

Reporte Técnico

Base de datos de las observaciones de las Boyas Oceanográficas y de Meteorología Marina (BOMM)



Grupo de Oleaje

Departamento de Oceanografía Física

Centro de Investigación Científica y de Educación Superior

Ensenada, B.C., México

8 de enero de 2019

Revisión 1.0

Índice

1. Introducción	3
1.1. Niveles de procesamiento	3
1.2. Formato NetCDF	3
2. Descripción de la base de datos	4
2.1. Nivel 0	4
2.2. Nivel 1	5
2.3. Nivel 2	9
2.4. Nivel 3	9

1. Introducción

En este reporte técnico se presenta una descripción detallada del conjunto de datos observados con las Boyas Oceanográficas y de Meteorología Marina (BOMM).

1.1. Niveles de procesamiento

Para los fines de este trabajo, se definieron cuatro niveles de procesamiento, los cuales se describen a continuación:

Nivel 0 En este nivel están los datos binarios que se escriben en la memoria interna de los instrumentos y los datos que son transmitidos a través de comunicación entre los instrumentos y la computadora de la boya. Estos datos están en formato ASCII. Estos datos no tienen ningún control de calidad ni tienen incorporados los metadatos.

Nivel 1 El primer nivel de procesamiento consiste en generar la base de datos en formato netCDF4 a partir de los datos de Nivel 0. Se hace una acomodación de los datos en una secuencia creciente del tiempo de acuerdo a su tasa de muestreo. Adicionalmente, se llenan los espacios vacíos con banderas de datos no válidos (NaN). Se incorporan los metadatos e información adicional para realizar correcciones y calibraciones. **Ejemplos:** *aceleraciones y tasas de cambio de los ángulos en el marco de referencia del sensor de movimiento, velocidad del viento sin corregir por el movimiento de la boya, posición de los WaveStaff para el cálculo de los espectros direccionales.*

Nivel 2 En este nivel se presentan las variables derivadas a la misma tasa de muestreo de los datos del Nivel 1. Se aplican las correcciones y calibraciones necesarias. Se aplican un control de calidad de los datos de acuerdo a criterios físicos y estadísticos. **Ejemplos:** *elevación de la superficie libre del mar, velocidad del viento corregida por el movimiento de la boya, perfil de corrientes superficiales en el marco de referencia inercial.*

Nivel 3 En el último nivel se presentan las variables resultantes de agrupar y combinar los datos en un tiempo determinado. Específicamente, se presentan promedios de las variables provenientes de diferentes sensores. Se aplican metodologías establecidas para el cálculo de variables derivadas. **Ejemplos:** *espectros direccionales, esfuerzos de Reynolds, rapidez y dirección del viento, parámetros integrales del oleaje, etc.*

1.2. Formato NetCDF

Según el SMID de CIGOM (<http://smid.cigom.org/smid-docs/>), en la Línea 1 del proyecto se estableció que la entrega de conjuntos de datos será en formato NetCDF ya que es un formato ampliamente utilizado por la comunidad científica y además es compatible con una gran variedad de software de análisis y visualización. NetCDF (*Network Common Data Form*) es un conjunto de bibliotecas de software y

estándares de formato de datos, el cual es independiente de la arquitectura de la máquina, es de código abierto, y permite la creación, acceso y distribución de datos científicos orientados a arreglos. Una de sus mayores ventajas es que es autodescriptivo, es decir, los archivos contienen los datos y los metadatos. Otra de sus ventajas es la portabilidad y la escalabilidad, que permiten una fácil distribución y el acceso a subconjuntos más pequeños de los datos que están alojados en un servidor remoto. Específicamente la versión 4 del formato NetCDF permite la separación de los datos en grupos, lo cual es muy útil cuando se trabajan con datos de diferentes sensores y medidos a diferentes tasas de muestreo en el mismo dataset, como es el caso de los datos de las boyas oceanográficas.

2. Descripción de la base de datos

2.1. Nivel 0

Se consideran como datos crudos los datos que se escriben directamente en el disco duro de la boya. Estos datos no tiene ningún tipo de procesamiento, y se escriben tal cual como salen de cada uno de los sensores. Los datos crudos son los datos de nivel 0 y tienen la siguiente estructura:

```
data/
|-- level0/
|   |-- acelerometro/
|   |-- anemometro/
|   |-- binary/
|   |-- gps/
|   |-- marvi/
|   |-- maximet/
|   |-- msg/
|   |-- proceanus/
|   |-- rbr/
|   |-- signature/
|   |-- vector/
|   `-- wstaff/
|-- level1/
|-- level2/
`-- level3/
```

Se genera una carpeta por cada sensor. En la mayoría de los sensores (los de alta frecuencia) se escribe una carpeta por año, una por mes, una por día y una por hora; y se escribe un archivo cada 10 minutos. Por ejemplo, la ruta del archivo que corresponde a las 10:20 am del 5 de enero de 2018 del anemómetro sónico es:

`./data/level0/anemometro/2018/01/05/10/anemometro-1801051020.csv`

Por otra lado, los sensores de baja frecuencia, es decir, los que reportan promedios de los datos, como el sensor de CO₂ y el CTD, se escribe una carpeta por año y una carpeta por mes, y se escribe un archivo por día, el cual contiene los datos a la tasa de muestreo específica.

En la carpeta `binary` se almacenan los archivos binarios originales de los sensores que así lo permiten. Por ejemplo, en el caso de la BOMM1-ITS que estuvo midiendo cerca de la Isla de Todos Santos, entre noviembre de 2017 y febrero de 2018, los sensores que permitieron almacenar los datos binarios fueron el acelerómetro (Ekinox2-M), el MARVI (Módulo de Adquisición y Regulación de Voltaje Inteligente) y el velocímetro Vector. En la carpeta `msg` se almacenan los promedios de los datos que son enviados vía satélite. En las carpetas `level1`, `level2` y `level3` se almacenan los datos procesados en los diferentes niveles, los cuales se describen en las siguientes secciones.

2.2. Nivel 1

Análisis de los datos

Se realizó un análisis y procesamiento de los datos crudos para convertirlos al formato `netCDF`. Para esto, primero se hace una acomodación de los datos en una secuencia creciente de tiempo de acuerdo a su tasa de muestreo. Por ejemplo, algunos sensores presentan imprecisiones en su reloj interno, lo que implica que la tasa de muestreo a la que se programan no sea constante en el tiempo lo cual genera pequeñas variaciones, que hacen que en ocasiones se tengan más o menos datos de los que se esperan en cierto intervalo de tiempo. Por ejemplo, los alambres de capacitancia se programan para medir a una tasa de muestreo de 20 datos por segundo, pero debido a la deriva del reloj, en ocasiones se tiene 19 o 21 datos en un segundo. Para solucionar esto, se genera una arreglo del tiempo de acuerdo a la tasa de muestreo del instrumento y se aplica una interpolación lineal de los datos. Si hay más del 10 % de datos inválidos dentro de un intervalo de 10 minutos, ese intervalo se considera como inválido. Cuando son menos de 10 % los datos inválidos y estos no están de forma continua en el tiempo, los datos inválidos se reemplazan por promedios del intervalo de 10 minutos con el fin de aplicar la interpolación. Finalmente, cuando se tiene espacios vacíos, es decir, intervalos de tiempo donde la boya no registró datos, se genera un arreglo con datos no válidos (NaN) del mismo tamaño y se llenan dichos espacios vacíos. En los datos de Nivel 1 se eliminan también los datos en los que la boya no estaba instalada. Estos datos si se conservan en el Nivel 0.

Estructura de los archivos

Los datos de nivel 1 de procesamiento se escriben en archivos `netCDF4`, usando la capacidad de generar grupos de datos, característica de este formato. Se genera un archivo `netCDF4` por cada día de datos. El nombre cada archivo consiste en la fecha en el formato `'yyyymmdd'` más la extensión `'nc'`, por ejemplo:

```
data/
`-- level1/
    |-- 20171117.nc
    |-- 20171118.nc
    |-- 20171119.nc
    |--      .
    |--      .
    |--      .
    |-- 20180131.nc
    `-- 20180201.nc
```

Cada archivo agrupa los datos en grupos por cada sensor. Por ejemplo, los datos de la BOMM1-ITS se clasifican en los siguientes grupos.

```
groups: ekinox, sonic, gps, marvi, maximet, proceanus,
        rbr, signature, vector, wstaff
```

Los grupos contienen las variables y las dimensiones. En este caso solo se tiene dos tipos de dimensiones, el tiempo y el número de celdas, este último es específico del perfilador de corrientes Signature 1000 kHz. El siguiente es un ejemplo de la estructura del grupo asociado con los datos del sensor de movimiento Ekinox2-M:

```
group /ekinox:
    sampling_frequency: 100
    serial_number: 5242914
    description: MRU Subsea Ekinox2-M
    convention: X positive towards north buoy, Y eastward and Z downward
    dimensions(sizes): time(8640000)
    variables(dimensions): float64 time(time), float64 accel_x(time),
                                float64 accel_y(time), float64 accel_z(time),
                                float64 gyro_x(time), float64 gyro_y(time),
                                float64 gyro_z(time), float64 delta_vel_x(time),
                                float64 delta_vel_y(time), float64 delta_vel_z(time),
                                float64 delta_ang_x(time), float64 delta_ang_y(time),
                                float64 delta_ang_z(time), float64 temp(time)
```

Cada una de las variables tiene atributos, para los cuales se sigue la convención CF-1.7 y las recomendaciones de los manuales de usuario de cada uno de los instrumentos. Por ejemplo, los atributos de la tasa de cambio del ángulo al rededor del eje x , se presentan a continuación:

```
float64 gyro_x(time)
  _FillValue: nan
  standard_name: gyro_x
  long_name: rate of change of the angle in X direction
  units: rad/s
path = /ekinox
unlimited dimensions:
current shape = (8640000,)
filling on
```

Los metadatos de las mediciones que se incorporan en el archivo netCDF se presentan en un archivo aparte con un formato YAML (<http://yaml.org/>) ya que es el formato más amigable para este tipo de información. Por ejemplo, los atributos globales que identifican la información de la boya son:

```
title: >
  This dataset presents air-sea physical and chemical variables from a BOMM
  (Oceanographic and Marine Meteorology buoy) near the Isla Todos Santos,
  Ensenada, BC, México, from Nov 2017 to Jan 2018.
```

```
history: >
  The BOMM1-ITS was deployed on 2017/11/16 and recovered on 2018/02/02.
```

```
source: Ocean surface observations
convention: CF-1.7
institution: CICESE - CIGOM
creator_name: The waves group - CICESE
creator_url: https://www.cicese.mx/
email: oleaje@cicese.mx
acknowledgments: >
  This research has been funded by Fondo Sectorial CONACYT-SENER
  Hidrocarburos, Project 201441.
```

```
comments: |
  Sampling frequency is shown in Hz, in all cases.
```

Igualmente, en el formato YAML se presentan los metadatos asociados con los atributos de cada variable. Para esto se separan en grupos de acuerdo con cada sensor como se muestra a continuación:

```
ekinox:
  sampling_frequency: 100
```

```
seconds_per_file: 600
serial_number : 024000042
description: MRU Subsea Ekinox2-M
convention: X positive towards north buoy, Y eastward and Z downward
variables:
  accel_x:
    column: 6
    long_name: acceleration in X direction
    units: m/s^2
  accel_y:
    column: 7
    long_name: acceleration in Y direction
    units: m/s^2
  .
  .
  .
  delta_ang_z:
    column: 18
    long_name: coning output in Z direction
    units: rad/s
  temp:
    column: 15
    long_name: sensor temperature
    units: degrees_celsius
```

En este caso, los metadatos son los mismos que se escribieron en el archivo NetCDF, excepto por algunos como `column` que es el número de la columna en la que se presentan los datos en el archivo ASCII del nivel 0.

Ejemplos de acceso a los datos

Los archivos netCDF4 son de fácil acceso en la mayoría de los lenguajes de programación, por ejemplo, en Python se usa la paquetería `netcdf4-python` (<http://unidata.github.io/netcdf4-python/>). A continuación se presenta un ejemplo simple de cómo leer los datos de nivel 1 en python3.6.5.

```
import numpy as np
import netCDF4 as nc
#
filename = "../data/20171117.nc"
data = nc.Dataset(filename) #<-- carga el dataset completo
```



```
ekx = data["ekinox"]          #<-- lee el grupo del ekinox
met = data["maximet"]         #<-- lee el grupo de la maximet
#
a, b = 0, 180000              #<-- indices para el tiempo 30 mins
ax = ekx["accel_x"][a:b]      #<-- extrae la aceleracion en x
```

2.3. Nivel 2

TODO

2.4. Nivel 3

TODO
