# A.I. Cup: Sound of Climate Change

C. Millet, S. Oger

CEA DAM DIF, France
April 07, 2022

## 1. Introduction

Low frequency sounds, also known as infrasound, are generated worldwide by a variety of natural and anthropogenic sources including ocean waves, volcanoes, mountains, chemical explosions, mining blasts and meteorites. For a frequency less than a few hertz, these sounds travel through atmospheric waveguides that extend from the Earth's surface up to the upper atmosphere, where complex physical processes occur. The combination of the low-frequency content and atmospheric waveguides facilitates long-range infrasound propagation and allows its detection by ground-based arrays. However, the dependence of waveform morphology on atmospheric structures results in signals that show a large variety of shapes, which makes infrasound signal classification a difficult problem.

Infrasound signals are now routinely detected on sparse ground-based infrasound networks, such as the International Monitoring System (IMS), which was developed by the Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO). While the infrasound signal classification remains of particular concern for national security and hazard mitigation, the effect of atmospheric variability on infrasound propagation has received a growing level of research interest over the last decades. Many data centers, including the International Data Center (IDC) which is part of the CTBTO, are building acoustic event catalogs, resulting in a large amount of data. With the advent of regional infrasound arrays, and when combined with large scale monitoring environments, such catalogs present a great potential for climate-related scientific applications.

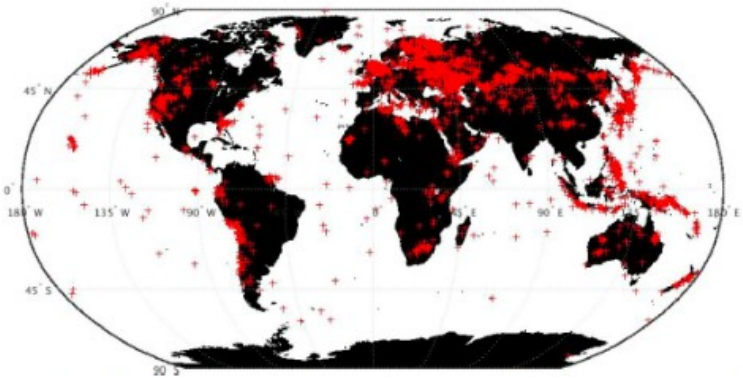

**Figure 1** : Localization of events from the database Reviewed Event Bulletin (REB) produced by IDC

Most studies based on machine learning techniques are applied to datasets that are limited in geographic area, waveguide complexity and/or source diversity. It is unclear how these methods transfer to datasets that are representative of realistic time-evolving atmospheres encountered in real-time monitoring operations. Another limitation of the existing algorithms is that they fail to capture the stochastic small-scale atmospheric processes. Such processes are essential to better represent the inadequacy in climate models and to improve the quantification of forecast uncertainty. Operational weather centers now routinely use numerical schemes, also known as stochastic parameterizations, to represent these processes in short-term, medium-range, and seasonal forecasts. Developed initially for numerical weather prediction, the inclusion of stochastic parameterizations is also extremely promising for reducing long-standing climate biases and is relevant for determining the climate response to external forcing.

**Multi-class classification of infrasound-like signals.** The objective of this challenge is to classify infrasound-like waveforms, according to the small-scale random atmospheric structures infrasound has encountered along the source-receiver path.

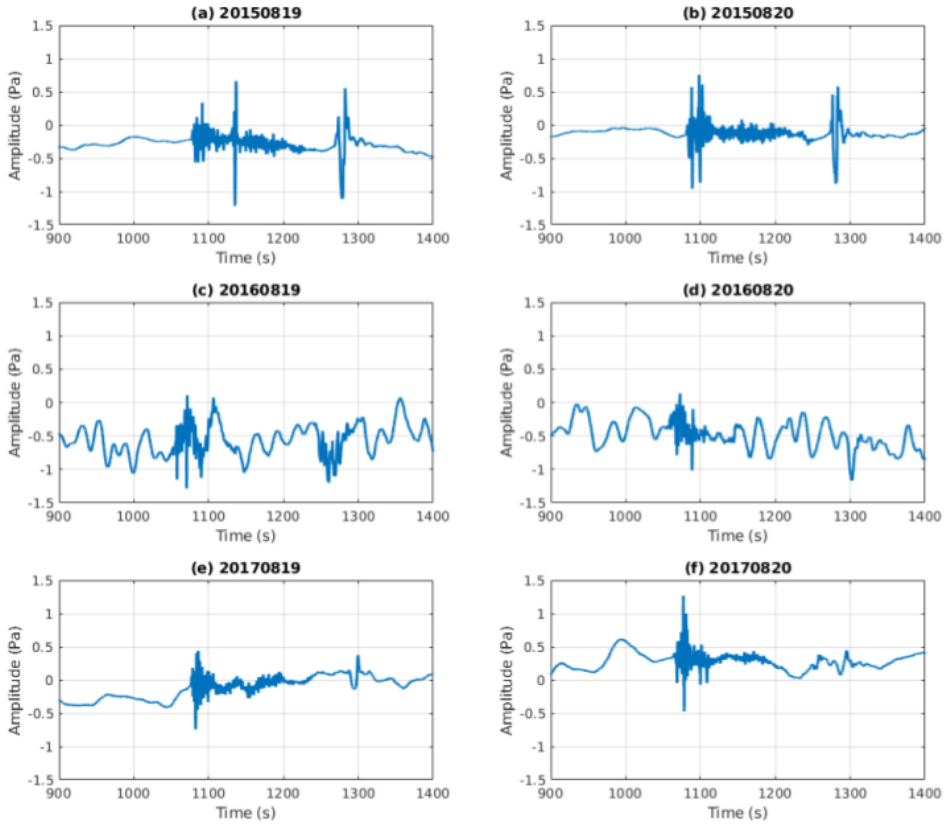


**Figure 2.** Recorded signals at a given IMS station for several dates (in YYYYMMDD format) for a recurring event located 320 km away from the IMS station. Each signal gives the overpressure (in Pa) as a function of time. Since the source is fixed, the waveform variability can be directly related to that of the atmospheric structures between the source and the IMS station.

## 2. Data collection

### 2.1 Description

Infrasound data record atmospheric events of interest, such as atmospheric nuclear explosions, but also a lasting ambient noise dependant on the local environment (e.g., the ocean). To make the most realistic infrasound-like data, we added these two components, that we will refer below as signal and noise.

The signals were produced using the numerical plaform FLOWS (Fast Low-Order Wave Simulation) developed and extensively used at CEA, for operational activities, and in collatboration with the IDC. The signal computation is based on the wave equation, starting from the classical normal mode technique and applying CEA expertise-based techniques for accounting atmospheric absorption and ground reflection. Such long-range infrasound propagation problems are characterized by a large number of length scales and a large number of propagating modes. In the atmosphere, these modes are confined within waveguides causing the sound to propagate through multiple paths to the receiver. FLOWS computes these modes and the effect of random atmospheric fluctuations on these modes. It is worthwhile to notice that the overall performance of this plateform has been demonstrated intensively using ground truth events of specific concern for the verification regime of CTBTO and CEA.

Within the frame of the IACup, 8 sets of stochastic parameterization were used to describe the random atmospheric component, which is superimposed onto the background state. The signals were then computed using a low-pass filter with a cutoff frequency of 0.2 Hz.

The noise result from infrasound data recorded at the IMS station I48TN in Sicily over the period 03/12/2015-02/11/2016. It was resampled to fit the signal (2.5s), and cut into bins of 94 time steps. A bin was then randomly taken and added to each signal.

### 2.2 Metadata

As a result, the dataset consist of 4000 overpressure signals (in Pa) as a function of time (s), over 94 time steps, and labelled as 'noisy_SignalDATA'. The target is the perturbation class, as a binary matrix for the 8 classes, and labelled as 'ClassDATA'. The dataset was split such as 2400 signals were used for the trainings and 1600 for the tests. The table 1 summarize the labels and dimensions of the data.

| | Training Set | Testing Set |
|---|---|---|
| Predictor | 'noisy_SignalDATA' [2400 x 94] | 'noisy_SignalDATA' [1600 x 94] |
| Target | 'ClassDATA' [2400 x 8] | - |

**Table 1** : Description of the data

## 3. Performance metrics

The classification is evaluated with (i) the **precision** and (ii) the **confusion matrix**.

The **precision** is the number of true cases devided by the number of predicted cases.

The **confusion matrix** is a table which describes the model results of the test set, counting the true classes agains the predicted classes (see Figure 2).

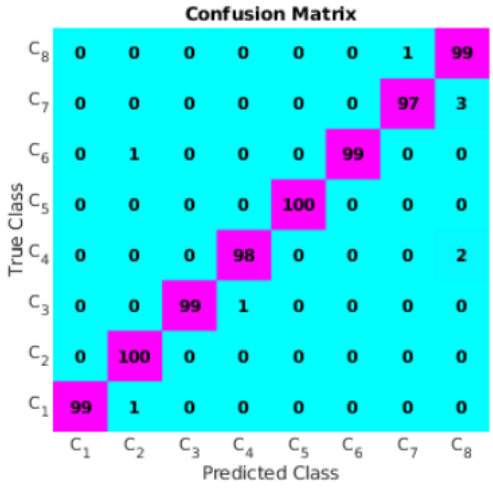


**Figure 3** : Example of a confusion matrix, counting the true classes Ci agains the predicted classes Ci.