UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE MEDELLÍN
FACULTAD DE MINAS

DYNA

http://dyna.medellin.unal.edu.co/

Variabilidad de la velocidad de flujo durante el ciclo anual a partir de técnicas de inteligencia artificial.

Paula Andrea Espinosa Ordoñez [a]

[a] *Facultad de Minas, Universidad Nacional de Colombia, Medellín, Colombia. paespinosao@unal.edu.co*

Resumen

La aleatoriedad de las variables oceánicas y de los procesos que las generan dificultan el entendimiento físico de los fenómenos que ahí toman lugar, por ello se recurre a modelos numéricos los cuales resuelven las ecuaciones de la física bajo unas consideraciones, pero en muchas ocasiones estos modelos representan un gasto computacional grande frente a los resultados que entregan. Es así como en el presente trabajo se proponen el uso de modelos basados en datos con el fin de ampliar sus aplicabilidades dependiendo de los problemas específicos, y disminuir el gasto computacional que los modelos numéricos representan. De esta forma, se evaluaron modelos de Machine Learning paramétricos y no paramétricos con el fin de determinar su capacidad al representar la variabilidad de las velocidades de flujo a lo largo del ciclo anual. En los resultados se encontró que el modelo KNN representó con gran aproximación estos patrones, sin embargo, los modelos de regresión lineal lograron capturar la variabilidad en el rango medio de los patrones de velocidad, por lo tanto la decisión de aplicar uno u otro dependerá del problema específico.

1 Introducción

Las observaciones del océano son de particular interés para el entendimiento en áreas como clima, diseño de obras offshore, y estudios de riesgo, ante eventos naturales extremos en la atmósfera/océano, una de las variables más importantes son las velocidades de las corrientes oceánicas, estas permiten cuantificar la transferir de momentum, y el transporte advectivo de propiedades del océano (salinidad, temperatura), sustancias contaminantes o migración de especies a lo largo del año.

Diversos estudios (Escobar et al., 2015; Liu & Weisberg, 2005; Posada et al., 1997) se enfocan en entender los patrones de movimiento que las velocidades generan. Diversas metodologías son aplicadas en cada uno de estos estudios, desde el uso de modelos hidrodinámicos (Jouon et al., 2006) acoplados hasta el uso de modelos basados en la estadística y el aprendizaje automático (Jirakittayakorn et al., 2017; Liu & Weisberg, 2005). La ventaja de los modelos hidrodinámicos se centra en aplicar métodos numéricos; diferencias finitas, volúmenes finitos; para solucionar las ecuaciones que representan la transferencia de momentum y masa en el océano, sin embargo el gasto computacional y el tiempo de modelado suele ser alto, aunque estos modelos permiten comprender la física de los fenómenos si no se tiene conocimiento de cada uno de los parámetros que muchos de ellos suelen integrar se pierde el sentido físico y los resultados pueden ser sesgados. Por su parte los modelos basados datos cuentan con ventaja frente a la optimización en el gasto computacional y tiempo de cómputo, pero se pierde la causalidad explicada a través de las ecuaciones que representan la física. Por lo tanto, la aplicabilidad de los modelos dependerá de la finalidad y el nivel de detalle que se precise en un problema, por ejemplo, la compresión y análisis de procesos altamente detallados, posiblemente requiere del uso de modelos hidrodinámicos, sin embargo en campos como la oceanografía operacional donde normalmente se requieren los valores o patrones de variables oceánicas en el régimen medio, y en un rango temporal corto, o para predicciones climatológicas a escala horaria – diaria – mensual, los modelos basados en datos suelen representar las magnitudes de estas variables, aplicando el metodologías que parten de previas predicciones hechas con modelos basados en la física o datos registrados en campo.

Se ha encontrado que los modelos basados en datos se impulsan a partir del entendimiento de la física y las variables que correlacionan la predicción del fenómeno, así como de la cantidad de información disponible. Por ejemplo, muchos centros de investigación a través de los años han usado modelos hidrodinámicos calibrados y validados con información real, el uso de estos datos validados puede ser de gran utilidad para crear modelos basados por datos, el propósito del presente trabajo por lo tanto es poder desarrollar un modelo basado en datos que represente la variabilidad de la velocidad de flujo en una zona insular del caribe, a partir de datos recolectados de bases de datos oceanográficas y atmosféricas.

1.1 Área de Estudio

La zona de estudio se eligió en un área del mar Caribe limitada entre Colombia y los países centro americanos; Panamá, Costa Rica y Nicaragua **Figura 1**. Las velocidades de flujo normalmente vienen del noreste – este, sin embargo, a lo largo de la costa se presentan cambios en la dirección que permiten la formación de patrones, como el giro de panamá.

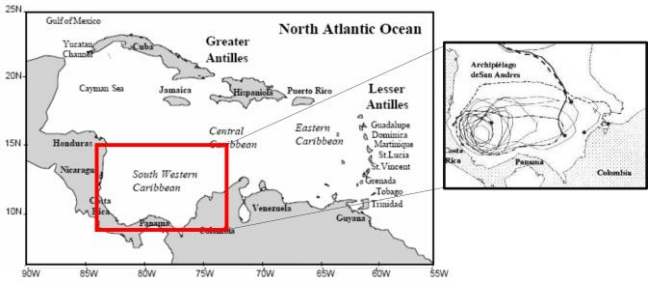


Figura 1. Área de estudio. Tomada y modificada de (Alberto & Amaya, 2001).

2 Metodología

En la presente sección se describen procedimientos usados para la formación de la base de datos, el análisis general de las variables independientes, y los modelos usados para analizar el objetivo planteado.

2.1 Variables y análisis exploratorio de datos

La configuración de la base de datos se realizó a partir de la descarga de datos de las bases de datos HYCOM (link de descarga) para información oceanográficas; temperatura, salinidad, nivel del agua y velocidad del flujo (gráficas a), b), c) y f) en la **Figura 2**), y la base de datos ERA 5 (link de descarga) para la altura de ola y la velocidad del viento (d) y e) en **Figura 2**). Las celdas de cada ráster tienen una resolución de 8 km × 4 km y una resolución temporal mensual, del año 2021.

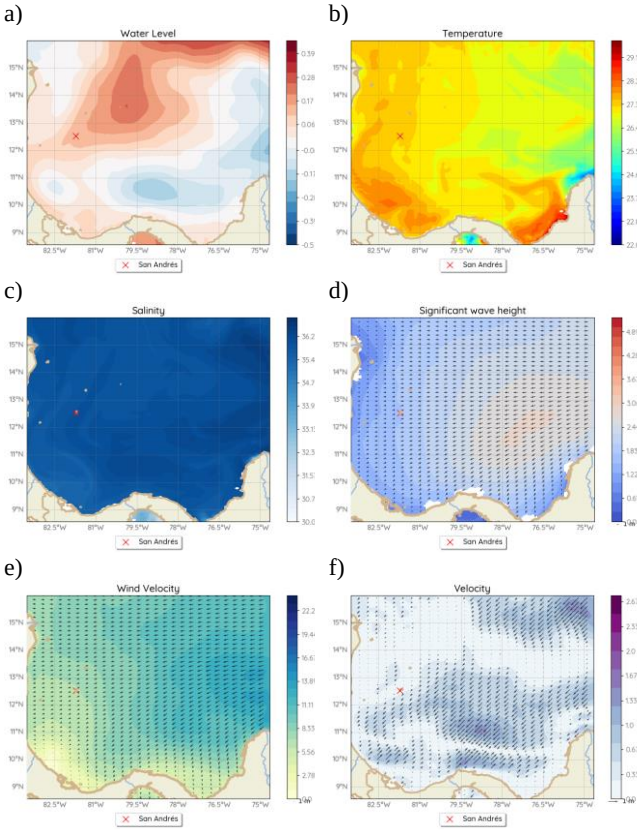


Figura 2. Representación de las variables usadas; variables independientes a), b), c), d), e); variables dependientes f).

Para la construcción del DataFrame que representa la base de datos inicialmente se enumeraron la cantidad de celdas para los ráster pertenecientes a cada mes, posteriormente se hizo un ajuste a la forma de cada ráster para que se convirtiera en un vector, luego se apilaron los correspondientes al mes 1, seguidos por el mes 2 hasta alcanzar el mes 12, esto con el fin de indicar el mes al que pertenecía cada celda. Posteriormente se hizo una máscara usando la máscara base de las variables a predecir, luego se verifico que todas las celdas tuvieran información. Esta configuración se hizo con el fin de garantizar la espacialidad y temporalidad de cada ráster. En la siguiente tabla se muestra el resultado obtenido.

Tabla 1. Data Frame de la base de datos.

	index	waterlevel	temp	sal	hs_wave	dir_wave	vel_wind	dir_wind	u_wind	v_wind	vel_flow	dir_flow	u_flow	v_flow	mes
0	46	0.260220	0.260220	0.260220	0.116837	0.116837	0.116837	45.000000	0.116837	0.116837	0.368006	45.0	0.260220	0.260220	1
1	47	0.263163	0.263163	0.263163	0.122465	0.122465	0.122465	45.000000	0.122465	0.122465	0.372168	45.0	0.263163	0.263163	1
2	48	0.258976	0.258976	0.258976	0.130494	0.130494	0.130494	45.000000	0.130494	0.130494	0.366247	45.0	0.258976	0.258976	1
3	49	0.256825	0.256825	0.256825	0.140012	0.140012	0.140012	45.000000	0.140012	0.140012	0.363206	45.0	0.256825	0.256825	1
4	50	0.258829	0.258829	0.258829	0.149988	0.149988	0.149988	45.000000	0.149988	0.149988	0.366040	45.0	0.258829	0.258829	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
221371	20939	0.216121	28.110162	35.755665	1.887047	87.019027	9.213161	259.490423	-8.908047	-1.652549	9.781339	225.0	-0.552490	-0.552490	12
221372	20940	0.222498	28.105264	35.748895	1.888027	87.102830	9.229263	259.623510	-8.930887	-1.635659	9.802535	225.0	-0.567478	-0.567478	12
221373	20941	0.226166	28.097134	35.737414	1.888999	87.187034	9.245683	259.752242	-8.953703	-1.618732	9.822991	225.0	-0.581518	-0.581518	12
221374	20942	0.229069	28.081057	35.720507	1.889970	87.271393	9.257756	259.873328	-8.972129	-1.602489	9.838176	225.0	-0.592680	-0.592680	12
221375	20943	0.231113	28.077571	35.710580	1.890939	87.355596	9.262286	259.979580	-8.983224	-1.587286	9.848334	225.0	-0.599862	-0.599862	12

221376 rows × 15 columns

Como variables independientes **X** se definen todas indicadas en data frame a excluyendo la columna index que representa el indicador de la celda, y las componentes de la velocidad de flujo u_flow y v_flow. Como variables dependientes inicialmente se definieron dos **Y1**: velocidad de flujo (vel_flow en **Tabla 1**) y **Y2**: dirección del flujo (dir_flow en **Tabla 1**), finalmente luego de hacer el análisis exploratorio de datos se encontró que la variable **Y2** solo se distribuía entre dos valores, por lo tanto tiene excusa variabilidad, lo cual representaría un sobre ajuste a estos dos valores. Por ende se decidió tomar como variable independiente únicamente la velocidad del flujo **Y1**.

2.1.1 Análisis exploratorio de datos.

Una vez se construyó la base de datos asegurándose que no existan datos faltantes se realiza un análisis de las variables. En la **Figura 3** se presenta la matriz de correlación donde se resalta que las variables más correlacionadas son temperatura y la salinidad y la altura de ola y velocidad del viento. Físicamente esta correlación se sustenta en que el viento es el principal forzador de la generación de olas en la superficie del océano y por su parte la salinidad y la temperatura determinan la densidad de éste.

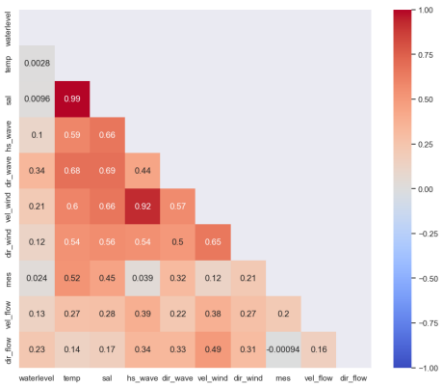


Figura 3. Matriz de correlación.

Alusivo a la distribución de las variables se grafican los diagramas de caja presentados en la **Figura 4**, donde se destaca que en general todas las variables presentan concentraciones en los cuartiles de los extremos, a excluyendo de la temperatura y la altura de ola cuyos datos se concentran en el cuartil dos. Esto podría representar inconvenientes en la aplicación de modelos, principalmente los paramétricos.

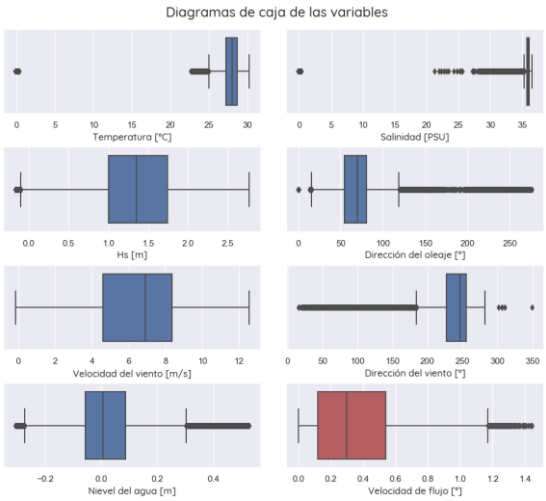


Figura 4. Diagramas de caja.

2.1.2 Selección de variables

Para disminuir la complejidad del modelo se realiza un análisis de selección de variables, logrando reducir las dimensiones del modelo. De los modelos aplicados se decidió aplicar los resultados entregados por el método "Recursive Feature Elimination", el cual utiliza un modelo de "Machine Learning" para seleccionar las variables, eliminando las de menor importancia en un proceso iterativo, finalmente las variables independientes de mayor importancia fueron las siguientes:

```
from sklearn.feature_selection import RFE
from sklearn.linear_model import LinearRegression
rfe=RFE(estimator=LinearRegression(),n_features_to_select = 4, step = 1)
fit=rfe.fit(X,Y1)
print(fit.n_features_)
print(X.columns[fit.support_])
print(fit.ranking_)
```

✓ 0.2s

```
4
Index(['waterlevel', 'temp', 'hs_wave', 'vel_wind'], dtype='object')
```

Es decir, nivel del agua, temperatura, altura de ola y velocidad del viento. La selección de estas variables se hizo bajo el criterio de garantizar que la física del fenómeno a representar no se afectara, debido a que estas variables son forzadoras de corrientes.

3 Aplicación de modelos

Para la aplicación de los modelos se llevo a cabo un esquema en el cual a partir de seleccionar las variables que más representaran el problema, se llevara a cabo una evaluación general de cada modelo a partir de aplicar **K-Fold** y **Cross Validation**, la aplicación de esta combinación de algoritmos entrega una métrica del R2 que se obtiene con el modelo. Posteriormente, si durante la etapa de evaluación se observaba una métrica aceptable se procede a realizar un análisis de sensibilidad de los hiperparámetros ejecutando el algoritmo **RandomizedSearchCV**. Finalmente, si se obtenía una buena métrica se aplica el modelo, reportando la métrica otorgada durante la etapa de validación (**Figura 5**).

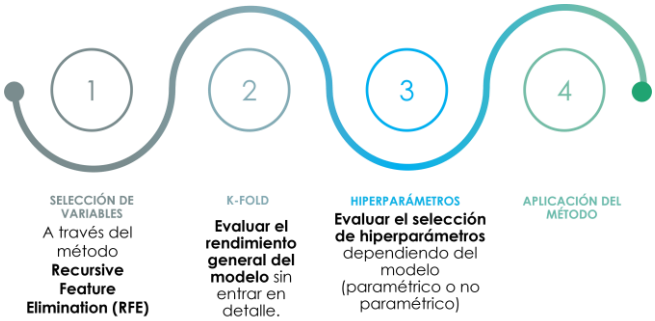


Figura 5. Esquema de aplicación de los modelos.

3.1 Regresión lineal multivariada

Las métricas otorgadas para regresión lineal (<0.2) representaban un problema caracterizado por alto bias, lo cual indicaría que este modelo no sería capas de representar la velocidad del flujo, sin embargo, se eligió este modelo como base para observar la evolución que tendría si incluía funciones de regularización o variables categóricas.

En la **Tabla 2**, se puede observar que el R2 obtenido durante la etapa de entrenamiento fue de 0.164, por su parte la métrica durante la validación fue de 0.16. Además, se puede observar que todas las variables consideradas son estadísticamente significativas. Cabe resaltar que se incluyó el intercepto porque en ausencia de los forzadores de corrientes, por mínimo que sea existen pequeñas velocidades de flujo rezagadas por lo ocurrido en tiempos predecesores.

Tabla 2. Resultados para regresión lineal OLS

OLS Regression Results

Dep. Variable:	vel_flow	R-squared:	0.164
Model:	OLS	Adj. R-squared:	0.164
Method:	Least Squares	F-statistic:	8475.
Date:	Tue, 06 Dec 2022	Prob (F-statistic):	0.00
Time:	15:30:12	Log-Likelihood:	637.40
No. Observations:	173085	AIC:	-1265.
Df Residuals:	173080	BIC:	-1214.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0824	0.002	38.779	0.000	0.078	0.087
waterlevel	0.1837	0.005	40.305	0.000	0.175	0.193
temp	0.0022	9.91e-05	21.707	0.000	0.002	0.002
hs_wave	0.1346	0.003	49.670	0.000	0.129	0.140
vel_wind	0.0051	0.001	8.801	0.000	0.004	0.006

Omnibus:	13838.957	Durbin-Watson:	2.005
Prob(Omnibus):	0.000	Jarque-Bera (JB):	17485.093
Skew:	0.740	Prob(JB):	0.00
Kurtosis:	3.485	Cond. No.	222.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

En la **Figura 6** se muestran los diagramas de dispersión entre el valor real y la variable predicha por el modelo durante la etapa de testeo, para hacer estos diagramas se tomaron las celdas que más se repetían en el tiempo (meses) durante el periodo de testo (**Figura 6a**), por su parte para representar la gráfica de la **Figura 6b** se tomaron las celdas que tuvieron velocidades mayor a 1m/s durante el testo, esto se hizo con el fin de representar como era el ajuste del modelo con los valores extremos de la velocidad. En general, se puede observar que para ambos casos la dispersión no se ajusta a una línea recta, por lo tanto, se rectifican los problemas de bias.

a) b)

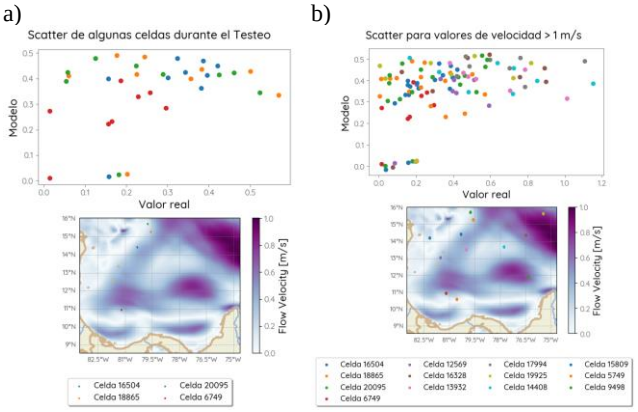


Figura 6. Diagramas de dispersión del modelo de regresión lineal.

3.2 Regresión lineal con variables categóricas

Para intentar mejorar las métricas obtenidas se decide aumentar la dimensionalidad del problema incluyendo variables categóricas, estas representarían la pertenencia de cada celda a un mes en particular. Como se indica en la

Así mismo, se puede observar que todas las variables fueron estadísticamente significativas, por consiguiente, se puede afirmar que la temporalidad asignada como variable de clasificación repercutió en el desempeño del modelo, sin embargo, algunos meses fueron estadísticamente más significativos que otros y las correlaciones entre las variables continuas se invirtieron, dado que en el modelo descrito anteriormente los coeficientes eran positivos indicando una correlación positiva.

Tabla 3 la métrica obtenida durante la etapa de entrenamiento fue **0.324** y la obtenida durante la etapa de validación fue muy similar (**0.32**).

```
rfe=RFE(estimator=LinearRegression(),n_features_to_select = 4, step = 1)
fit=rfe.fit(X,Y1)
print(fit.n_features_)
print(X.columns[fit.support_])
print(fit.ranking_)
```

```
4
Index(['waterlevel', 'temp', 'hs_wave', 'vel_wind'], dtype='object')
```

Así mismo, se puede observar que todas las variables fueron estadísticamente significativas, por consiguiente, se puede afirmar que la temporalidad asignada como variable de clasificación repercutió en el desempeño del modelo, sin embargo, algunos meses fueron estadísticamente más significativos que otros y las correlaciones entre las variables continuas se invirtieron, dado que en el modelo descrito anteriormente los coeficientes eran positivos indicando una correlación positiva.

Tabla 3. Resultados aplicando regresión lineal con variables categóricas.

OLS Regression Results

Dep. Variable:	vel_flow	R-squared:	0.325
Model:	OLS	Adj. R-squared:	0.325
Method:	Least Squares	F-statistic:	5559.
Date:	Tue, 06 Dec 2022	Prob (F-statistic):	0.00
Time:	16:21:25	Log-Likelihood:	19191.
No. Observations:	173085	AIC:	-3.835e+04
Df Residuals:	173069	BIC:	-3.819e+04
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
waterlevel	0.2695	0.004	63.718	0.000	0.261	0.278
temp	-0.1125	0.002	-60.829	0.000	-0.116	-0.109
hs_wave	0.4674	0.003	141.349	0.000	0.461	0.474
vel_wind	-0.0361	0.001	-60.680	0.000	-0.037	-0.035
mes__1	0.0905	0.002	47.339	0.000	0.087	0.094
mes__2	2.7566	0.052	52.817	0.000	2.654	2.859
mes__3	2.8541	0.052	55.072	0.000	2.753	2.956
mes__4	2.9366	0.052	56.320	0.000	2.834	3.039
mes__5	3.0290	0.053	57.031	0.000	2.925	3.133
mes__6	3.0972	0.054	57.611	0.000	2.992	3.203
mes__7	3.0584	0.054	56.688	0.000	2.953	3.164
mes__8	3.3107	0.054	60.904	0.000	3.204	3.417
mes__9	3.3975	0.055	62.122	0.000	3.290	3.505
mes__10	3.3879	0.055	61.507	0.000	3.280	3.496
mes__11	3.2966	0.055	60.354	0.000	3.190	3.404
mes__12	2.9788	0.054	55.485	0.000	2.874	3.084

Omnibus:	6149.329	Durbin-Watson:	2.004
Prob(Omnibus):	0.000	Jarque-Bera (JB):	7122.612
Skew:	0.441	Prob(JB):	0.00
Kurtosis:	3.459	Cond. No.	9.51e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.51e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Con relación a los diagramas de dispersión se observa en la **Figura 7** que tratan de ajustarse a una línea recta con mayor precisión que en la **Figura 6**, sin embargo se siguen presentando problemas de alto bias, pues los valores altos de velocidad el modelo los sobrestima.

a) b)

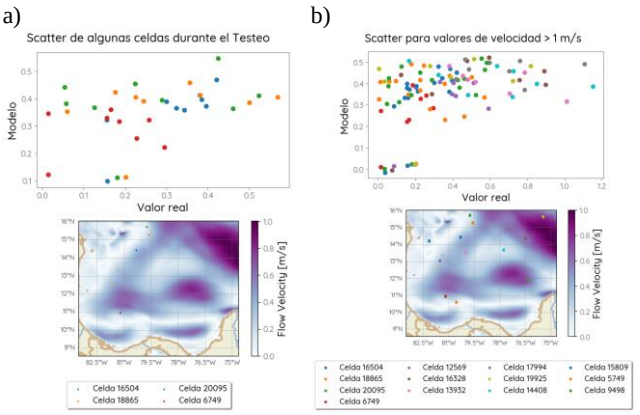


Figura 7. Diagrama de dispersión de regresión lineal incluyendo variables categóricas.

Para entender el desempeño del modelo, se observa el comportamiento de los residuales y su distribución; tomando como base las consideraciones bajo las cuales se rigen los modelos de regresión lineal:

1.	Tener una varianza constante de los residuales
2.	Residuales aproximadamente normalmente distribuidos
3.	Ser independientes el uno del otro.

La **Figura 8** indica las dispersiones de los residuales frente cada una de las variables independientes, se puede observar que la varianza de los residuales no es constante. Además, en la **Figura 9**

se representa la distribución de los residuales destacando en el diagrama Q-Q plot que en los cuartiles de los extremos se alejan del ajuste lineal, lo cual puede representan problemas en la aplicación del método ya que no estarían normalmente distribuidos.

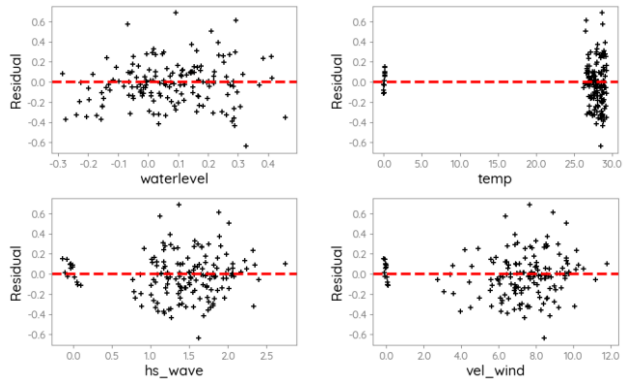


Figura 8. Diagramas de dispersión de los residuales en función de las variables independientes.

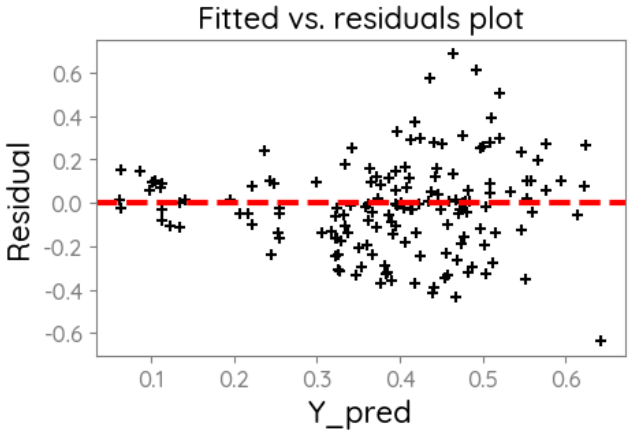


Figura 9. Diagrama de dispersión de los residuales en función de la variable predicha.

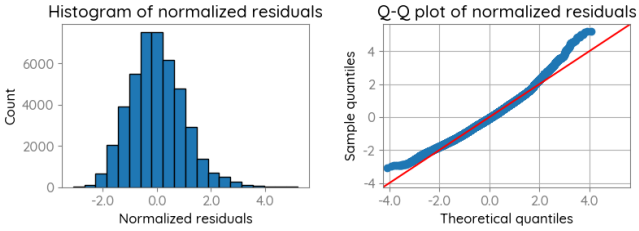
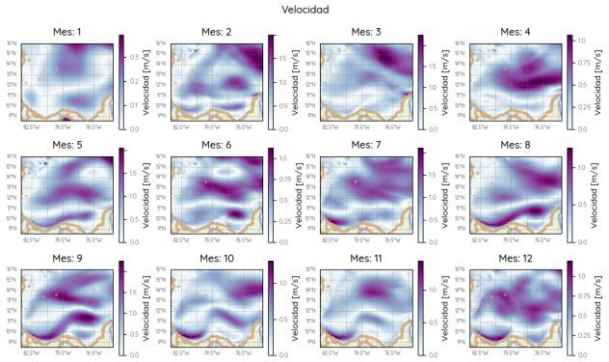


Figura 10. Distribución de los residuales.

Para representar espacial y temporalmente los residuales y compararlos con los valores de los flujos reales se presenta la Figura **11**, de esta figura se puede observar que los residuales más altos se presentan en las celdas con valores más altos de velocidad, es decir que el modelo no logra representar adecuadamente los valores más grandes de velocidad.

a)



b)

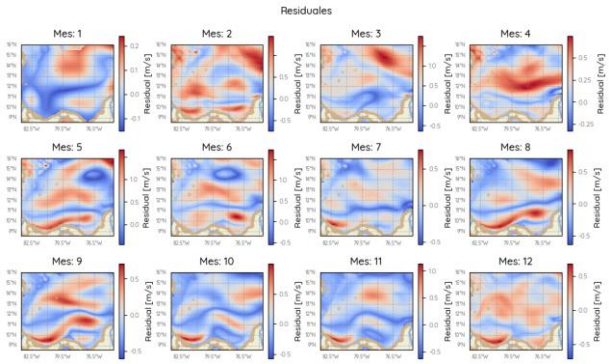


Figura 11.Comparación de los valores de velocidad de cada celda con los residuales de cada celda, a) velocidad, b) residuales.

3.3 KNN

Para evaluar el desempeño de este modelo supervisado basado en el vecino más cercano, primero se hizo una validación que diera una evaluación general del modelo con los datos, obteniendo métricas bastante altas (R2 > 0.9):

```
kfold = ShuffleSplit(n_splits=5)
model = KNeighborsRegressor()
results = cross_val_score(model, X_train.iloc[:,1:], y_train1, cv=kfold)
print(results.mean())
print(results.std())
```

✓ 1.5s

```
0.9129901863252652
0.0032970304178354636
```

Posteriormente se realizó un análisis sobre el principal hiperparámetro; número de vecinos cercanos, para ello se usó **RandomizedSearchCV**, con el fin de definir el valor optimo, obteniendo que:

```
# Buscada de los vecinos
from sklearn.model_selection import RandomizedSearchCV
from scipy.stats import uniform

k_neighbors = np.arange(2,115,10)
param_grid = {'n_neighbors': k_neighbors}
rsearch = RandomizedSearchCV(estimator=KNeighborsRegressor(), param_distributions=param_grid, n_iter=100, rando
rsearch.fit(X_train.iloc[:,1:], y_train1)
print('Mejor R2',rsearch.best_score_)
print('Mejor estimador', rsearch.best_estimator_.n_neighbors)
```

✓ 1m 0.6s Python

```
Mejor R2 0.9126437799970694
Mejor estimador 2
```

Además, se realizaron las curvas de validación mensuales mostradas en la **Figura 12**, tratando de evaluar cómo era la métrica con respecto al número de vecinos cercanos para cada mes, notando que en todos los meses se presentan problemas de sobreajuste o alta varianza y a su vez que el número de vecinos cercanos es menor a 50.

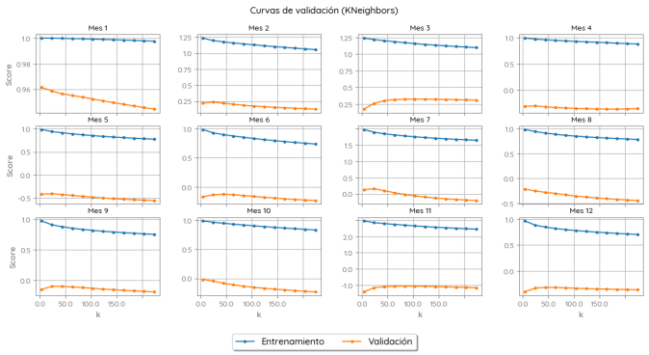


Figura 12. Curva de validación mensual.

Aplicando los resultados de **RandomizedSearchCV** se ejecuto el modelo con 2 vecinos cercanos, lo cual dio como resultado un valor de R2 de **0.97** durante el entrenamiento y de **0.92** durante la validación. En la **Figura 13** se puede observar que la dispersión de las celdas que más veces se encuentran en las muestras de testo y son mayores a 1 m/s, se ajustan a una línea, indicando que el modelo logra predecir los valores reales de la velocidad de flujo.

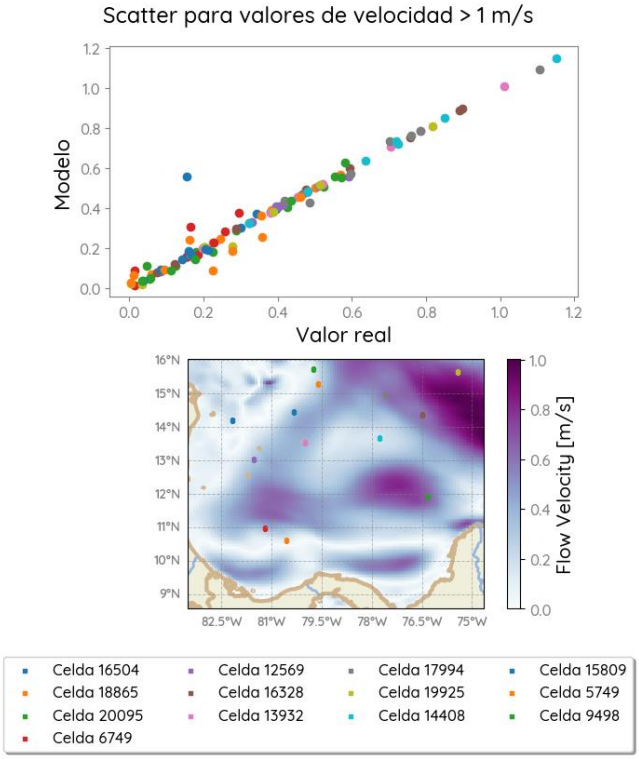


Figura 13. Diagrama de dispersión del modelo KNN

3.4 Suport Vector Machine

Para aplicar el modelo en cuestión al problema de tipo regresión se presentaron limitaciones con el número de muestras. Dado que en el presente estudio se cuenta con más de 10000 muestras, el modelo limita el uso de sus kernel a una cantidad de datos, para el presente análisis recomienda el uso de los kernel: 'Linear' o 'SDGRegressor'. En este estudio se usó el kernel 'Linear', al cual se le hizo optimización de hiperparámetros mediante **RandomizedSearchCV**, para el valor de C, el cual se ajusta para reducir o aumentar problemas de tanto de bias como de varianza, el valor optimó obtenido fue de 1, con una estimación de la métrica preliminar de 0.13:

```
from sklearn.svm import LinearSVR
c_params = np.arange(1,100,10)
param_grid = {'C': c_params}
rsearch = RandomizedSearchCV(estimator=LinearSVR(), param_distributions=param_grid, n_iter=100, random_state=1)
rsearch.fit(X_train_sd.iloc[:,1:], y_train1)
print('Mejor R2',rsearch.best_score_)
print('Mejor estimador', rsearch.best_estimator_.C)
```

9m 45.2s Python

```
Mejor R2 0.12871066863776942
Mejor estimador 1
```

Es importante resaltar que para ejecutar el modelo fue necesario estandarizar todas las variables independientes, dado que en caso contrario el modelo no lograba representar ni el valor medio.

Durante la búsqueda del parámetro C, se obtuve un valor de R2 muy deficiente, aplicando el modelo este valor fue muy similar (0.13) tanto para los datos de entramiento como testeo. Igualmente esto se logra ver en el diagrama representado en la siguiente figura.

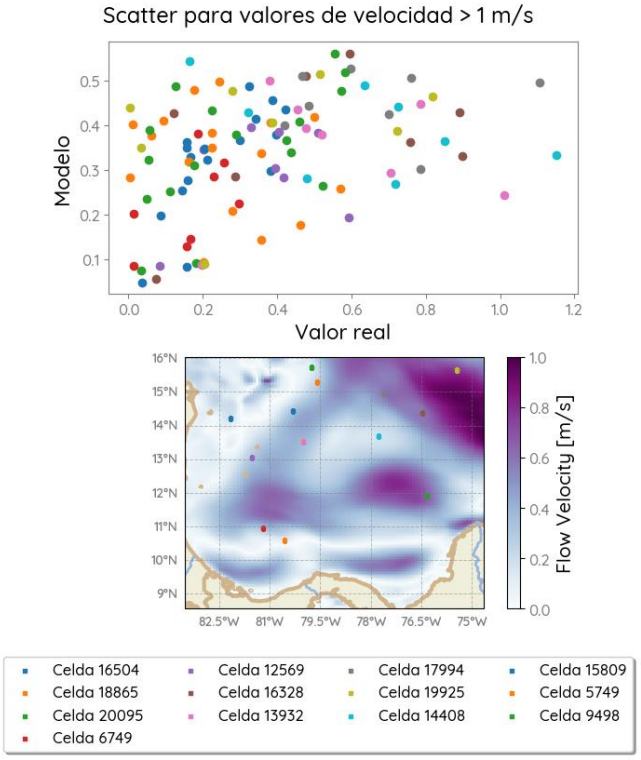


Figura 14. Diagrama de dispersión del modelo SVR Linerar

Si bien con el modelo Linear se tienen problemas de bias, al aplicar el modelo 'SDGRegressor' basado en la función de perdida gradiente descendente implica más proceso y para la cantidad de información se vuelve no optimo, sin embargo, es tarea de futuros trabajos evaluar el mismo ejercicio en una zona de menor dimensión.

4 Resultados

En la **Figura 15** se presentan los patrones de velocidad de flujo reales, a su vez, en la **Figura 16**, se muestran los patrones encontrados a partir de usar el **modelo de regresión lineal**, así mismo en la **Figura 17** se presentan los patrones encontrados con el **modelo KNN** y en la **Figura 18** se indican los patrones determinados con el modelo de **Suport Vector Regressor (Linear)**.

En general se puede identificar que el modelo KNN, representa muy bien los patrones indicados en la **Figura 15**, sin embargo, se logran identificar unos pequeños patrones con velocidades

superiores a las reales, especialmente en los meses 6, 9 y 12, las magnitudes de la velocidad son levemente mayores. Aunque la métrica de la validación fue acertada posiblemente se presenten problemas de varianza, sin embargo, es un buen modelo para representar problemas que incluyan espacialidad y temporalidad.

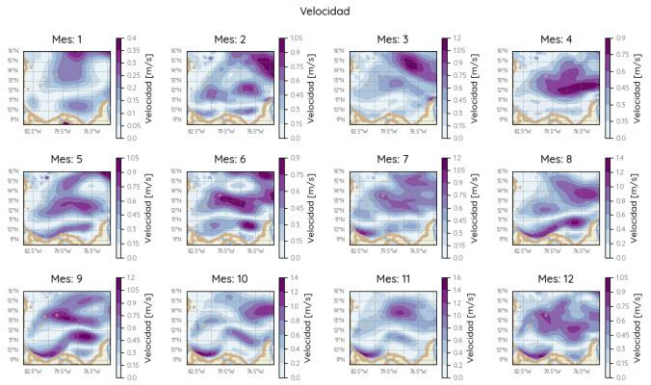


Figura 15. Patrones de velocidad de flujo en un año.

A diferencia de KNN, y consecuente con la **Figura 11**, los patrones determinados usando regresión lineal, logran representan el régimen medio de la velocidad y se quedan limitados en la representación de los valores extremos, lo cual dependiendo de aplicabilidad de los resultados del modelo podría ser útil o no, por ejemplo ante el paso de eventos extremos naturales el transporte de momentum tiende a incrementar por lo tanto las corrientes superficiales, si estos eventos son instantáneos las magnitudes de la velocidad aumentaran, por lo cual esto se podría ver reflejado en los datos como un "outlier" que ni los modelos de regresión lineal y SVR (**Figura 18**) implementados serían capaces de reproducir. Ahora si se desarrolla un modelo para predecir órdenes de magnitud medias en una escala de tiempo grande (anual, multianual) estos modelos podrían ser útiles.

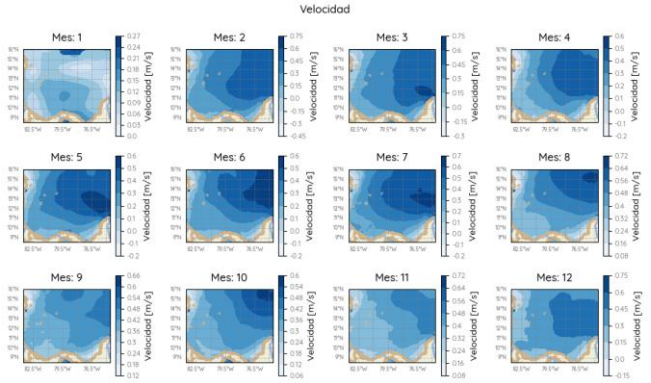


Figura 16. Patrones de velocidad encontrados con el modelo de Regresión lineal.

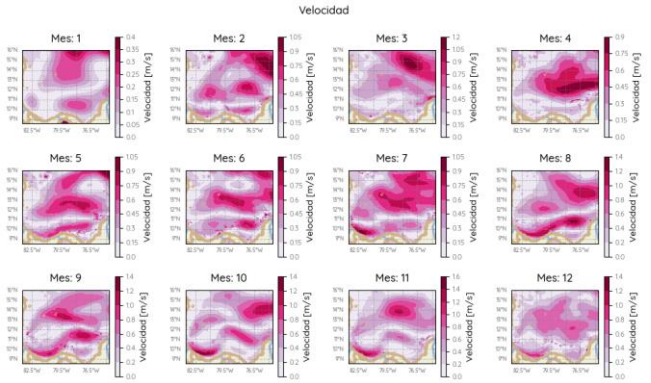


Figura 17. Patrones de velocidad determinados a partir del modelo KNN

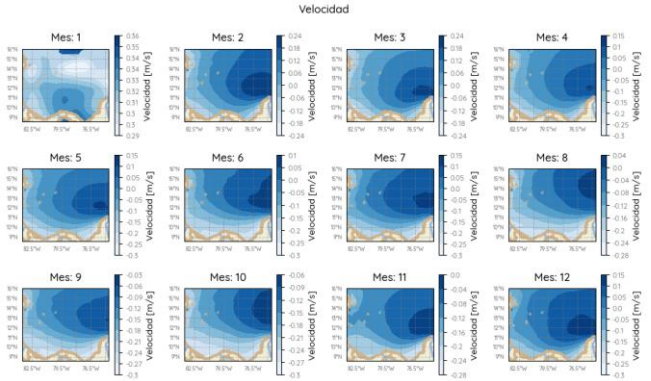


Figura 18. Patrones de velocidad encontrados con el modelo de SVR

5 Discusión

Los modelos basados en datos requieren un nivel de entendimiento de los procesos matemáticos o estadísticos que ellos consideran, desde la relación entre variables hasta la introducción de hiperparámetros. Entre los modelos evaluados se encontró que los modelos supervisados representan, con mejor métrica (R^2), la variabilidad de los patrones de flujo a lo largo del ciclo anual, principalmente el modelo KNN, fue el modelo que más se logró ajustar la representación de la variable dependiente, sin embargo, problemas de alta varianza pueden generar incertidumbre en la validación del modelo o en la aplicación en otra zona, por lo tanto, para solucionar este problema se hizo un análisis de sensibilidad al parámetro número de vecinos, y a la ponderación de los pesos de la cantidad de vecinos elegidos, esta sensibilidad redujo el valor de R^2 de 0.90 a 0.80 lo cual puede controlar un poco los problemas de varianza, por tanto se destaca el modelo como acertado representando los patrones de flujo.

Entre los modelos paramétricos, se usó regresión lineal multivariable, encontrando que la mejor métrica R^2 alcanzada fue 0.31, indicando que el modelo tiene problemas de presión (bias) dado que no se logra representar acertadamente los patrones de la velocidad, y de acuerdo con las consideraciones que se asumen para aplicar regresión lineal, se observó que se violan principalmente por no tener una varianza constante de los residuales y por el desajuste de los residuales en los extremos. Para futuros estudios se puede considerar el uso de funciones de regularización como (Laso y Ridge) incluyendo la temporalidad asociada a pertenencia a un mes del ciclo anual teniendo en cuenta la sensibilidad sobre el parámetro alfa.

De los modelos ensamblados como Random Forest, o redes neuronales, se concluye que no fueron los más óptimos para representar el objetivo del presente articulo puesto que la cantidad de datos que representa el domino del ráster acumulan a lo largo del año muestras superiores a 10000 celdas, por lo tanto el gasto computacional tan solo realizar la sensibilidad a los parámetros e hiperparámetros del modelo tomó un tiempo superior a 60 minutos, sin embargo, resultados preliminares de las métricas a partir de validación cursada con kflod se obtuvieron desempeños de 0.90, bajo estos resultados se establece que un modelo ensamblado bien estructurado, sería aplicable en la predicción de la variabilidad anual para ráster de domino medido (41km x 41 km).

6 Referencias bibliográficas

Alberto, C., & Amaya, A. (2001). Las Corrientes Superficiales En La Cuenca De Colombia Observadas Con Boyas De Deriva. *Revista de La Academia Colombia de Ciencias Exactas y Natura*.

Escobar, C. A., Velásquez, L., & Posada, F. (2015). Marine Currents in the Gulf of Urabá, Colombian Caribbean Sea. *Journal of Coastal Research*, *31*(6), 1363–1374. https://doi.org/10.2112/JCOASTRES-D-14-00186.1

Jirakittayakorn, A., Kormongkolkul, T., Vateekul, P., Jitkajornwanich, K., & Lawawirojwong, S. (2017). Temporal kNN for short-Term ocean current prediction based on HF radar observations. *Proceedings of the 2017 14th International Joint Conference on Computer Science and Software Engineering, JCSSE 2017*. https://doi.org/10.1109/JCSSE.2017.8025921

Jouon, A., Douillet, P., Ouillon, S., & Fraunié, P. (2006). Calculations of hydrodynamic time parameters in a semi-opened coastal zone using a 3D hydrodynamic model. *Continental Shelf Research*, *26*(12–13), 1395–1415. https://doi.org/10.1016/j.csr.2005.11.014

Liu, Y., & Weisberg, R. H. (2005). Patterns of ocean current variability on the West Florida Shelf using the self-organizing map. *Journal of Geophysical Research: Oceans*, *110*(6), 1–12. https://doi.org/10.1029/2004JC002786

Posada, F., Escobar, C. A., & Vela, L. (1997). ´, *Colombian Marine Currents in the Gulf of Uraba Caribbean Sea*. https://doi.org/10.2112/JCOASTRES-D-14-00186.1