inferiores al resto (especialmente ésta última).

En cualquier caso estos errores por sí solos no son muy significativos, por lo que hemos decidido dibujar la evolución de cada una de las predicciones con respecto a la producción. De aquí en adelante vamos a incluir únicamente las gráficas relativas a Sotavento, si bien hemos comprobado que los resultados y las conclusiones obtenidas se pueden extrapolar también a los parques de REE. A continuación se pueden ver algunos días en concreto, en los cuales tendremos tanto producciones altas como bajas y situaciones con mucha y poca dispersión:

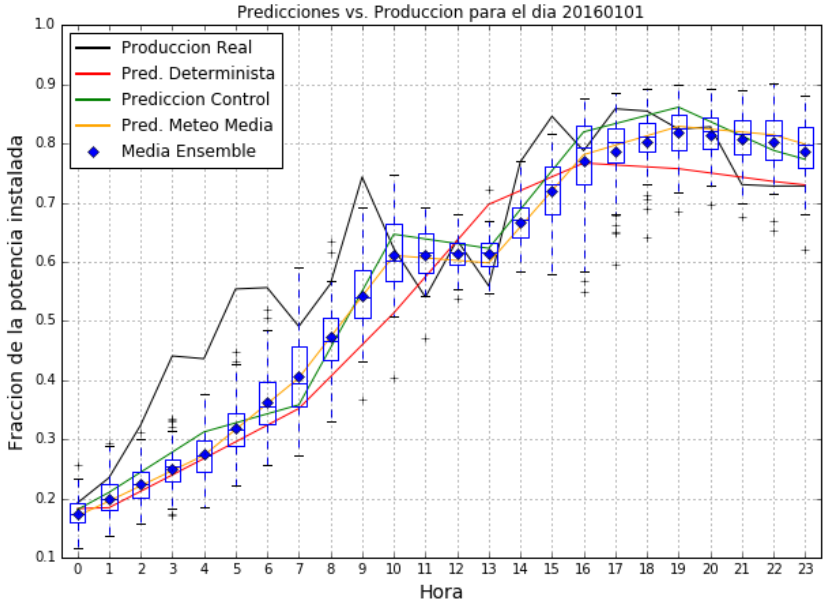

Figura 5.1: Evolución de las predicciones para Sotavento el 01/01/2016.

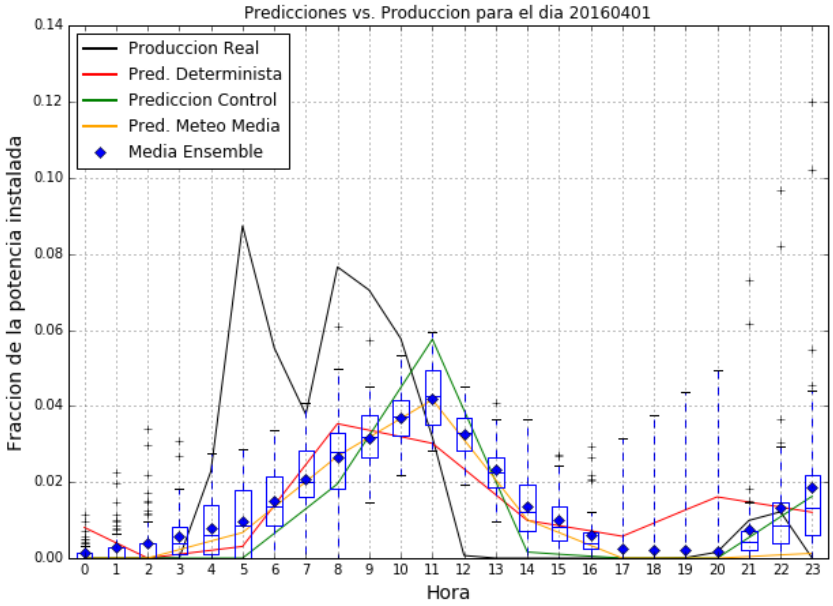

Figura 5.2: Evolución de las predicciones para Sotavento el 01/04/2016.

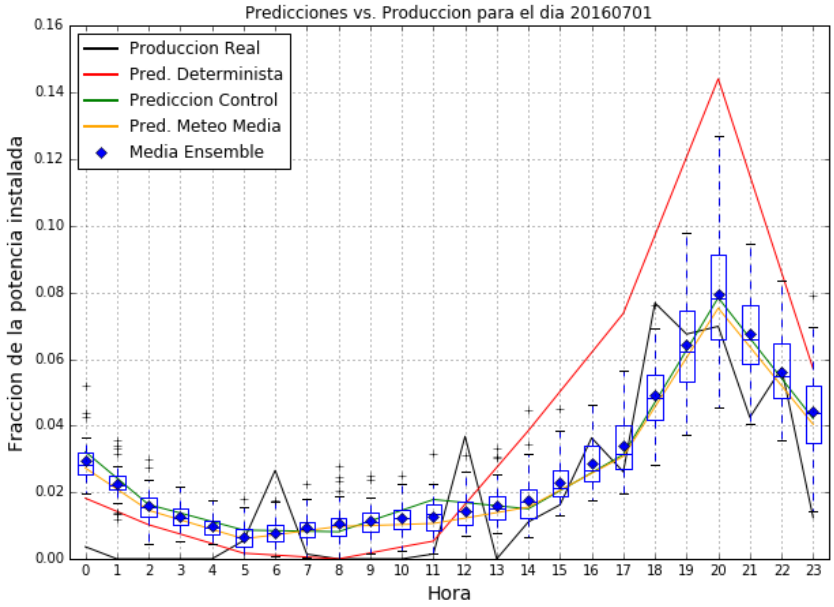

Figura 5.3: Evolución de las predicciones para Sotavento el 01/07/2016.

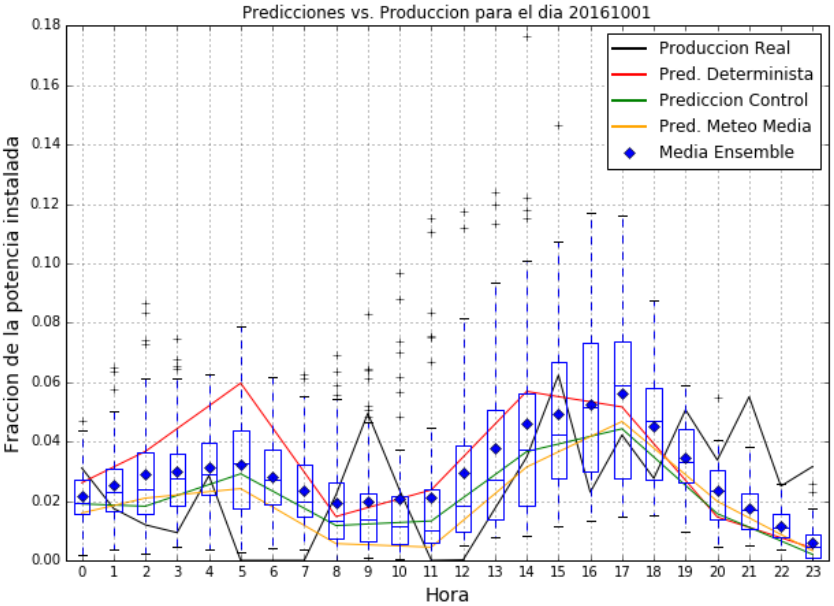

Figura 5.4: Evolución de las predicciones para Sotavento el 01/10/2016.

Es interesante ver cómo las predicciones no siguen patrones muy marcados, así como el hecho de que la predicción basada en la media de la meteorología y la de control se asemejan bastante, siendo algo mejor la primera en períodos de alta variabilidad. Esto es de esperar si tenemos en cuenta que ambas meteorologías son parecidas y se ejecutan con el mismo modelo. Por otro lado, podemos observar que las horas finales del día tienen una varianza mayor que las iniciales. Esto se debe a que descargamos la meteorología a las 00, por lo que las primeras horas del día están más actualizadas y tienen una menor dispersión.

Ahora bien, analizar lo que ocurre para algunos días en concreto tampoco es muy representativo, por lo que hemos estudiado las predicciones de las meteorologías determinista, de control y media para un período mayor, de nuevo Enero de 2016:

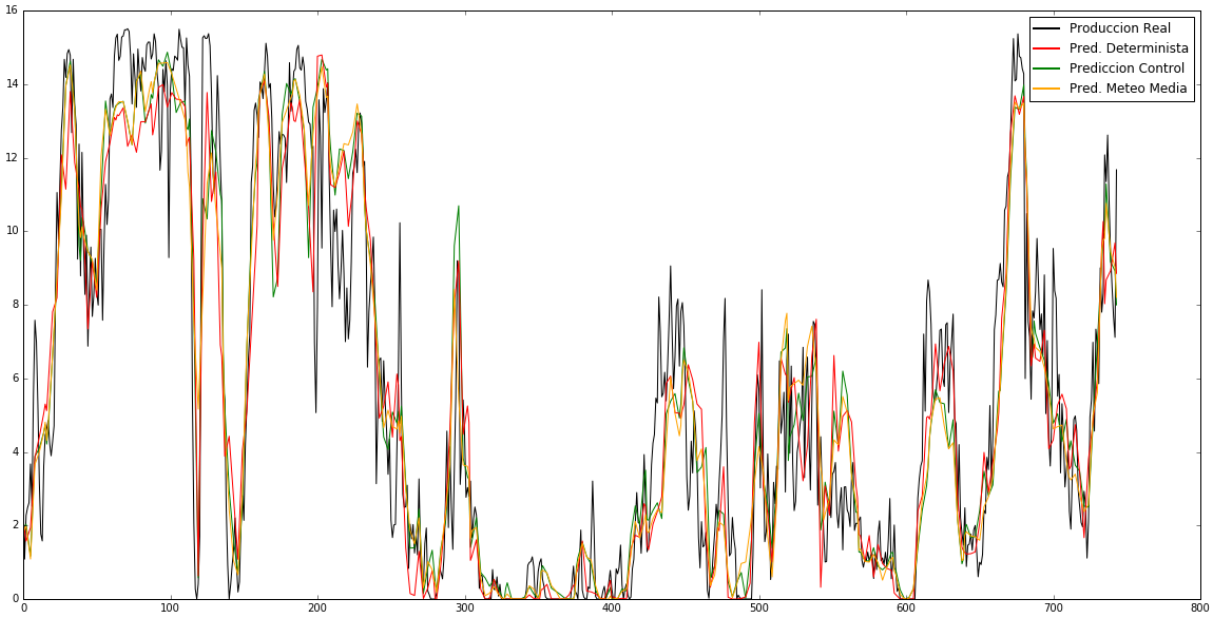

Figura 5.5: Evolución de las predicciones para Sotavento y Enero de 2016.

Como esperábamos la meteorología de control y la media del ensemble nos siguen dando resultados muy similares. Por otro lado, podemos apreciar que ninguna de las predicciones se asemeja verdaderamente a la producción. Debido a ello vamos a analizar el conjunto de todas las predicciones a un mismo tiempo, para lo cual hemos comenzado creando un intervalo de confianza de todas las predicciones. Como límites de dicho intervalo hemos seleccionado el primer y tercer cuartil del conjunto de todas las predicciones, con lo que uno esperaría que la producción se encontrase dentro de dicho intervalo en un 50 % de las veces. Sin embargo, los resultados no se asemejan a lo esperado ni mucho menos, habiendo obtenido que la producción queda con respecto al intervalo:

Por encima	32.564 %
En el intervalo	21.902 %
Por debajo	45.534 %

Tabla 5.6: Relación entre la producción y el intervalo de confianza.

Esto mismo se puede ver más fácilmente en la siguiente gráfica, la cual por proseguir con lo visto hasta ahora hace referencia a Enero de 2016:

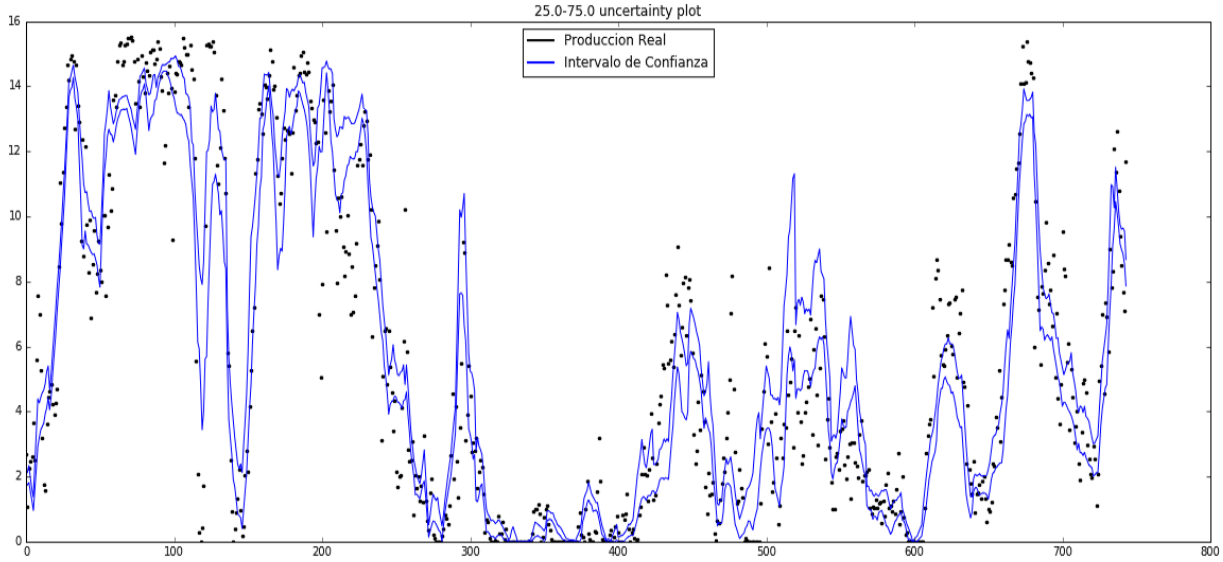

Figura 5.6: Intervalo de confianza para Sotavento y Enero de 2016.

Vemos que la anchura del intervalo no es especialmente grande, lo cuál indica que la dispersión de las predicciones tampoco lo es. Esto podría deberse a que hubiéramos delimitado demasiado el intervalo al tomar el primer y tercer cuartil. Si en su lugar tomáramos el máximo y el mínimo de las predicciones veríamos como este intervalo crece notablemente, incluyendo a la producción un número mucho mayor de horas:

Por encima	14.609 %
En el intervalo	63.051 %
Por debajo	22.340 %

Tabla 5.7: Relación entre la producción y el máximo y el mínimo de las predicciones.

Y la gráfica correspondiente:

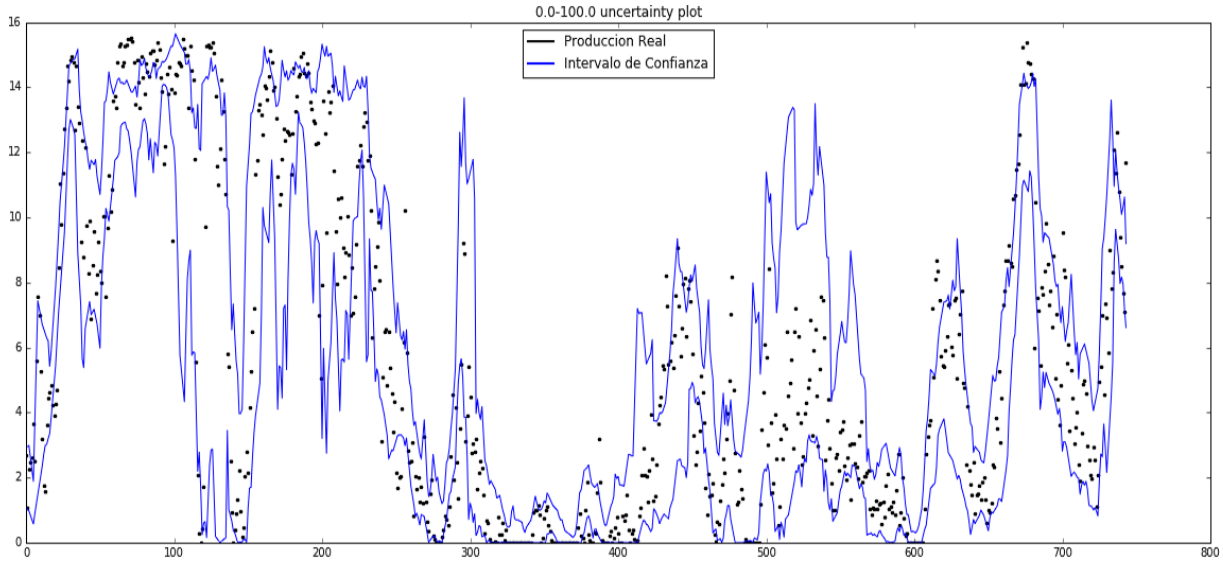

Figura 5.7: Máximos y mínimos para Sotavento y Enero de 2016.

En este caso la longitud del intervalo aumenta considerablemente, englobando a la producción en un porcentaje mucho mayor de ocasiones. Esto nos indica que, al quedarnos únicamente con los miembros más representativos del ensemble y descartar los outliers, estamos perdiendo una información de gran valor, en ocasiones más relevante que la que recabamos de las predicciones centrales, supuestamente más fiables. Pese a ello, el porcentaje de ocasiones en que la producción se encuentra dentro del intervalo sigue estando muy lejos del valor teórico esperado, que en este caso sería del 100 %.

Los resultados y conclusiones anteriores se referían exclusivamente al parque eólico Sotavento, si bien son perfectamente válidos también para el conjunto de parques de REE. El análisis, las tablas y las gráficas respectivas se pueden encontrar en el apéndice B.

6

Conclusiones y trabajo futuro

Conclusiones y trabajo futuro

Este TFG tenía como objetivo verificar la utilidad del Ensemble Prediction System para la predicción de la generación de energía eólica. Para ello se ha comenzado estudiando exclusivamente la meteorología, a fin de encontrar posibles problemas que pudieran aparecer al proseguir con el trabajo. Una vez visto que no debería haberlos se ha adaptado el sistema de predicción propio del IIC para su funcionamiento con esta fuente de meteorología. Finalmente, se ha realizado la extracción y el análisis de los resultados.

Como se puede observar en el capítulo anterior, en ningún momento hemos incluido el error de cada uno de los miembros del ensemble por separado, sino tan solo el de la predicción de control. La razón se encuentra en que en este trabajo no buscábamos medir la calidad individual de cada uno, sino el potencial de utilizar el conjunto de sus predicciones. A pesar de ello, sí hemos visto que cada una de las predicciones por separado nos daba un rendimiento peor que la predicción determinista. Esto era algo de esperar, ya que en principio la meteorología más ajustada a la realidad debiera ser ésta última, y por lo tanto sus resultados.

Por otro lado hemos observado que incluso dentro del propio ensemble existen diferencias en el nivel de acierto en las predicciones, siendo mejores los resultados de la predicción de control que los del resto. Esto también era perfectamente lógico, ya que si bien en ocasiones puede que la realidad se asemeje más a algún otro miembro del ensemble que a la de control, lo habitual es que esto no ocurra. Además, debemos tener en cuenta que hemos utilizado el mismo modelo para todas las predicciones, el cual entrenábamos únicamente para la de control, de forma que las características o patrones internos de cada miembro se pierden. Quedaría como trabajo pendiente por lo tanto estudiar la posibilidad de crear un modelo diferente para cada uno, lo cual en la actualidad supondría un problema tanto de memoria como de tiempo computacional (supondría entrenar y predecir 50 veces para cada parque y aproximación).

A continuación habíamos calculado la media de las predicciones de producción, obteniendo errores bastante asequibles en general, y si bien no tan buenos como con la meteorología determinista, sí mejores que con la de control. Esto demuestra que efectivamente la predicción basada en conjuntos mejora a aquella basada en una sola predicción, pese a lo cual las diferencias en el cálculo de la meteorología determinista y el EPS impiden que se alcance el mismo nivel de acierto. Es por ello que se deberían valorar otros sistemas de meteorología basados en ensembles,

como puede ser GEFS, que a ser posible utilizaran la misma resolución y modelo para todas las predicciones. Por otro lado, la media ponderada apenas mejora a la uniforme. Esto último tiene sentido si recordamos que al calcular esta media nos estamos centrando en las predicciones más probables, pese a que como ya hemos visto con los intervalos de confianza en muchas ocasiones las más representativas son aquéllas que no se esperan.

Por último habíamos probado a crear la media de la meteorología y ejecutar los modelos con dicha media. En esta ocasión podíamos advertir como los resultados variaban notablemente según el día, siendo algo mejores en períodos con alta variabilidad en la atmósfera. Se debería considerar la opción de crear una aproximación o modelo específico para esta meteorología, lo cual no supondría tanto problema como hacerlo para cada una de las predicciones del ensemble.

Una vez lanzadas todas las opciones anteriores habíamos generado unos intervalos de confianza de la producción. Sin embargo, incluso tomando como límites el máximo y el mínimo de las predicciones, en más de un 20-30 % de las horas la producción se encontraba fuera del intervalo. Esto nos indica que la dispersión del ensemble, o al menos de las predicciones obtenidas con cada uno de sus miembros, no es lo suficientemente alta como para conseguir englobar a la producción. Por lo tanto, parece difícil que podamos sacar resultados mucho mejores que los anteriores mediante este sistema. Pese a ello, no podemos descartar que se puedan conseguir siguiendo alguna de las vías aún abiertas, o más adelante según vayan mejorando los sistemas de predicción de este tipo.

Por otro lado me gustaría recordar que para este trabajo hemos empleado únicamente dos tipos de algoritmos: Support Vector Machine y Multilayer Perceptron. No obstante se deberían valorar otras opciones como Random Forest o Gradient Boosting, las cuales parecen estar dando buenos resultados para la meteorología determinista.

La elaboración de este TFG me ha permitido abordar temas que hasta ahora desconocía y que me han resultado de gran interés, como son el caos atmosférico y el (todavía oscuro) funcionamiento del mercado eléctrico. Tanto para la elaboración de las gráficas y los resultados, como para la conversión de la meteorología, hemos utilizado el lenguaje Python y sus librerías, mientras que para la ejecución de los modelos nos hemos servido de C, SQL y el lenguaje de scripts de bash. Esto me ha permitido retomar y mejorar mis conocimientos en problemas propios de la ingeniería informática, como la programación o el manejo de bases de datos. Finalmente, el análisis de las variables meteorológicas y las predicciones, así como el estudio de los algoritmos de aprendizaje automático, me ha permitido introducirme en el mundo de la ciencia de datos, el cual considero de gran relevancia para el futuro.

Bibliografía

[1] Red Eléctrica de España. Nota de prensa: La demanda de energía eléctrica aumenta un 2 % en diciembre. http://www.ree.es/es/sala-de-prensa/notas-de-prensa/2016/12/la-demanda-de-energia-electrica-aumenta-un-2-en-diciembre.

[2] Asociación Eólica Empresarial. ¿Quieres saber por qué cuando sopla el viento baja el precio de la luz? https://www.aeeolica.org/es/sobre-la-eolica/la-eolica-y-el-precio-de-la-luz/.

[3] ECMWF. The ECMWF Ensemble Prediction System. The rationale behind probabilistic weather forecasts. https://www.ecmwf.int/sites/default/files/elibrary/2012/14557-ecmwf-ensemble-prediction-system.pdf.

[4] Jason Brownlee. Crash Course On Multi-Layer Perceptron Neural Networks. http://www.machinelearningmastery.com/neural-networks-crash-course/.

[5] Edward N. Lorenz. Deterministic Nonperiodic Flow. *Journal of Atmospheric Sciences*, 20(2):130–141, 1963.

[6] Edward N. Lorenz. Three approaches to atmospheric predictability. *Bulletin of American Meteorological Society*, 50:345–351, 1969.

[7] Eugenia Kalnay. "Fighting chaos" in weather and climate prediction. https://www.atmos.umd.edu/~ekalnay/pubs/Chaos-Predictability-EnKF-WMOtalk.pdf.

[8] Martin Ehrendorfer. The Liouville Equation and Its Potential Usefulness for the Prediction of Forecast Skill. Part I: Theory. *Monthly Weather Review*, 122(4):703–713, 1994.

[9] Martin Ehrendorfer. The Liouville Equation and Its Potential Usefulness for the Prediction of Forecast Skill. Part II: Applications. *Monthly Weather Review*, 122(4):714–728, 1994.

[10] Parque Eólico Sotavento. http://www.sotaventogalicia.com/es/datos-tiempo-real/historicos.

[11] Boletín Oficial del Estado. *Resolución de 1 de agosto de 2013, de la Secretaría de Estado de Energía, por la que se aprueban las reglas de funcionamiento del mercado diario e intradiario de producción de energía eléctrica y el cambio de hora de cierre del mercado diario.* Ministerio de Industria, Energía y Turismo, Agosto 2013.

[12] World Meteorological Organization. *Guidelines on Ensemble Prediction Systems and Forecasting.* Geneva, Switzerland, March 2012.

[13] D. S Richardson. Ensemble forecasting, Febrero 2015.

[14] National Centers for Environmental Prediction. *Ensemble Prediction Systems. A basic training manual targeted for operational meteorologists*, Julio 2016. http://www.wpc.ncep.noaa.gov/ensembletraining/.

[15] European Centre for Medium-Range Weather Forecasts. `http://www.ecmwf.int/en/forecasts/charts/product-descriptions/Medium-range%20forecasts`.

[16] Alex J. Smola and Bernhard Schölkopf. A Tutorial on Support Vector Regression. *Statistics and Computing*, 14(3):199–222, Agosto 2004.

[17] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[18] Martin Riedmiller. Machine Learning: Multi Layer Perceptrons. `http://ml.informatik.uni-freiburg.de/_media/teaching/ss10/05_mlps.printer.pdf`.

[19] James McCaffrey. Test Run - L1 and L2 Regularization for Machine Learning. *MSDN Magazine*, 30(2):1415–1442, Febrero 2015.

[20] Corinna Moehrlen. *Uncertainty in wind energy forecasting*. PhD thesis, University College Cork, 2004.

[21] European Wind Energy Conference (EWEC). *Application of a Multi-Scheme Ensemble Prediction System for wind power forecasting in Ireland and comparison with validation results from Denmark and Germany* . WEPROG, 2006.

[22] ECMWF. MARS user documentation. `https://software.ecmwf.int/wiki/display/UDOC/MARS+user+documentation`.

[23] Jeff Whitaker. Python interface for reading and writing GRIB data. `https://www.github.com/jswhit/pygrib`.

A

Descarga y conversión de los datos de meteorología

A.1. Descarga de los datos

El ECMWF dispone de un sistema de acceso y descarga de datos propio, conocido como MARS catalogue [22]. Para acceder a dicho sistema y obtener los datos se necesita (además de una cuenta propia) realizar peticiones con un formato concreto al servidor. Dentro de estas consultas se pueden especificar diversos parámetros para la descarga como las fechas, la zona o la resolución.

La estructura general de las consultas es la siguiente:

```
verb,
keyword1 = value 1,
...      = ...,
keywordN = value N
```

, donde:

- Verb: Hace referencia a la acción que se va a llevar a cabo. En nuestro caso será siempre retrieve, que es aquella que nos permite extraer los datos del servidor, si bien existen otras como compute o read para manejo de los datos.
- Keyword: Hace referencia a alguna de las directrices o variables que podemos especificar para la acción escogida. Algunas de ellas son imprescindibles para seleccionar los daos de referencia por ejemplo.

Algunos de los parámetros que podemos especificar para el comando retrieve son:

- Class: Especifica el tipo de datos a descargar según la clasificación dada por el ECMWF. Pueden ser datos rutinarios (od), de experimentos de investigación (rd), de proyectos o países en concreto asociados al centro, etc.
- Stream: Identifica el sistema utilizado para generar la predicción. Habitualmente tomaremos como tal oper, que hace referencia al modelo atmosférico.

- Expver: Versión de los datos. En el caso de los datos generados periódicamente su identificador será 1 o 2.
- Type: Determina el tipo de datos, es decir, si son observaciones o predicciones. Si es la predicción de siempre (fc), si es la de control (cf) o si son las obtenidas de perturbar las condiciones iniciales (pf).
- Levtype: Denota el nivel al que se calculan los datos. En nuestro caso todas las variables se toman a nivel de superficie (sfc) si bien existen otras opciones como al nivel del modelo (ml) o profundo (dp).
- Param: Especifica las variables de meteorología que se quieren descargar. El formato debe ser var1/var2/.../varN, donde el var es un código único de la variable. Por ejemplo para la presión tomaremos 134.128 y para la temperatura 167.128.
- Date: Fechas en las que se ejecutaron los modelos y, por tanto, crearon los datos. Debe seguir el formato fecha1/fecha2/.../fechaN o bien fecha1/to/fechaN.
- Time: Horas y minutos de las ejecuciones, con formato time1/time2/.../timeN. Habitualmente se toman las horas 00:00, 06:00, 12:00 o 18:00.
- Step: Especifica los horizontes para los que se van a descargar los datos. El formato puede ser hora1/hora2/.../horaN o bien hora1/to/horaN/by/salto.
- Grid: Especifica el formato de la salida. Puede ser tanto Gaussian Grid como Lat/Lon grid, en cuyo caso especificaremos la resolución de la latitud seguida de la de la longitud.
- Area: Se refiere a la zona para la que descargamos los datos. Puede ser un área predefinida o bien una propia, para lo cual escribiremos los límites de la latitud y longitud como Norte/Oeste/Sur/Este. En el caso de la Península tendríamos por ejemplo 44/-9.5/35.5/4.5,
- Target: Selecciona un fichero en el que guardar la descarga, por supuesto en formato GRIB.

Es posible obtener los datos de meteorología con resoluciones mayores a las especificadas por el ECMWF, para lo cual a nivel interno se lleva a cabo una interpolación de los datos, si bien no es recomendable hacerlo con resoluciones menores que éstas ya que serían engañosas. En nuestro caso hemos optado por trabajar con resolución 0.125° para la predicción determinista y 0.25° para las del ensemble. La razón se encuentra en que aunque en la actualidad se pueden obtener ambas con resoluciones menores este cambio se hizo a partir de Marzo de 2016, por lo que el histórico se había calculado previamente con esas resoluciones.

Por otro lado en este trabajo tan solo vamos a trabajar con las siguientes variables, si bien del cara al futuro se deberían considerar otras que pudieran tener una alta correlación con la generación de energía eólica:

- 10V: Velocidad de la componente V del viento a 10 metros (m/s). Id 166.128.
- 10U: Velocidad de la componente U del viento a 10 metros (m/s). Id 165.128.
- 100V: Velocidad de la componente V del viento a 100 metros (m/s). Id 246.228.
- 100U: Velocidad de la componente U del viento a 100 metros (m/s). Id 247.228.
- 2 metre temperature: Temperatura a 2 metros (K). Id 167.128.
- Surface pressure: Presión a nivel de superficie (Pa). Id 134.128.

A continuación se pueden ver ejemplos de consultas para toda la Península, tanto para la predicción de control como para el resto del ensemble:

```
retrieve,
class=od,
stream=enfo,
expver=1,
type=cf,
levtype=sfc,
param=134.128/165.128/166.128/167.128/246.228/247.228,
date=2014-01-01/to/2016-10-31,
time=00/12,
step = 0/to/48/by/3,
grid = 0.25/0.25, # Resolución latitud/longitud
area = 44/-9.5/35.5/4.5,  # Peninsula
target="/path/ControlPeninsulaHistorico.grib"
```

Listing A.1: Petición para la predicción de control.

```
retrieve,
class=od,
stream=enfo,
expver=1,
type=pf,
levtype=sfc,
param=134.128/165.128/166.128/167.128/246.228/247.228,
date=2015-11-01/to/2016-10-31,
time=00,
step = 0/to/48/by/3,
grid = 0.25/0.25, # Resolución latitud/longitud
area = 44/-9.5/35.5/4.5,  # Peninsula
target="/path/RestoEnsemblePeninsulaHistorico.grib"
```

Listing A.2: Petición para los demás miembros del EPS.

Como se puede ver, la petición para la predicción de control difiere un poco con respecto a la del resto de miembros del ensemble. El primer cambio consiste en pasar de cf a pf en el tipo, lo cuál es obligatorio. Las modificaciones siguientes en cambio se refieren al date y al time. En el caso de las fechas pasamos de descargar los datos desde el 01/01/2014 a empezar en el 01/11/2015. La razón se encuentra en que como se explica en el capítulo 5, únicamente vamos a entrenar los modelos para la predicción de control, por lo que no necesitamos ese período para predecir. Además, en este caso time pasa de 00/12 a 00, lo que quiere decir que ya no vamos a descargar los datos obtenidos a las 12. Esto implica que al predecir para las horas de la tarde tomaremos la meteorología obtenida a las 00, ya que será la más reciente, mientras que para la predicción de control y la determinista tomaremos la de las 12 (de forma que las horas de la tarde podrían ser algo más precisas para estas dos últimas predicciones). No obstante, la idea de utilizar un conjunto de predicciones tiene como objetivo precisamente minimizar el error al alejarnos en el tiempo y, además, en 12 horas apenas hay tiempo para que las diferencias se noten demasiado.

A.2. Conversión de los datos

Una vez finalizada la descarga de los datos obtendremos un fichero con formato GRIB binario para la predicción de control y otro para las demás predicciones del EPS. Para poder utilizar dichos ficheros será necesario por tanto convertirlos a un formato legible y sencillo de ejecutar. Lo más fácil en este caso ha sido crear un archivo para cada cada día y predicción, de forma que se puedan ejecutar los modelos de aprendizaje de forma diaria y para cada miembro del ensemble por separado. En el caso de la predicción de control este paso se ha realizado mediante el conversor habitual del IIC, el mismo que para la predicción determinista. Este conversor no nos servía para el otro fichero, que consta de 50 predicciones, por lo que su conversión se ha realizado mediante la librería Pygrib [23] de Python, partiendo del trabajo previo realizado en la Cátedra IIC-UAM.

Además de para la conversión directa de los datos nos hemos servido también de la librería Pygrib para obtener la media de las predicciones del ensemble. Dicha media nos ha servido en ocasiones como datos meteorológicos de partida para la ejecución de los modelos, siendo más fiable que la predicción de control en épocas de gran variabilidad. Además, partiendo de lo anterior no ha sido necesario realizar apenas modificaciones sobre el código.

B

Resultados de los parques de REE

El mismo análisis que realizamos para Sotavento lo hemos llevado a cabo para las dos aproximaciones de REE. Aunque los resultados y las conclusiones son muy similares hemos decidido incluir aquí sus gráficas y sus tablas.

B.1. Aproximación con todas las variables: SVR-1

A continuación podemos ver la evolución de las predicciones y la producción para los mismos días que vimos para Sotavento, es decir, 01/01/2016, 01/04/2016, 01/07/2016 y 01/10/2016:

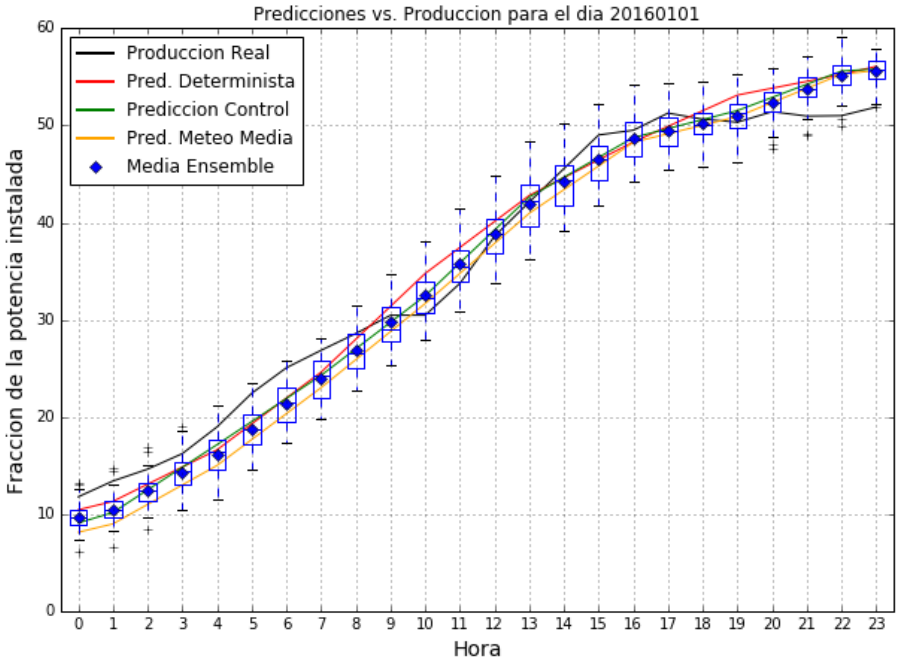

Figura B.1: Evolución de las predicciones para la primera aproximación de REE el 01/01/2016.

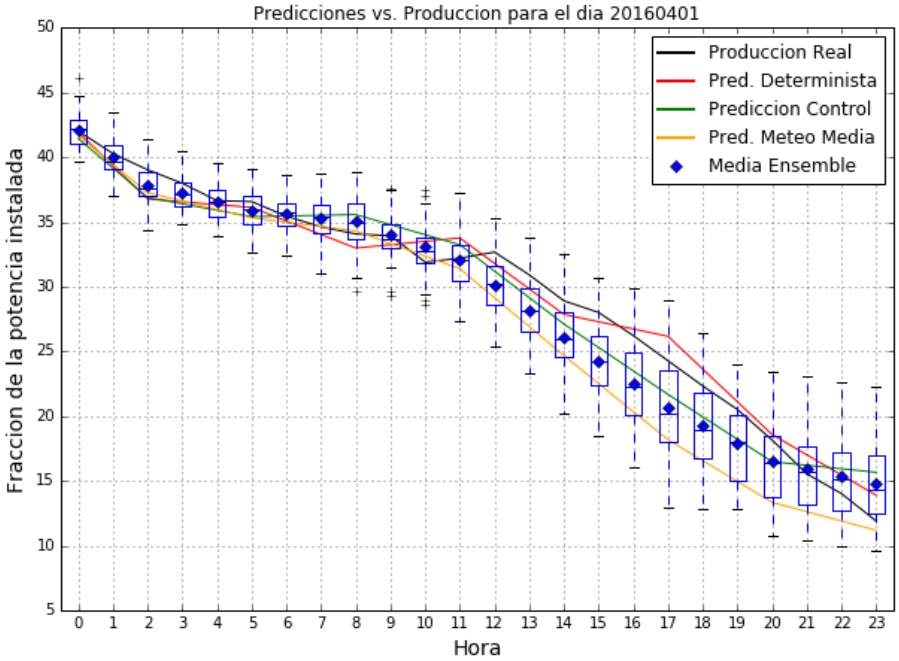

Figura B.2: Evolución de las predicciones para la primera aproximación de REE el 01/04/2016.

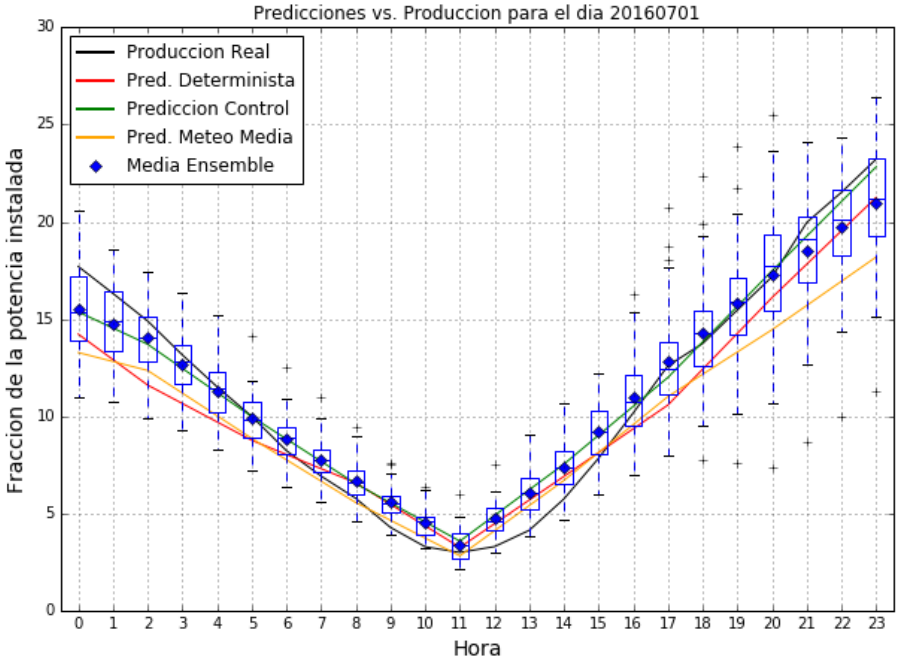

Figura B.3: Evolución de las predicciones para la primera aproximación de REE el 01/07/2016.

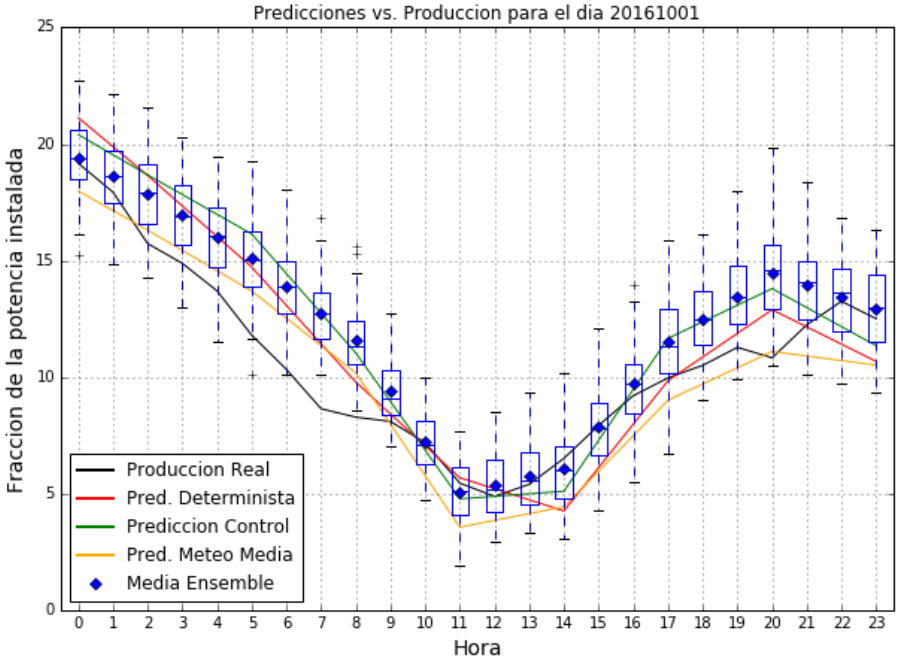

Figura B.4: Evolución de las predicciones para la primera aproximación de REE el 01/10/2016.

Igualmente hemos incluido las predicciones para todo el mes de Enero de 2016:

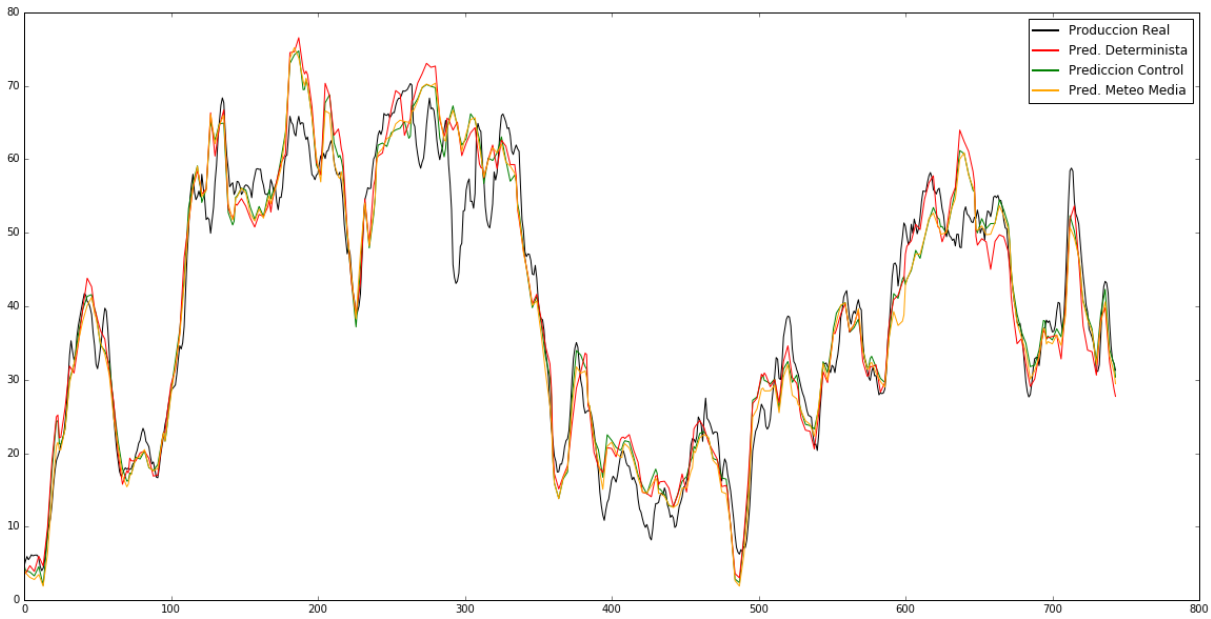

Figura B.5: Evolución de las predicciones para la primera aproximación de REE.

Y la relación con el intervalo de confianza, tomando como límites los cuartiles:

Por encima	28.488 %
En el intervalo	34.517 %
Por debajo	36.995 %

Tabla B.1: Relación entre la producción y el intervalo de confianza.

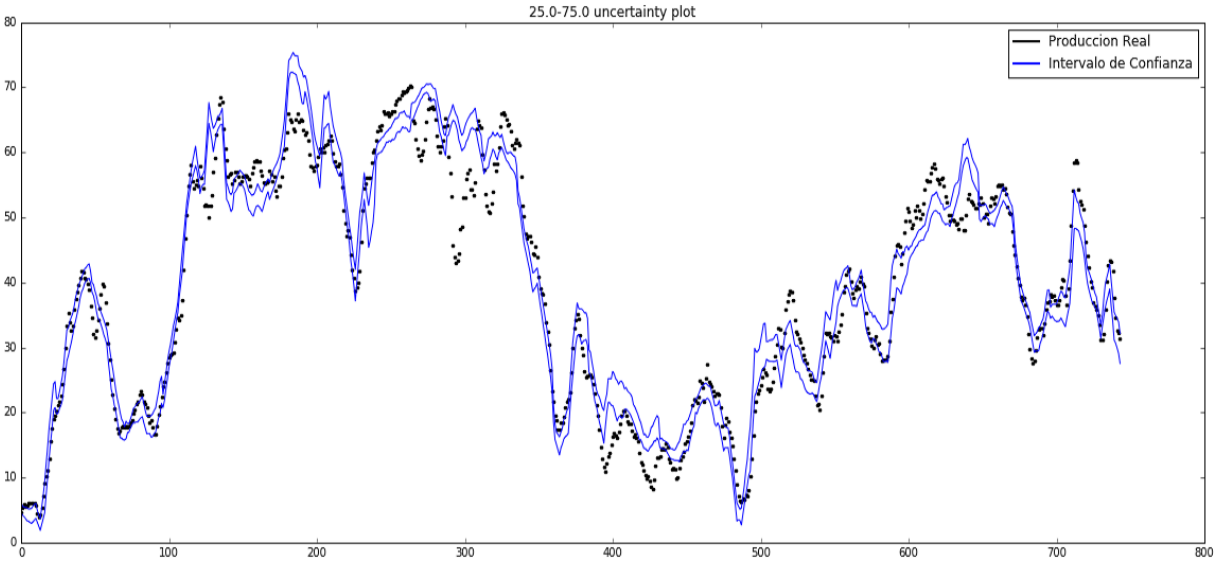

Figura B.6: Intervalo de confianza para la primera aproximación de REE y Enero de 2016.

Y el máximo y el mínimo:

Por encima	5.926 %
En el intervalo	86.264 %
Por debajo	7.810 %

Tabla B.2: Relación entre la producción y el máximo y el mínimo de las predicciones.

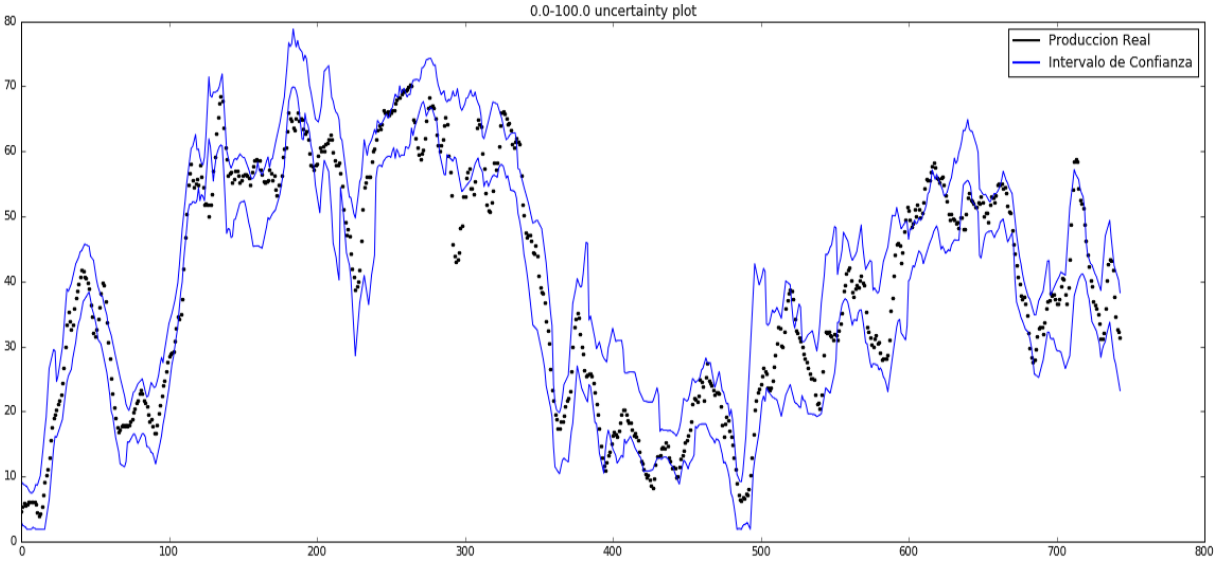

Figura B.7: Máximos y mínimos para la primera aproximación de REE y Enero de 2016.

Al igual que ocurría en Sotavento los intervalos con el primer y tercer cuartil son demasiado pequeños. No obstante al tomar los máximos y mínimos como límites el porcentaje de horas con producciones dentro del intervalo sube hasta el 86.264 %. Esto nos indica que las predicciones no se encuentran tan lejos de la producción como en Sotavento, lo cual se corresponde con el hecho de que los errores sean mucho menores en REE.

B.2. Aproximación sin las variables de viento a 10 metros: SVR-2

La evolución de las predicciones y la producción para los mismos días que hasta ahora:

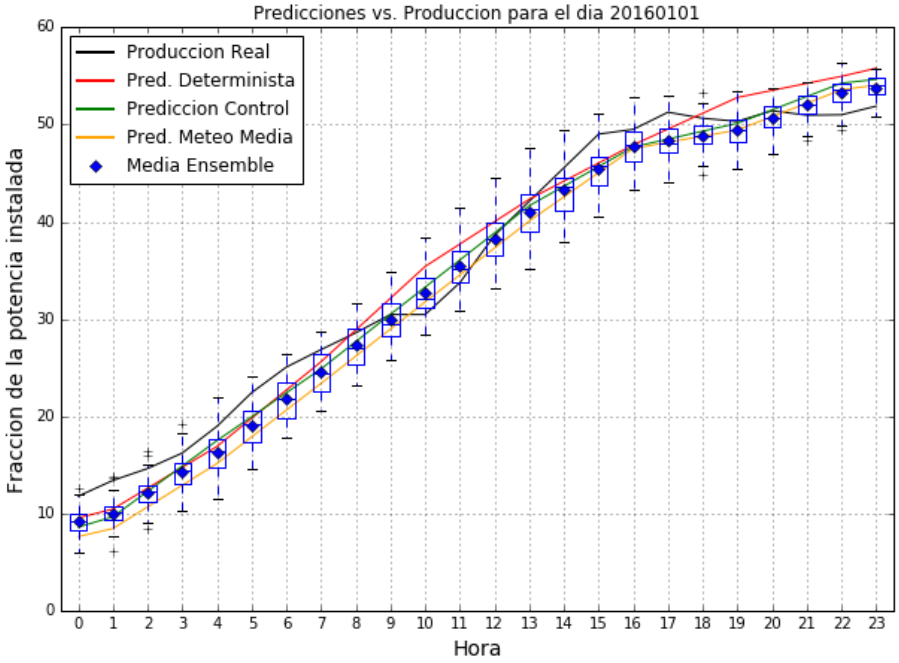

Figura B.8: Evolución de las predicciones para la segunda aproximación de REE el 01/01/2016.

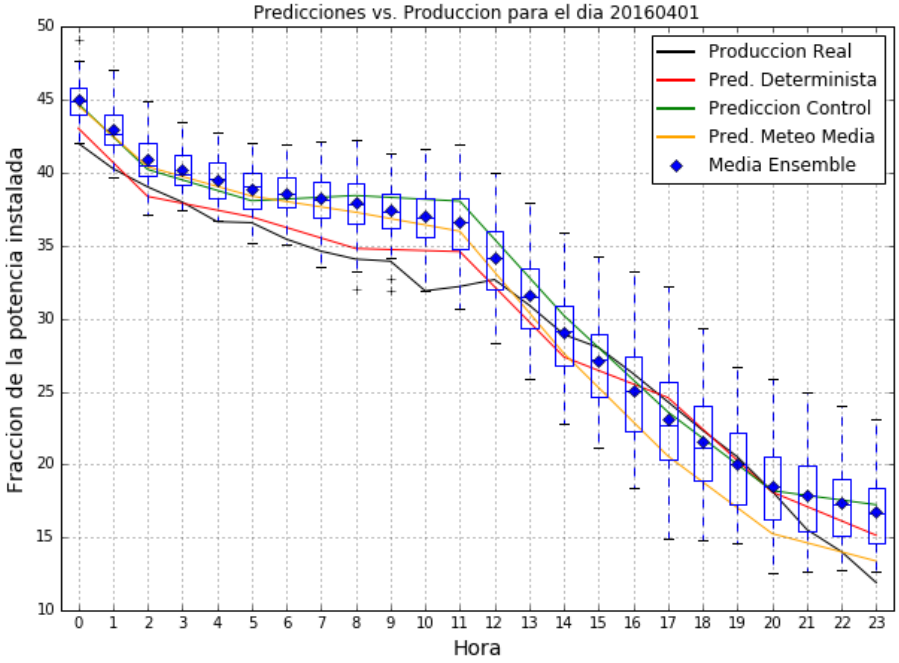

Figura B.9: Evolución de las predicciones para la segunda aproximación de REE el 01/04/2016.

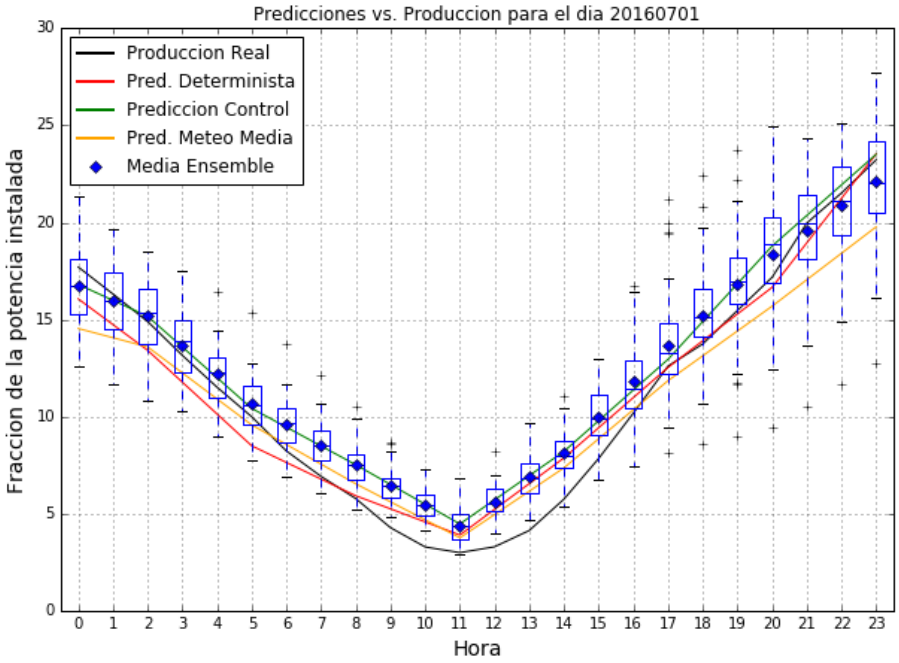

Figura B.10: Evolución de las predicciones para la segunda aproximación de REE el 01/07/2016.

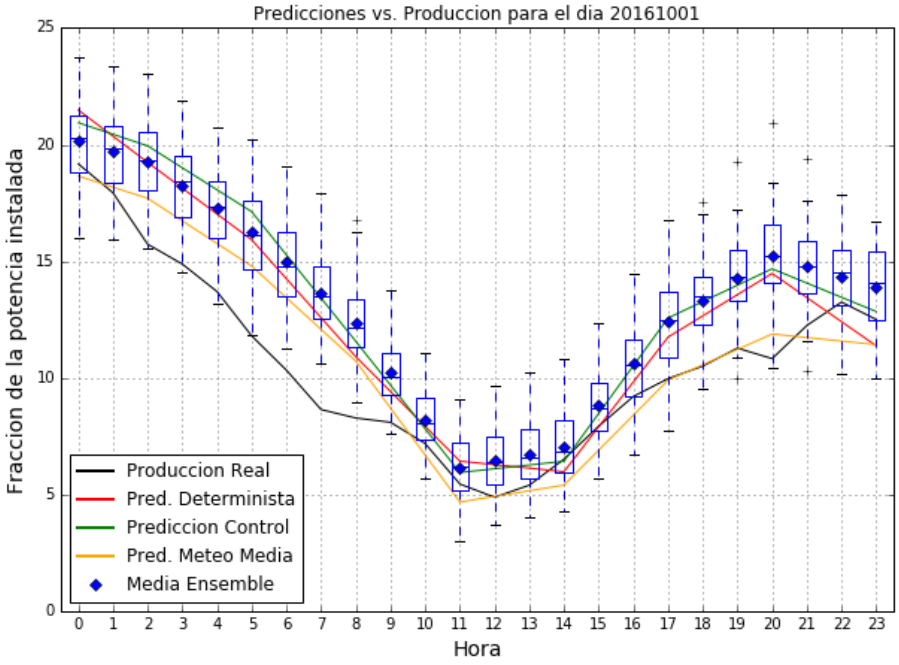

Figura B.11: Evolución de las predicciones para la segunda aproximación de REE el 01/10/2016.

La predicción para Enero de 2016 al completo:

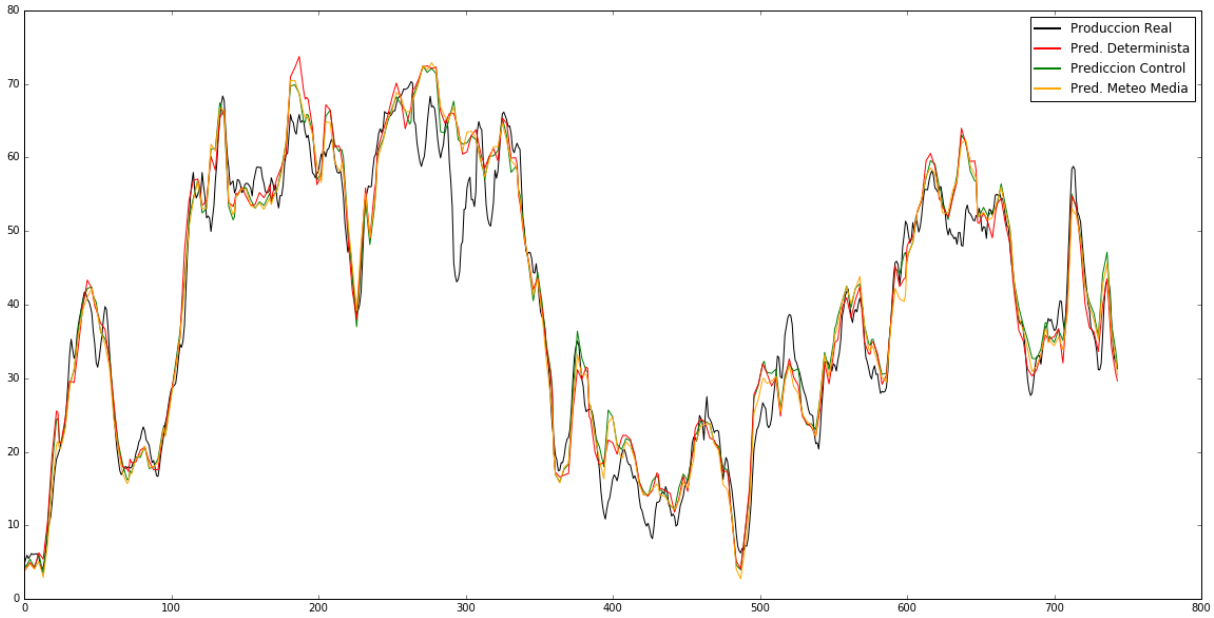

Figura B.12: Evolución de las predicciones para la segunda aproximación de REE.

El intervalo de confianza tomando los cuartiles:

Por encima	21.829 %
En el intervalo	33.170 %
Por debajo	45.001 %

Tabla B.3: Relación entre la producción y el intervalo de confianza.

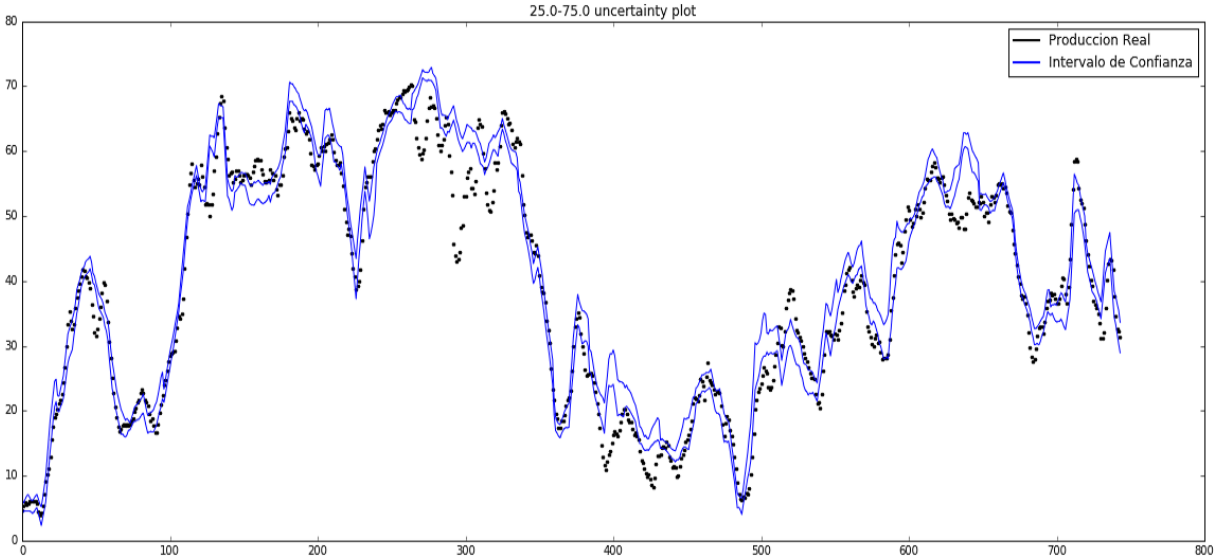

Figura B.13: Intervalo de confianza para la segunda aproximación de REE y Enero de 2016.

Y tomando como límites el máximo y el mínimo:

Por encima	3.758 %
En el intervalo	84.275 %
Por debajo	11.968 %

Tabla B.4: Relación entre la producción y el máximo y el mínimo de las predicciones.

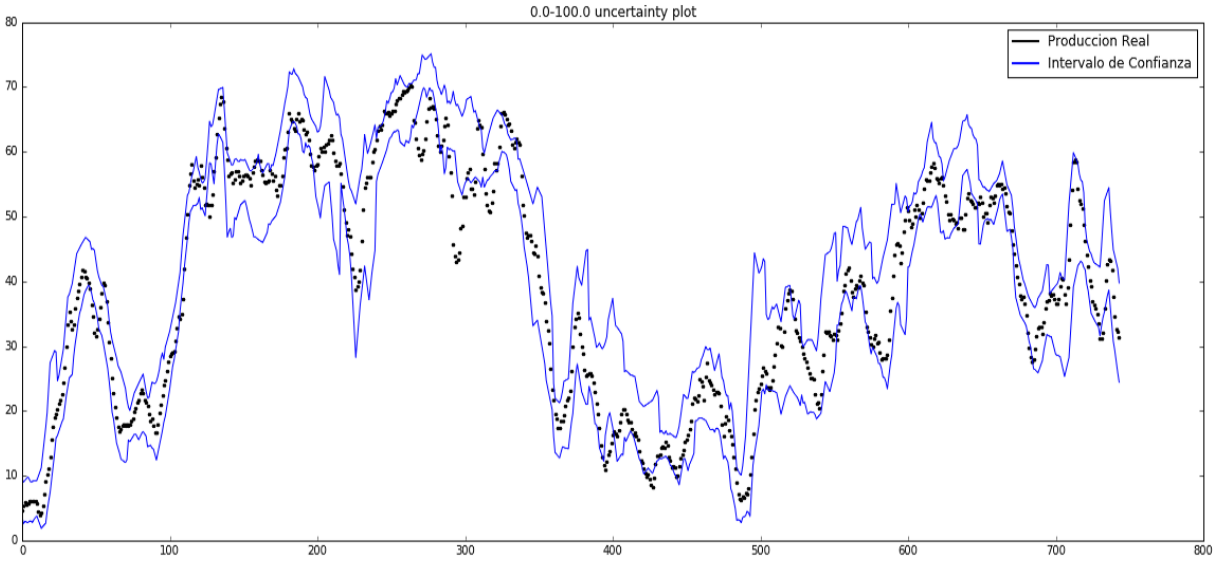

Figura B.14: Máximos y mínimos para la segunda aproximación de REE y Enero de 2016.

En esta aproximación el número de producciones por debajo de los intervalos aumenta considerablemente con respecto a la anterior. Esto se corresponde con el hecho de que el NMBE de las predicciones sea mayor que en la primera aproximación, es decir, las predicciones toman valores superiores en este caso. La razón podría encontrarse en que no hemos incluido las variables de viento a 10 metros, que podrían compensar esta sobrestimación.