Alma Mater Studiorum - Università di Bologna

DOTTORATO DI RICERCA

MODELLISTICA FISICA PER LA PROTEZIONE DELL'AMBIENTE

Ciclo XXII

Settore scientifico-disciplinare di afferenza: FIS/06

Probabilistic Tomography of Atmospheric parameters from GNSS data

Presentata da:
Dott. Alberto Ortolani

Relatore:
Prof. Rolando Rizzi

Coordinatore:
Prof. Rolando Rizzi

Esame finale anno 2011

Alla mia piccola Irene,
al tempo che non abbiamo avuto
e a quello che passeremo insieme,
altrove.

To my little Irene,
to the time that we haven't had
and to the one we will spend together,
elsewhere.

Contents

List of Figures

List of Tables

List of Acronyms

APL: Applied Physics Laboratory

A-S: Anti-Spoof

AUTONAV: AUTOnomous NAVigation

BPSK: Binary Phase Shift Keying

CDMA: Code Division Multiple Access

COSMEMOS: COoperative Satellite navigation for MEteo-marine MOdelling and Services

CS: Control Segment

CS: Commercial Service

DCB: Differential Code Bias

ECEF: Earth Centred Earth Fixed

ECI: Earth Centred Inertial

ECMWF: European Centre for Medium Range Weather Forecasts

EO: Earth Observation

EOF: Empirical Orthogonal Function

FDMA: Frequency Division Multiple Access

FP7: 7^{th} Framework Programme

GA: Ground Antenna

GFS: Global Forecast System

GLONASS: GLObal NAvigation Satellite System

GNSS: Global Navigation Satellite System

GPS : Global Positioning System

IEEE: Institute for Electrical and Electronic Engineers

IGS: International GPS Service for Geodynamics

IR: InfraRed

IRNSS: Indian Regional Navigational Satellite System

IWV: Integrated Water Vapor

LAM: Limited Area Model

MCS: Master Control Station

MEO: Medium Earth Orbit

MS: Monitor Station

MW: MicroWaves

NAVSTAR: NAVigation System with Time And Ranging

NAVWAR: NAVigation WARfare

NGS: National Geodetic Survey

NCST: Naval Center for Space Technology

NDS: Nuclear detonation Detection System

NNSS: Navy Navigation Satellite System

NRL: Naval Research Laboratory

NWP: Numerical Weather Prediction

OCS: Operational Control Segment

OCXO: Oven Controlled Crystal Oscillator

OS: Open Service

PPS: Precise Positioning Service

PRN: Pseudo Random Noise

PRS: Public Regulated Service

PW: Precipitable Water

RC: Ranging Codes

RH: Relative Humidity

RHCP: Right-Hand Circularly Polarised

SA: Selective Availability

SAR: Search And Resque

SBAS: Satellite Based Augmentation System

SOL: Safety Of Life

SSM: Spread Spectrum Modulation

SPS: Standard Position Service

TCXO: Temperature Compensated Crystal Oscillator

TEC: Total Electron Content

TIMATION: TIMe navigATION

UV: UltraViolet

WMO: World Meteorological Organisation

WRF: Weather Research and Forecasting

ZHD: Zenith Hydrostatic Delay

ZWD: Zenith Wet Delay

Introduction

Problem statement

In the next few years, as soon as Galileo will be deployed, more than 50 GNSS (Global Navigation Satellite System) and SBAS (Satellite Based Augmentation System) satellites will be operational, emitting precise microwave (MW) L-band spread spectrum signals, and will remain in operation for several decades. As it will be comprehensively described in this work, these signals, can be used for remote sensing of the Earth, distinctively for atmospheric monitoring.

A number of works are available for measuring vertically integrated water-vapour, sometime referred as precipitable water, from fixed GNSS receiving stations (see for instance Bar-Sever 1997). They generally propose different solutions to the problem, but the basic idea they are based on is essentially the following: computing the signal delay due to the geometric distance between satellites and the receiver station, whose position is know with great accuracy (being fixed); then measuring the actual phase-delay of the GNSS signal (measured by the receiver); finally subtracting the delay for the geometric distance, in order to obtain the delay due to atmospheric effects. The problem of estimating water vapour is however much more complicated, because such delay contains also various errors due to uncertainties in satellite and receiver clocks, relativistic effects, receiver processing steps etc. (J. J. Spilker, Jr, 1980; B.W. Parkinson, J.J. Spilker, 1996), that often are of the same order of magnitude of the atmospheric effects we want to measure, and very difficult to model with the necessary accuracy. In addition the atmospheric delay consist of a first major term from ionospheric effects, a second term from the dry component of the

troposphere and a third term from the water vapour (i.e. the wet part), that have to be decoupled. The Ionosphere is dispersive at the GNSS frequencies, thus a receiver processing the different GNSS frequencies allows to decouple the ionospheric components, by forming the ionospheric-free combination. In order to decouple the dry tropospheric effects, some ancillary information are needed. Normally a number of strong approximations are done in order to solve the problem. Typically, mapping functions are used to map slant tropospheric delay into zenith delay, and inferences on the temperature and pressure vertical profiles are made from local measurements in order to close the system equations in the retrieval process (Bevis et al. 1992; Rocken et al. 1995).

The aim of the present work is to demonstrate that, under certain conditions and through a proper bayesian approach, profiles (instead of integrated values) of water-vapour partial pressure, of temperature and pressure are achievable from GNSS data. We will show that a 3D retrieval (into a limited volume) of such atmospheric parameters could be achievable through a simultaneous processing of GNSS signals from different receivers. These are key parameters in meteorology, because they are the basic tropospheric state parameters, driving the local dynamics, the precipitation and the exchange of latent and sensible heat between atmosphere and earth surface. A probabilistic algorithm is thus proposed, that goes beyond the "mapping function approach" for measuring zenith integrated atmospheric parameters, i.e. averaged parameters in a strict plane parallel approximation for atmosphere. The issue we want to address is analogous to the "classical" tomography problem, but from measurements of GNSS delays sampled from station points and for slant directions that are inhomogeneous in space and variable in time: as a consequence the properties of discrete Fourier transforms are not (at least straightforwardly) exploitable, as normally it is done to solve the tomographic problem.

It will be analysed on which extent the accuracy of retrievals depends on precisions and number of available signals (i.e. of satellites simultaneously in view and receivers in the target area) and on ancillary information eventually

available, such as measurements of surface pressure, temperature and relative humidity. As a main achievement however it will be assessed where information is effectively gained due to GNSS data, with respect to a priori information or in situ measurements. Not surprisingly the assessment of water vapour is the most positively constrained by GNSS data, but, under some conditions and measurement configurations, also temperature and even pressure can gain information, especially for non surface values.

The core part of such analysis is performed in a "controlled environment", i.e. it is done by means of synthetic data that however are generated on the basis of a previous accurate analysis of main characteristics of real data, errors included.

The proposed approach has the advantage to retrieve parameters each time with their probability density functions (i.e. uncertainties), as mandatory, for example, for a correct assimilation into models. In the perspective of an eventual operational implementation the method has a second advantage, that is the robustness, because number, distribution and precision of measurements affect result accuracy but not the feasibility of the retrieval process, that, at worst, it gives back a priori distributions. Conversely it can be very machine-time consuming, if some expedients are not adopted, especially in case of large number of measurements to be processed.

Structure of the thesis

The structure of the thesis is the following:

- Chapter 1 recalls the basics characteristics of the atmosphere from a stationary point of view, with more details on troposphere and ionosphere, that are relevant for the present work as the basis of the a priori knowledge of the parameters we want to retrieve.

- Chapter 2 analyses the properties of an electromagnetic L-band wave crossing a gas medium (neutral and slightly ionised), up to evaluating approxi-

mations enclosed in the geometrical optics approach we adopt as solution to the problem.

- Chapter 3 describes the principles of GNSS positioning signals, and, referring mainly to GPS, the main error sources in the computation of the tropospheric component of the signal delay from fixed receiving stations.

- Chapter 4 explains the algorithm designed for retrieving the atmospheric parameters from the GNSS tropospheric delays, addressing the problem from 0D to 3D, and it describes also the dataset generated for testing performances and information content achievable through different measurement configurations.

- Chapter 5 looks beyond the present work, suggesting future development of the approach and application to new EO contexts.

Subjects of relevance for the work, but that are not essential in its logic flow, are dealt in specific notes instead of appendixes, in order to keep them close to the main arguments they are related to.

Chapter 1

Stationary view of the Earth atmosphere

1.1 Basic assumptions

The atmosphere is essentially a gaseous shell surrounding the (solid and liquid) Earth. It is a mixture of different gases, neutral for more than 99% in mass (cf. [47]). Thermodynamics and Earth gravitation determine the gas density at different height: a limit height of the shell can be arbitrarily set around 1000 km, where the gravitational field is about 30% of the surface value, the average particle velocity (mainly due to UV solar radiation absorption) equals the escape threshold ($\sim 11\ km/s$), and the gas density is approximately $10^{-15}\, kg/m^3$ (with respect to about $1.2\, kg/m^3$ at the sea level) (cf. [15]). The goodness of such limit definition is of course dependent on the problem we are dealing with: in our case we will see in § 2 that the effects on the signal delay at GPS (more generally GNSS) frequencies ($1 \div 2\ GHz$) due to the neutral and charged parts of the atmosphere vanish even at lower heights than 1000 km.

In spite of the complexities of atmospheric phenomena, the majority of the properties we are interested in can be described in the ideal gas approximation, thus through the basic thermodynamics variables, i.e. temperature and partial pressures of the different gas species (both neutral and charged ones).

1.2 Neutral atmosphere

1.2.1 Thermal classification of the atmosphere

At the equilibrium, a gas in a gravitational field follows the hydrostatic law, with pressure locally balancing the gravitational force. It can be written as a vector differential equation and simplified to the scalar form if the problem is essentially one dimensional, as it happens for the Earth atmosphere where the gravitational field is (quasi) homogeneous (Fig. 1.1):

$$dP = -\rho \cdot g \; dz \tag{1.1}$$

with P, ρ and g, respectively pressure, density and gravitational acceleration at height z. As a consequence the P gradient depends on g and on the local temperature T, through ρ, according to the ideal gas law (cf. [18]):

$$P = R \; \rho \; T \tag{1.2}$$

being R the gas constant[1].

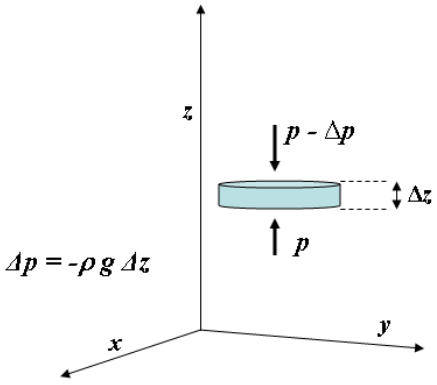


Figure 1.1: Atmospheric hydrostatic vertical balance.

In any case from 1.1 we see that $\frac{dP}{dz}$ is always negative, thus P has a monotonic decreasing profile with z, typically as in Fig.1.2.

[1]When ρ is expressed as mass density, R depends on the gas specie, or mixture of species. For the dry air it is assumed $R = R_d \simeq 0.29 \; kPa \cdot K^{-1} \cdot m^3 \cdot kg^{-1}$. Wet air is less dense than dry air and consequently it is $R_w > R_d$. The opposite happens if there is a liquid part in air. Generally we prefer to maintain $R = R_d$ and account for such differences introducing the virtual temperature, T_v, instead of T in Eq. (1.2), with $T_v = T \cdot (1 + 0.61 \cdot r - r_L)$, and r and r_L water-vapour and liquid-water mixing ratios respectively.

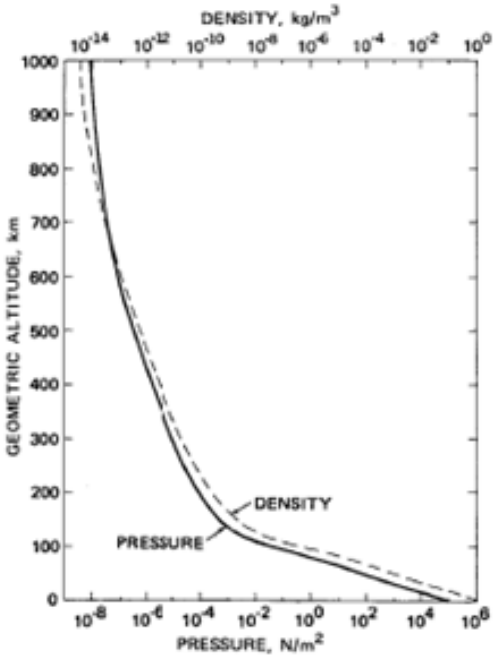


Figure 1.2: Typical vertical profiles of density and pressure (after [43]).

The temperature vertical gradient, dT/dz, changes the sign at a number of height levels, as it depends on the energy absorption and transfer mechanisms that are effective at different heights. The average profile allows to distinguish

Gas	Concentration
N_2	$78,08\%$
O_2	$20,95\%$
CO_2	320 ppm
Ne	18 ppm
He	5 ppm
CH_4	2 ppm
H_2	$0,5$ ppm
K_r	$0,11$ ppm
X_e	$0,08$ ppm
O_3	$0,04$ ppm

Table 1.1: Concentrations of main atmospheric gases from [47]. Values are average quantities at the sea level for the dry atmosphere. Water vapour is very variable and can contribute from 0 to 4% of the total atmospheric composition.

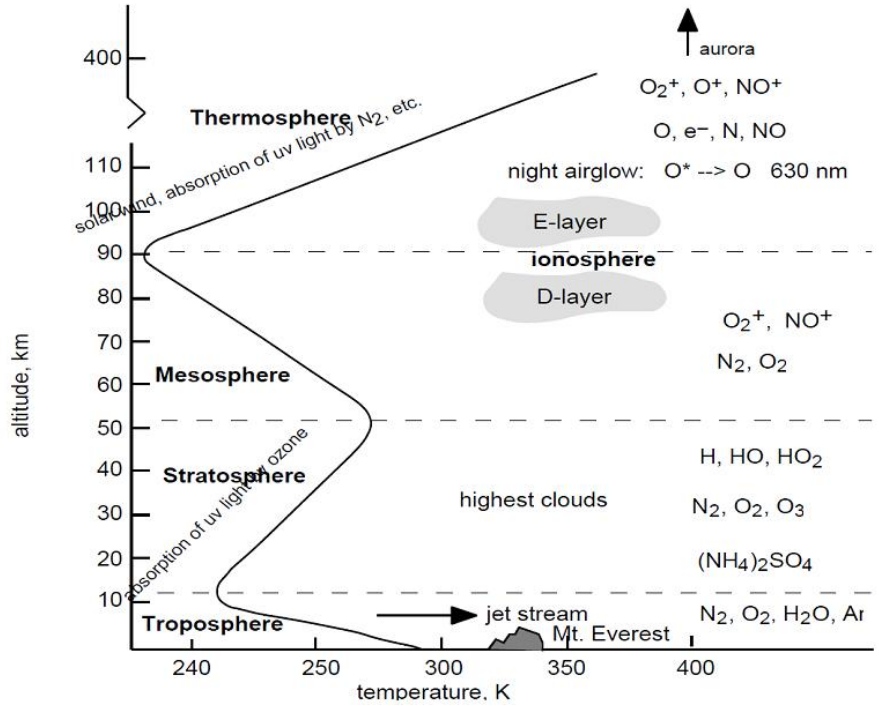


Figure 1.3: Temperature profile and principal chemical components in atmosphere (after [27]).

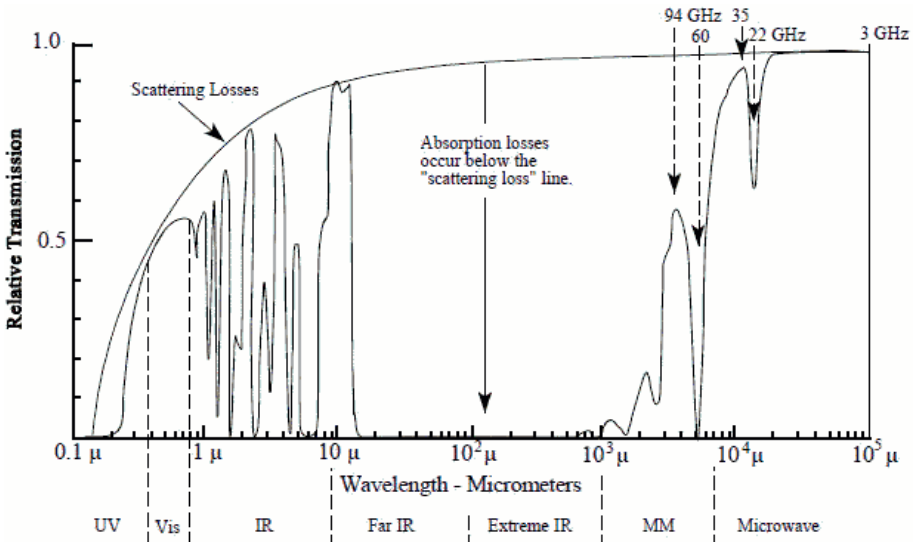


Figure 1.4: Atmospheric radiative transmission.

four atmospheric regions (Fig.1.3): troposphere, stratosphere, mesosphere and thermosphere.

At the sea level the gas composition of the dry atmosphere is as in Tab. 1.1. In the atmosphere there can be also a component of water vapour very variable, spanning from 0 to 4% of the total atmospheric composition.

The amount of solar radiation reaching the Earth surface depends critically on the amount of condensed vapour (i.e. cloud coverage, fog, etc.) that can be present in the low atmosphere, but, on average, the atmosphere is relatively transparent to solar radiation around its peak frequencies (Fig.1.4). The main absorber of solar radiation is thus the Earth surface: hence the surface layer is on average the warmer layer of the lower atmosphere and (still on average) temperature monotonically decreases up to a height of 8-14km (cf. [47, 27]). This height (tropopause) limits the lower part of the atmosphere, named troposphere: it is minimum at the Earth poles and maximum at the tropics, and however it shows seasonal fluctuations. Transfer of heat from the surface to the upper layer of troposphere happens little through conduction, much more through irradiance at IR frequencies (i.e. close to the pick of the warmed Earth emission), that is partially absorbed by molecules of CO_2 and N_2O (greenhouse effect), and through convection. When the lapse rate $(-dT/dz)$ is higher than a threshold value, convection can be activated. Such value depends on the presence of water vapour: namely it is greater for dry/unsaturated atmospheric parcels than for wet saturated ones[2]. When convection starts, we have transfer of heat due to

[2]For a simple justification of this phenomenon we can proceed imaging a theoretical experiment of a dry air parcel raising up from a starting layer A, with ambient temperature T_A and pressure P_A, to an upper layer B, with T_B and P_B. Due to the fact that air is a bad heat conductor (see Fig. 1.5), we can reasonably assume that a raising parcel has an adiabatic behaviour. This means that while the parcel pressure P_p pressure instantaneously rearranges towards the ambient one, its temperature T_p does not, and simply follows an adiabatic transformation due to the change of pressure. Thus in A we have $(P_p, T_p) = (P_A, T_A)$ and in B we have $(P_p, T_p) = (P_B, T_{p_B})$ where:

$$T_{p_B} = T_A \left(\frac{P_B}{P_A}\right)^{1-\frac{1}{\gamma}}$$

and where we have used the known ideal gas laws $PV/T = \text{const}$ (equivalent to Eq. (1.2) and $PV^\gamma = \text{const}$, the latter valid for quasi-static (i.e. always at the equilibrium) adiabatic transformations, with $\gamma = C_P/C_V$, ratio between the specific heats at constant pressure and volume respectively ($\gamma \simeq 1.4$ for dry air). For Eq. (1.2), if $T_{p_B} > T_B$ the parcel results less dense (i.e. less heavy) than the surrounding and continue rising triggering convection. On the

vertical transport (thus mixing) of mass, ad *in primis* of water vapour, that evaporates from the Earth surface and that sometimes can be lifted to layers whose temperature and pressure conditions can cause condensation and then precipitation. This transport mechanism can be effective at different atmospheric levels up to the whole troposphere, sometimes even overshooting the tropopause, into the lower stratosphere, in these cases often associated to extreme precipitation events.

When activated, convection is very efficient in transferring heat and in a certain way it acts as stabiliser of the tropospheric lapse rate, preventing to overpass the limit thresholds. In fact diurnal cycles and horizontal dynamics (i.e. advections of air masses), continuously change the lapse rate with respect to the average profiles, generating what are defined as meteorological events.

Above the tropopause the atmosphere is essentially stable, no matter of the lapse rate sign. In particular the stratosphere has a positive temperature gradient (i.e. a negative lapse rate), resulting exceptionally stable to vertical motions, that practically are totally inhibited. Stratospheric warming is not linked to Earth surface phenomena, but it depends more directly on solar radiation, namely UV ($0.1 \div 0.2\ \mu m$) absorption due to oxygen photodissociation: $O_2 + h\nu \rightarrow 2\,O$, with $h\nu$ energy of UV photons. In a height range of $20 \div 50\ km$ the amount of molecular oxygen sustains an exothermal reaction producing ozone, $O_2 + O \rightarrow O_3$, that in turn is an efficient UV radiation absorber, through the dissociation $O_3 + h\nu \rightarrow O_2 + O$.

---

contrary if $T_{p_B} \leq T_B$ convection is inhibited.
If parcel is moist the mechanism is the same, except if the raising parcel meets saturation conditions that provoke vapour condensation due to the decreasing temperature: in this case latent heat is released and it warms the parcel. If convection conditions are satisfied, parcel warming accelerates its lift and the consequence can be a positive feedback, with heavy convective precipitation, until the most of all water vapour is consumed, and equilibrium conditions re-established. Using also the hydrostatic law 1.1 we find the threshold gradient for the stability of the dry atmosphere as $\frac{dT}{dz} \simeq 9.8\,K/km$. For the saturated moist air we have to add the contribution of the water latent heat. However it strongly depends on the temperature thus its contribution is very variable and it leads to a lapse rate varying from 4 to $9.5\,K/km$. This means also that average wet lapse rates depend on latitude. Of course different heat mechanism and horizontal motions complicate the problem, but previous results reveal fairly good in explaining several atmospheric phenomena.

Over the stratopause the lapse rate changes its sign again, as density allows oxygen to be essentially stable, and the previous reactions are much less efficient with respect to radiative cooling. This region, between about 50 and 85 km is named mesosphere. At the mesopause we find the lower temperature value (around $-80\ ^oC$) of the whole atmosphere.

Beyond this height begins the thermosphere, where temperature raises again due to different absorption mechanisms of high energy radiation, whose principal are:

$$\begin{cases} O + h\nu & \rightarrow O^+ + e^- \\ H + h\nu & \rightarrow H^+ + e^- \end{cases} \tag{1.3}$$

These are ionisation reactions (the latter increasing its relative efficiency for increasing heights) that generate the non neutral part of the atmosphere. Here densities are so low ($P \ll 10^{-1}\ mbar$) that the thermodynamic definition of temperature does not hold any more, and temperature has to be intended as a parameter expressing the average kinetic energy of particles (hence named kinetic temperature), which are no more in local thermodynamic equilibrium. These low densities have also a negligible effect on MW signal delay, for what concerns the neutral part of the gas. On the contrary we will see that effects of free electrons are definitely non negligible. Thermosphere is considered extending up to about 600 km: thus at the thermopause we are still at a height less than 10% of the Earth radius.

Above thermosphere we can say that the gas kinetic temperature is very high but strongly coupled to solar irradiance (i.e. shows great latitudinal and diurnal variations from about 500 to 1700 ^oC) but essentially it does not change with height. The high particle energy and the low gravitation field make the gas vanishing ($P < 10^{-27}\ mbar$); however its charged part is still relevant with respect to lower layers, up to thousands of kilometres: this transition zone between the thermosphere and the interplanetary vacuum is named exosphere.

1.2.2 Standard atmosphere

In 1962 a standard model for the atmosphere was introduced by the WMO (World Meteorological Organisation), successively upgraded in 1976 (cf. [43]), as reference in meteorology and in aerospace instrumentation development. Here we will derive some basic characteristics of this standard model, that are useful to understand some typical atmospheric features, relevant for our work.

The standard atmosphere is a stationary model for the dry atmosphere at a mean latitude (around 45^o), from 0 to 1000 km, that assumes hydrostatic equilibrium (Eq. (1.1)) and ideal gas law (Eq. (1.2)) (cf. [47]). The latter holds for the whole atmosphere, as gas density is low enough even at its highest values (i.e. at the sea level), and the charged part maintains always a small part with respect to the neutral one, at any ionospheric level. Eq. (1.2) is linear with respect to the gas numerical density[3], and it consequently implies the Dalton law:

$$P_{tot} = \sum_i P_i \tag{1.4}$$

with P_i partial pressure of the i_{th} gas specie. Eq. (1.4) can be written in the troposphere as:

$$P_{tot} = P_d + e_w \tag{1.5}$$

with e_w water-vapour partial pressure and P_d pressure due to the dry atmosphere.

In the hydrostatic equation, g, the gravitational acceleration, is not constant, but slightly varies with height, due to the reduction of the gravity field as $1/D^2$, with D distance from the Earth centre. This effect can be negligible for a large part of the troposphere, but towards and especially above the tropopause it must be included in the equilibrium equations. Models commonly account for it maintaining g constant to the value at the sea level (g_0) and rescaling the height, by means of the introduction of the geopotential height, h, (instead of

[3]If for ρ we mean the number of particles for unit volume (instead of mass density), in Eq. (1.2) for a mixture of gases (labelled with i) we can write $\rho_{tot} = \sum_i \rho_i$. R is constant and thus does not depend on i. T does not depend on i too, provided we have local thermodynamic equilibrium for all gas species (i.e. we have not a multitemperature fluid).

the physical height) defined as:

$$h = \frac{1}{g_0} \int_{z_0}^{z} g(\zeta, \phi) d\zeta \tag{1.6}$$

where $g(z, \phi)$ is the gravity acceleration at z at the latitude ϕ (neglecting longitudinal effects). From the universal gravitational law we can write $g(z, \phi) = g_0 \cdot \frac{R_T}{R_T + z}$, with $R_T = R_T(\phi)$, the Earth radius at the latitude ϕ, and consequently:

$$h = \frac{z \cdot R_T}{R_T + z} \tag{1.7}$$

In the troposphere the difference between z and h is less than 2%. From Eq. (1.2) and (1.6) we obtain the hypsometric equation:

$$h_2 - h_1 = \frac{R_d}{g_0} \int_{P_1}^{P_2} T \frac{dP}{P} \tag{1.8}$$

with $R_d = R/M_d$ gas constant R for the dry atmosphere, being M_d the apparent molecular weight[4].

Thanks to the mean integral theorem, we can integrate Eq. (1.8) as:

$$P_2 \cong P_1 e^{-\frac{(z_2 - z_1)}{H_d}} \tag{1.9}$$

with $H_d = \frac{T \cdot R}{g_0 M_d}$ named *scale height* and T a mean temperature between the two height values. From Eq. (1.9) we can estimate that about the 80% of the atmospheric mass is concentrated in the tropopause. In addition, in stationary conditions, when dynamics mixing processes are not efficient (i.e. far above the stratopopause), the scale height gives an indication on how gases with different molecular weights distributes with height, the lightest increasing their relative concentration with height.

Air is a low efficient heat conductor (Fig. 1.5). This, and the lack of other efficient mechanisms of heat transfer among gas parcels, allows to reasonably assume that a moving gas parcel behaves adiabatically (cf. [18]):

$$\frac{T}{P^{\frac{\gamma - 1}{\gamma}}} = cost. \tag{1.10}$$

[4] $M_d = \sum_i m_i / \sum_i \frac{m_i}{M_i}$, with m_i and M_i, molecular mass and weight respectively, for the i^{th} gas specie of the dry mixture. For the dry atmosphere $M_d = 28.97$ Kmol.

with $\gamma = C_P/C_V$, ratio of specific heats at constant pressure and volume respectively, for the gas mixture. Differentiating Eq. (1.10) and using Eq. (1.1) and (1.2), we find:

$$T = T_0 - \Gamma_d(z - z_0) \tag{1.11}$$

with T_0 temperature at z_0 and $\Gamma_d = \frac{g}{R_d}\left(\frac{\gamma-1}{\gamma}\right)$ adiabatic lapse rate.

From statistical mechanics we know that, for a biatomic gas, $\gamma = \frac{7}{5}$ (cf. [49]) and $\Gamma_d \cong 9.8\ K/km$.

When condensing water vapour releases latent heat. A rising moist gas parcel, with saturated water vapour, cools slower than a dry one, as, in saturation conditions, a temperature reduction is accompanied by vapour condensation and thus by heat release within the parcel. The adiabatic lapse rate for wet saturated air is very variable, it strongly depends on temperature, thus, on average, it depends on latitude too[5]

The following relationships thus give the standard temperature profile for dry atmosphere in the troposphere (the one of main interest for this work):

$$\begin{cases} T(z) &= T_0 - \Gamma_s \cdot z \\ T_0 &= 288.15\ K \\ \Gamma_s &= 6.5\ K/km \end{cases} \tag{1.12}$$

Pressure is given by Eq. (1.1) and (1.8), assuming Eq. (1.12) for temperature:

$$\begin{cases} P(z) &= P_0\left(\frac{T_0}{T}\right)^{-\alpha} \\ P_0 &= 1013.25\ hPa \\ \alpha &= 5.25 \end{cases} \tag{1.13}$$

The main sources of profile departures from this standard are due water vapour and atmospheric horizontal and vertical dynamics, from large to local scales, that generate the meteorological phenomena. Mass air advections, convections and precipitations make the temperature values and lapse rate very

[5]See note 2.

variable, conditioned by latitudes, annual and diurnal cycles, up to local characteristics, such as orography, proximity of water basins, land coverage, etc. Especially the lower tropospheric layers can exhibit temperature behaviour very different from standard, including temporary lapse rate inversions. Pressure instead shows relevant changes in local values but not in the profile trends.

The standard atmosphere fixes the tropopause at 11 km: above it we have to include the different mechanisms of direct absorption of solar radiation (described in the previous section), that are efficient in the high atmosphere. The resulting profiles in the standard atmosphere are given by the following relationships, where T, P and h (geopotential height) must be expressed in K, kPa and km respectively (cf. [38]):

$$\begin{cases} T = 288.15 - 6.5 \cdot h & h \leq 11 \\ T = 216.65 & 11 < h \leq 20 \\ T = 196.65 + h & 20 < h \leq 32 \\ T = 228.65 + 2.8 \cdot (h - 32) & 32 < h \leq 47 \\ T = 270.65 & 47 < h \leq 51 \end{cases} \tag{1.14}$$

$$\begin{cases} P = 101.325 \cdot (288.15/T)^{-5.255877} & h \leq 11 \\ P = 22.632 \cdot \exp\left[-0.1577 \cdot (h - 11)\right] & 11 < h \leq 20 \\ P = 5.4749 \cdot (216.65/T)^{34.16319} & 20 < h \leq 32 \\ P = 0.868 \cdot (228.65/T)^{12.2011} & 32 < h \leq 47 \\ P = 0.1109 \cdot \exp\left[-0.1262 \cdot (h - 47)\right] & 47 < h \leq 51 \end{cases} \tag{1.15}$$

Patterns of T and P that are reported, are from the Earth surface up to the low mesosphere, and it largely comprehend all that shell of neutral atmosphere whose density brings effects on the GNSS signals of relevance[6].

[6]From the standard atmosphere values we can easily verify that at a height of 50 km density is about 1/1000 than on the Earth surface; then it rapidly decreases with height.

1.2.3 Water vapour distribution and laws

Water vapour is the main variable gas component of the atmosphere. Its presence is essentially limited to the troposphere and in particular to the low tropospheric layers. In fact it is part of the Earth water cycle, thus continuously released in atmosphere by the Earth surface (from oceans, soil, vegetation, etc.) and then removed as precipitation. Water vapour does not normally exceed a concentration of 4% in atmosphere (cf. [47]), but it is a main "ingredient" in meteorology because of the importance of precipitation phenomena and of the absorption and release of latent heat during the processes of evaporation and condensation respectively, that condition the atmospheric dynamics (cf. [4]).

Water vapour can be modelled in the ideal gas approximation for a large number of problems. The laws introduced for the dry atmosphere can consequently applied to water vapour too, provided we use proper values for the gas parameters: e.g. it holds the relationship:

$$e_w = R_w \ \rho \ T \tag{1.16}$$

with e_w vapour partial pressure and R_w gas constant for water vapour.

The saturation pressure e_s can be obtained from the semi-empirical Buck law (cf. [8]) (derived from the Clausius-Clapeyron law, cf. [18]):

$$e_s = A \cdot 10^{\frac{aT_c}{b+T_c}} \tag{1.17}$$

with $A = 6.11\,\mathrm{hPa}$, $a = 7.5$ and $b = 237.7$, and being T_c the temperature expressed in Celsius degrees.

Another parameter of common use is the relative humidity, RH, that expresses the percentage of vapour, w, with respect to the saturation value, w_s:

$$RH = 100\frac{w}{w_s} \tag{1.18}$$

w and w_s are expressed in term of mass mixing ratio, that is $w = m_w/m_d$ (commonly in g/kg).

The relation between vapour pressure and total pressure is (cf. [47]):

$$e_w = \left(\frac{w}{w+\epsilon}\right) \cdot P \tag{1.19}$$

with $\epsilon = R_d/R_w \cong 0.622$.

Assuming $e_w \ll P$ and $e_s \ll P$ we can write:

$$\begin{cases} w & \cong \epsilon \cdot \frac{e_w}{P} \\ w_s & \cong \epsilon \cdot \frac{e_s}{P} \\ e_w & = \frac{RH}{100} \cdot e_s \end{cases} \tag{1.20}$$

Finally we recall an important property of the water relevant for the present work: the built-in dipole moment, $\mu_w = 6.162 \cdot 10^{-30}(\mathrm{C} \cdot \mathrm{m})$. As a consequence water vapour molecules interact with the electromagnetic radiation differently with respect to the other gas components[7] (cf. [14, 10]), but this will be matter of § 2.2.2.

1.3 Non neutral atmosphere

1.3.1 Plasma component of the atmosphere

We have already seen in the thermosphere how oxygen and (for higher layers) hydrogen reactions are responsible for the absorption of high energy radiation (principally from the sun), and thus for gas heating. In this way they are also sources of ions, and specifically of free electrons. Free electrons are of interest for this work because they affect the group (and phase) velocity of MW signals, as a function of frequency (i.e. in a dispersive way). These particles constitute a component of plasma gas in the atmosphere. They populate a region globally identified as ionosphere (Fig. 1.3), from about 60 km over the sea level, to more than 1000 km, i.e. extending from the higher layers of the mesosphere and

[7] In effect there are a number of other polar molecules in atmosphere e.g. CO ($\mu = 0.39 \cdot 10^{-30}\,Cm$), NO ($\mu = 0.92 \cdot 10^{-30}\,Cm$), N_2O ($\mu = 1.002 \cdot 10^{-30}\,Cm$), O_3 ($\mu = 1.7 \cdot 10^{-30}\,Cm$), but they have $\mu \ll \mu_w$ and moreover their concentrations are marginal. In sum their effect is negligible for the aim of this work.

including all the thermosphere, up to a large part of the exosphere, where it is identified more specifically as protonosphere.

On average the absolute density of free electrons increases with height up to the thermopause, with a maximum of about $10^6\ cm^{-3}$, then it smoothly decreases in the exosphere, so that at 1000 km it is still relatively high (approximately $10^5\ cm^{-3}$). However the ionised matter is generally less than 1% of the neutral mass (cf. [31]).

At these heights light molecules prevail, such as oxygen, helium and hydrogen (cf. [47]), that are ionised by UV and X rays (mainly from solar radiation). Very low densities, and consequently large mean free paths, then reduce recombination probabilities of ions into neutral matter (Fig. 1.7), which increases the mean life-times for the charged matter.

The Total Electron Content (TEC) is the density of electrons along a vertical path, and it is a parameter commonly used for defining the ionisation degree of the ionosphere and modelling some major phenomena of interest also for our work (see § 2.2.3).

TEC, and more generally the local electron density, changes considerably according to the spatial and temporal variability of the high energy radiation flux from the sun, that is we observe latitudinal and diurnal cycle effects, and also longer term variations linked to solar activity. In addition free ions tend to follow magnetic field lines, thus their distribution is also driven by the magnetic field topology around the Earth. This dynamic behaviour is very difficult to model, but typical (average) features at different heights allow a classification of the ionosphere into regions that historically are distinguished mainly through their effects on radio communication at different frequencies.

1.3.2 Ionosphere classification

The main interest in understanding ionosphere took origin in its effect on radio transmission, due to the presence of free charges (cf. [22]). The historical classification of ionosphere is still used, and it consists of some regions flagged with

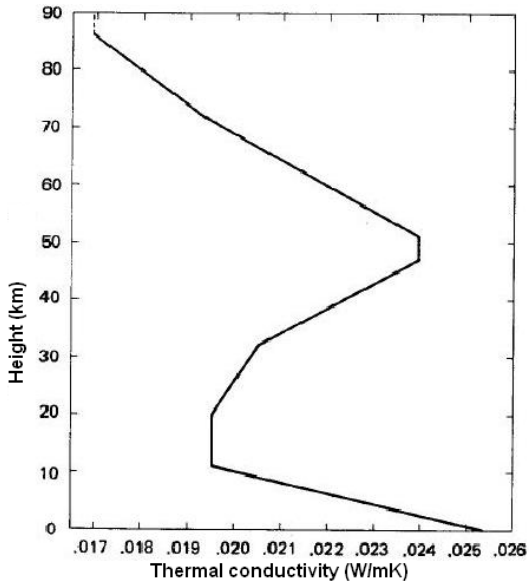


Figure 1.5: Vertical profile of thermal conductivity for dry air (after [43]. For comparison, water is about 0.58 ad iron 80 W/mK.

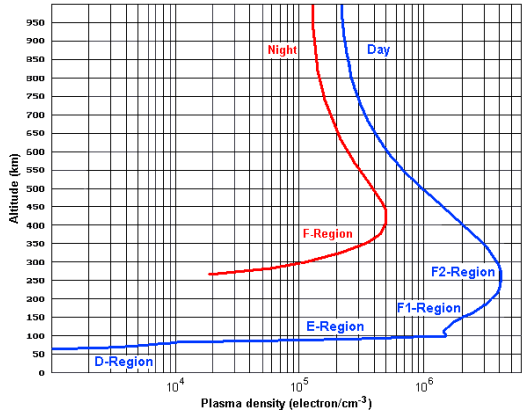


Figure 1.6: Average ionospheric densities. Note the difference between nocturnal plasma densities (red line) and diurnal ones (blue line).

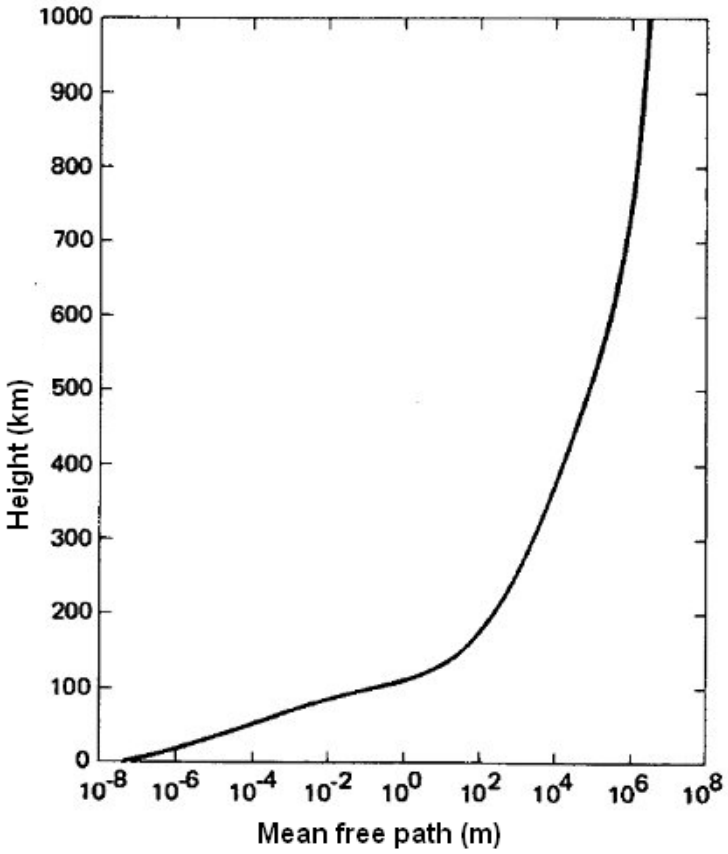


Figure 1.7: Mean free path of particles in atmosphere (after [43]).

alphabetic letters.

D is the lowest region, located between 75 and 95 km, and it is responsible for the propagation of radio signals at frequencies $\leq 2\,MHz$. This layer exists only during daytime, as at this height the density makes the recombination processes effective enough to neutralise the charged matter during night, when the ionisation engine is off.

E is the second region located between 95 and 150 km, and exists also during the night (due to the reduced efficiency of the recombination processes). ${O_2}^+$ are the principal ions of this region, that allow signal transmission up to $10\,MHz$.

F is the third region located from 150 to about 500 km. O^+ and also NO^+ are the prevalent ions, and, also due to its large extension (that increases during the daytime), is the most important region for high frequency radio transmission.

The *topside region* is the last region extending over the F region, which is composed mainly by H^+ ions. Low particle density however makes this region less relevant in radio transmission phenomena.

Chapter 2

Atmospheric effects on GNSS signals

2.1 Propagating electromagnetic signals

A GNSS signal is an electromagnetic wave packet with a carrier frequency at L band, i.e. between 1 and 2 GHz (according to the GNSS constellation and signal type), travelling from the satellite towards the Earth, where receivers decode and process it. Such signal comes from an essentially vacuum environment around the satellite, then crosses the rarefied but ionised gas of the high atmosphere (i.e. the ionosphere), and finally the low atmosphere, which is neutral but whose density brings non negligible effects on radiation, due to both the dry component and the water vapour, in different ways.

The aim of this chapter is to summarise the basic processes that affect an electromagnetic signal travelling in a medium like the atmosphere, in order to introduce the equations and the built-in approximations we will refer to, when dealing with ionospheric free GNSS signals. Such equations have been implemented in a simulator of GNSS signals travelling in the atmosphere, that has been built to implement the core part of the whole thesis, i.e. the tests on the algorithms for retrieving tropospheric profiles from GPS-like signals.

An electromagnetic wave in the vacuum comes as solution of the Maxwell equations for each component of the electric and magnetic field vectors, **E** and

B:

$$\begin{cases} \boldsymbol{\nabla} \cdot \mathbf{E} = 0 \\ \boldsymbol{\nabla} \cdot \mathbf{B} = 0 \\ \boldsymbol{\nabla} \wedge \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \\ \boldsymbol{\nabla} \wedge \mathbf{B} = \frac{1}{c^2}\frac{\partial \mathbf{E}}{\partial t} \end{cases} \tag{2.1}$$

If $u(\mathbf{r}, t)$ is the generic scalar component of the electric or magnetic field, the wave equation is:

$$\nabla^2 u - \frac{1}{v^2}\frac{\partial^2 u}{\partial^2 t} = 0 \tag{2.2}$$

where $v = \frac{c}{\sqrt{\mu\epsilon}}$ is the phase velocity, that in the vacuum equals the light velocity c.

A solution of Eq. (2.2) is the plane wave which is of particular interest when dealing with electromagnetic waves far from their source. In the linear polarisation case it can be expressed as:

$$\begin{cases} \mathbf{E}(\mathbf{r}, t) = E_0 e^{i(\mathbf{kr}-\omega t)}\epsilon_{\mathbf{1}} \\ \mathbf{B}(\mathbf{r}, t) = B_0 e^{i(\mathbf{kr}-\omega t)}\epsilon_{\mathbf{2}} \\ \epsilon_3 = \frac{\mathbf{k}}{|k|} \\ \epsilon_2 = \epsilon_3 \wedge \epsilon_1 \end{cases} \tag{2.3}$$

The first two equations give the fields through their real part, and the last two equations express the orthogonal relationship exiting among **E**, **B** and the propagation vector **k**, being ϵ_1, ϵ_2 and ϵ_3 a normal Euclidean basis (Fig. 2.1).

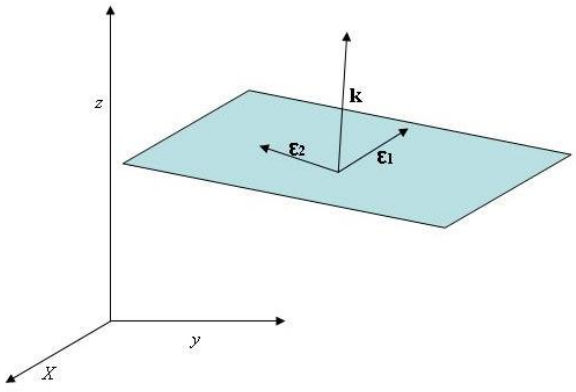


Figure 2.1: Geometry of a plane wave.

What we have written for a single wave frequency (i.e. in the ideal monochromatic case) can be straightforwardly generalised to generic wave packets, as real electromagnetic signals are, following the Fourier theorem, a linear combination of an infinite number of monochromatic waves, with given amplitude and phase (cf. [22]).

In the following sections we will address the issue of what happens to local fields when an electromagnetic wave propagates through a gas medium.

2.2 Electromagnetic signals in gas media

An electromagnetic wave propagates the **E** and **B** vector fields: in a given point this means we have time-varying fields. A gas particle interacts with these fields, both if it is neutral and furthermore if it is polarised or ionised, and due to this interaction the fields partially change.

A rigorous analysis of these phenomena would need a complex quantum mechanical approach, that is out of the scope of the present work. Nevertheless, the interaction can be modeled using a classical approach with results (at least qualitatively) coherent with the semi-empirical approach of wide use, based on the geometrical optics approximation and on experimental findings. However our approach will permit also to explicitly quantify the magnitude of some approximations that are included in the semi-empirical approach.

2.2.1 Interaction with a neutral non-polarised gas

The neutral non polarised part of the atmospheric gases is essentially what is called dry atmosphere, which is mainly concentrated in the troposphere and low stratosphere.

An electromagnetic wave interacts with neutral atoms or molecules through the mechanisms of excitation of the particle energy levels, responsible for absorption and (spontaneous plus stimulated) emission phenomena. As a consequence an electromagnetic wave crossing a gas layer will be generally subjected

to changes in intensity and phase.

A very rough but still useful description of the phenomena can be obtained by a simple classical representation of each gas particle as a negative charged cloud due to electron(s), surrounding a positive charged nucleus, both having spherical shapes with coincident centres. In this simplified view the presence of an electric field **E** induces a cloud deformation, and consequently the displacement of the negative cloud centre from the positive nucleus one. Such displacement in turn produces a restoring electric force, between the two centres.

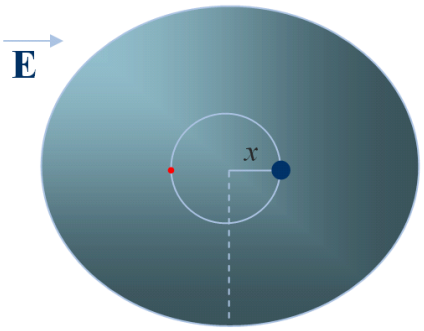


Figure 2.2: Simplified classical view of charge displacement in an atom immersed in an electric field.

We can make the following assumptions that are more than reasonable in many cases, including the one of interest for this work:

1. the displacement is instantaneous with respect to **E** variations (i.e. electron inertia is negligible with respect to the wave period);

2. the effect of **B** on this phenomenon is negligible;

3. the electron cloud displacement is much less than the cloud dimension (i.e. the intensity of **E** so low that induces just a perturbation in the particle charge symmetry).

In these hypotheses our problems reduces to the study of an harmonic oscillator forced by an electric force, periodic with time. In fact from item 2 we can neglect **B**; item 1 says that **E** and the charge cloud move in phase, and it is easy

to demonstrate that item 3 leads to a restoring force f that is elastic. In fact the charge responsible for the restoring force is only the one inside the sphere of radius x in Fig. 2.2, and thus we have[1]:

$$\mathbf{f} \simeq -\frac{Zq}{4\pi\epsilon_0} \cdot \frac{\rho_q \left(\frac{4\pi}{3}\right) x^3}{x^2} \cdot \frac{\mathbf{x}}{x} \tag{2.4}$$

with $\mathbf{x}$ cloud displacement vector of module x; q proton charge; Z atomic number of the gas particle; ρ_q mean charge density of the electron cloud roughly equivalent to $\frac{Zq}{\left(\frac{4}{3}\pi a^3\right)}$ with a mean particle radius. At the end we find $f \propto x$ that reduces the problem to the oscillator one:

$$\ddot{x}(t) + \omega_0^2 x(t) = \frac{q\,E}{m} \tag{2.5}$$

being $\omega_0 = \sqrt{\frac{Z^2q^2}{\frac{4\pi}{3}ma^3}}$ the oscillator own frequency, with m mass of the electron cloud, and $\mathbf{E}$ forcing field.

In the ideal oscillator case if f varies periodically with a frequency close to the resonant one (i.e. ω_0), the amplitude of the periodic solutions tends to diverge to infinity. In real cases (e.g. in mechanical oscillator) for large amplitudes the oscillator does no more behave as an harmonic one. Normally a damping term is added to the oscillator equation, that accounts for inelastic phenomena, dissipating the energy acquired (in excess) by the forcing term. The quantum counterpart of such inelastic interactions are the radiation absorption and stimulated emission that happen in real atoms (or molecules) when radiation frequencies are close to the resonant atomic (or molecular) frequencies. In our still simplified view we can write:

$$\ddot{x}(t) + 2\gamma\dot{x}(t) + \omega_0^2 x(t) = \frac{-q}{m} E_0 e^{i\omega[t-\frac{z}{c}]} \tag{2.6}$$

On the left side of the equation we have explicitly included the forcing term, or more correctly one of its Fourier component, with the time phase delay due to the distance z of the gas particle from the radiation source (being c the speed of

[1]The charge outside such sphere gives a null total contribution to f, thanks to the Gauss theorem.

light in the vacuum). The solution is:

$$x(t) = \frac{qE_0\, e^{i\omega[t-\frac{z}{c}]}}{\mathcal{A}(\omega, \omega_0, \gamma)} \tag{2.7}$$

with:

$$\mathcal{A}(\omega, \omega_0, \gamma) = m(\omega^2 - \omega_0^2 - 2i\gamma\omega) \tag{2.8}$$

GNSS systems have been designed in order to avoid problems of signal amplitude reductions, thus their operational frequencies are far from the absorption lines of atmospheric molecules (i.e. $|\omega - \omega_0| \gg 0$). In other words we could neglect the term with γ and simplify Eq. (2.7) with $\mathcal{A} \simeq m(\omega^2 - \omega_0^2)$.

Previous equation says that our neutral and non polarised gas particle becomes an oscillating dipole when "forced" by an electromagnetic radiation, with momentum $\mathbf{P}(r,t)$ proportional to the incident electric filed. Thus it becomes in turn a source of electromagnetic radiation, according to the following equations (cf. [19]).

$$\begin{cases} \mathbf{E}(r,t) &= -\frac{q}{4\pi\epsilon_0}\ \left[\frac{\mathbf{e}_{\mathbf{r}'}}{r'^2} + \frac{r'}{c}\frac{d}{dt}\left(\frac{\mathbf{e}_{\mathbf{r}'}}{r'^2}\right) + \frac{1}{c^2}\frac{d^2}{dt^2}\mathbf{e}_{\mathbf{r}'}\right] \\ \mathbf{B} &= \frac{\mathbf{e}'_{\mathbf{r}} \times \mathbf{E}_{\mathbf{r}}}{c} \end{cases} \tag{2.9}$$

The apex refers to features at the time they radiate what is measured at the time t in a generic point p (Fig. 2.3). If the source is far from p Eq. (2.9) simplifies in:

$$\mathbf{E}(r,t) \simeq -\frac{q}{4\pi\epsilon_0}\ \left[\frac{1}{c^2}\frac{d^2}{dt^2}\mathbf{e}_{\mathbf{r}'} + O\left(\frac{1}{r'^2}\right)\right] \tag{2.10}$$

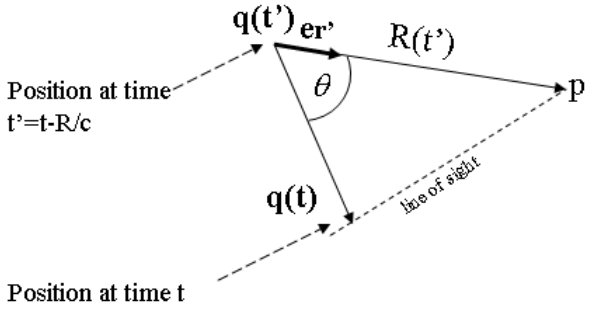


Figure 2.3: Geometry of the emitting dipole.

Eq. (2.10) gives the contribution to the measured field due to each gas particle, and it says that only the acceleration perpendicular to the direction from the particle to p gives a non negligible contribution to the far field.

The second time derivative of $\mathbf{e}'_{\mathbf{r}}$ is:

$$\frac{d^2}{dt^2}\mathbf{e}_{\mathbf{r}'} \simeq \frac{a_\theta}{r'}\mathbf{e}'_{\mathbf{r}} \tag{2.11}$$

where a_θ is the acceleration computed at time $t - z/d$.

In fact in p we have: $\mathbf{r}' = r'\mathbf{e}'_{\mathbf{r}}$, thus $\ddot{\mathbf{r}}' = (\ddot{r}' - r'\dot{\theta}^2)\mathbf{e}'_{\mathbf{r}} + (2\dot{r}'\dot{\theta} + r\ddot{\theta})\mathbf{e}_\theta$. As we know that only the component parallel to $\mathbf{e}_\theta$ contributes to the radiation field (cf. [22]), and neglecting contributes of order higher than one, we obtain: $\ddot{\mathbf{r}}' = a_\theta \simeq r\ddot{\theta}\mathbf{e}_\theta$.

For our induced dipole we can consequently write:

$$\mathbf{E}(r,t) = -\frac{q}{4\pi\epsilon_0 c^2}\frac{\ddot{x}(t - r/c)}{r} \cdot \cos\theta + O\left(\frac{1}{r^2}\right) \tag{2.12}$$

The overall effect of a number of induced oscillating dipoles is the sum of the single effects plus the source field (superposition principle, cf. [19]):

$$\mathbf{E_T}(p,t) = \mathbf{E_s}(p,t) + \mathbf{E_a}(p,t) \tag{2.13}$$

In (2.13) $\mathbf{E_s}$ is the source field and $\mathbf{E_a}$ is the dipole total contribution, that we want to compute neglecting the reciprocal dipole interactions, a very reasonable approximation in rarefied media.

For a general gas volume we can integrate all the gas particle contributions at different layer z of thickness Δz (Fig. 2.4), the contribution of each layer being:

$$\mathbf{E_a}(z,t) = \Delta z \cdot \frac{q^2\omega^2\ \mathbf{E_0}}{4\pi\epsilon_0\,\mathcal{A}(\omega,\omega_0,\gamma)\,c^2}\int_{-\infty}^{+\infty} dy \int_{-\infty}^{+\infty} dx\ \frac{\eta(x,y,z)}{r}\ \cos\theta\, e^{-i\omega(t-\frac{r}{c})} \tag{2.14}$$

with:

$$\begin{cases} r &= \sqrt{x^2+y^2+z^2} \\ \cos\theta &= z/r \end{cases} \tag{2.15}$$

and with: η density of gas particles; the origin of the reference system coincident with the observer position (e.g. the centre of the GNSS receiver antenna), and

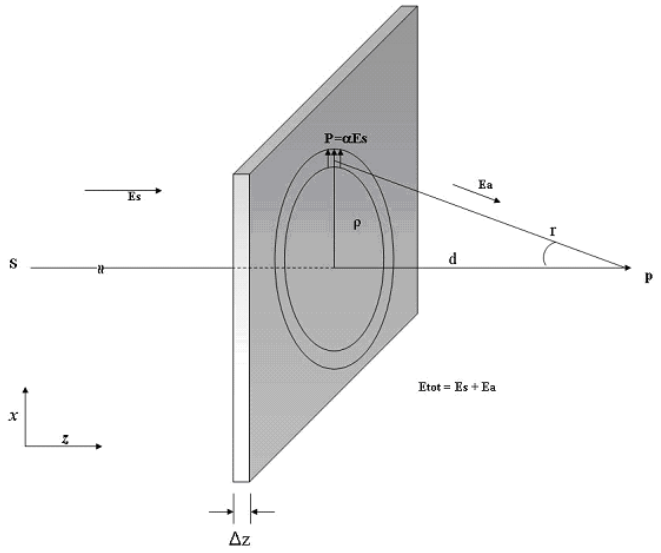


Figure 2.4: Layer of atmosphere of quasi-infinitesimal thickness, Δz, crossed by an electromagnetic signal measured in p.

the z axis pointing towards the signal source (e.g. the centre of the GNSS satellite antenna).

The integration in the whole space is apparently very complicated, but a closer look reveals that the integral in Eq. (2.14) contains an oscillating term that allows to simplify the part of integration orthogonal to z (i.e. in the xy plane) through the stationary-phase approximation (cf. [16, 40]).

The stationary phase method is a procedure for evaluation of integrals of the form:

$$I(\Omega) = \int_a^b g(x) e^{i\Omega f(x)} dx \tag{2.16}$$

where $\Omega f(x)$ gives a rapidly-varying function of x over most of the range of integration, and $g(x)$ is slowly-varying (by comparison). Rapid oscillations of the exponential term (or equivalently $\Omega \to \infty$) mean that I is approximately null over all regions of the integrand, except where $df/dx = 0$, i.e. at points of stationary phase. Regions around these points are the only that give significant non-zero contributions to the integral. Points of stationary phase are labelled x_s and defined by $f'(x_s) = 0$.

In the vicinity of the stationary phase points we have $g(x) \simeq g(x_s)$, since we remind that $g(x)$ is assumed to be slowly varying, and hence this term can be pulled outside the integral.

Expanding $f(x)$ in a Taylor series near x_s up to the second order and assuming $f''(x_s) \neq 0$, we have:

$$f(x) \simeq f(x_s) + \frac{1}{2} f''(x_s)(x - x_s)^2 \tag{2.17}$$

Substituting this into Eq. (2.16) we obtain:

$$I(\Omega) = \frac{1}{\sqrt{\Omega}} \cdot \left[\sqrt{\frac{2\pi i}{f''(x_s)}}\, g(x_s)\, e^{i\Omega f(x_s)} \right] + O(\frac{1}{\Omega}) \tag{2.18}$$

For integrals in two dimensions, we essentially proceed in the same way, and the result is:

$$I(\Omega) = \frac{1}{\Omega} \cdot \left[\frac{2\pi i \cdot g(x_s, y_s)\, e^{i\Omega f(x_s, y_s)}}{\sqrt{f_{xx}(x_s, y_s) \cdot f_{yy}(x_s, y_s) - f_{xy}^2(x_s, y_s)}} \right] + O(\frac{1}{\Omega^2}) \tag{2.19}$$

The integral of Eq. (2.14) becomes something that depends no more on the whole gas volume, but only on a cylinder of (variable) section $\mathcal{S}$ equal to:

$$\mathcal{S} = \Omega \cdot \left(\sqrt{f_{xx} f_{yy} - f_{xy}^2} \right) \Big|_{(x_s, y_s)} \tag{2.20}$$

This can be imagined as a sort of tube whose axis is on the stationary phase points. If the section is small we can approximate the tube with a line, in other words we can proceed in the geometrical optics approximation. In our problem "small" is what is less than the spatial resolution sought for our tropospheric parameters.

If η depends only on z it can be pulled outside the integrals in dx and dy, and we easily find that the ray path stays on the z axis. On the contrary if η has a gradient orthogonal to z that cannot be neglected with respect to the integral oscillating term, the stationary phase approximation needs to account for more terms than what we have done, and one of the differences in the final result is

that the ray path departs from a straight line, showing a bending, problem that we will address later in § 2.3.2.

For the moment if we assume our approximations as reasonable, from Eq. (2.14) and Eq. (2.15) we can write:

$$\begin{cases} g(x,y) &= \eta/r^2 \\ f(x,y) &= -r \\ \Omega &= \omega/c \end{cases} \tag{2.21}$$

For the derivatives of f we have:

$$\begin{cases} f_x &= -x/r \\ f_y &= -y/r \\ f_{xx} &= -(r^2-x^2)/r^3 \\ f_{yy} &= -(r^2-y^2)/r^3 \\ f_{xy} &= f_{yx} = -(xy)/r^3 \end{cases} \tag{2.22}$$

Stationary points are consequently in $(x,y) = (0,0)\ \forall z$ and they imply:

$$\begin{cases} f_{xx} &= f_{yy} = -1/z \\ f_{xy} &= f_{yx} = 0 \end{cases} \tag{2.23}$$

Eq. (2.14) becomes:

$$\mathbf{E_a}(z,t) = \Delta z \cdot \frac{i\Omega}{2\epsilon_0} \frac{q^2}{\mathcal{A}(\omega,\omega_0,\gamma)}\, \eta(0,0,z)\, E_0\, e^{-i\omega(t-\frac{z}{c})} \tag{2.24}$$

From Eq. (2.13) and Eq. (2.24) we have also:

$$\mathbf{E_T}(p,t) = \mathbf{E_s}(p,t) \left(1 + \Delta z \cdot \frac{i\Omega}{2\epsilon_0} \frac{q^2}{\mathcal{A}(\omega,\omega_0,\gamma)} \eta(0,0,z)\right) \tag{2.25}$$

It can be compared with the equations valid in the geometrical optics approximation for electromagnetic waves in non vacuum (but non ferromagnetic) media. Specifically the phase delay due to a medium of thickness Δz and refractive index n, is:

$$\mathbf{E_T}(p,t) = \mathbf{E_s}(p,t)\, e^{i\phi t} \tag{2.26}$$

with $\phi = \frac{\omega}{c}(n-1)\Delta z$. Comparing the two equations for $\Delta z \to 0$ we obtain a relationship between the macroscopic properties of the medium and the refractive index:

$$n = 1 + \frac{i\Omega}{2\epsilon_0} \frac{q^2}{\mathcal{A}(\omega,\omega_0,\gamma)}\, \eta(0,0,z) \tag{2.27}$$

Generally n is a complex number. Its imaginary part comes from the γ term in $\mathcal{A}(\omega, \omega_0, \gamma)$, and it reduces the module of the field, thus the amplitude of the wave. We have already commented that this aspect is negligible in our problem. The real part of n instead can be interpreted as a term $n = c/v$, thus a change in the propagation velocity of the wave from c, in the vacuum, to v, in the (gas) medium, and consequently in the travel time of the wave packet: this is the property at the basis of atmospheric sounding by means of GNSS signals.

From our derivation we find $n = n(\omega_0, \gamma, \omega, \eta)$ and specifically $n \propto \eta$. It means that n depends on the gas composition, through the atoms properties contained in ω_0 and γ, on the incident radiation through ω and on the local thermodynamic properties through η.

In real atoms and molecules the resonant frequencies are much more than one; they are related to emission/absorption lines and they are given by quantum mechanics. In the troposphere, non polar molecules typically have major resonant lines around the visible or near infrared frequencies ($\sim 10^{15}\, Hz$): if we assume $\eta(z) = \frac{1}{k_B}\frac{P}{T}$, we find $n \simeq 1.00016$.

From this point of view the classical approach we derived from Eq. (2.24) gives a good order of magnitude for n, whose typical values are about 1.0002. It is however too simplified to allow a precise quantitative computation of n. Other approaches still quasi-classical as the Lorentz and Debye ones (cf. [14, 45]) are necessary at this purpose. Nevertheless the linear dependence of n on η we have derived, finds experimental confirmations, and essentially this is what will be used in this work. In addition, the derivation we have performed is meaningful for evaluating the approximations included in the geometrical optics assumptions. In particular geometrical optics reduces a 3D problem to a 1D one, but Eq. (2.19) and Eq. (2.24) say that the 1D approximation along the ray path direction z, implies averaging the medium parameters on xy surfaces of finite dimension, that at GNSS frequencies are about $0.2\, z\ m^2$, for z measured in meters. This poses a theoretical limit to the possibility of resolving horizontal atmospheric structures, that at the top of the troposphere is around $50\ m$.

In the atmosphere we can assume valid the ideal gas approximation, thus we have $\eta \propto P/T$ and as a consequence, for a neutral non polarised gas, we find $n \propto P/T$. There are other two relevant properties of n in the troposphere (and stratosphere too) specifically related to radiation at GNSS frequencies:

1. n is real (i.e. the imaginary part is essentially negligible);

2. n does not depend on the radiation frequency.

If we want to explain these properties still through a quasi-classical scheme, we can say that ω is so far from the resonant frequencies that in n the contribution of the γ damping term vanishes, and n is such that $dn/d\omega \simeq 0$. In other words the neutral atmosphere at GNSS frequencies is essentially not absorbing and non dispersive.

Precise values of n for the neutral atmosphere have been assessed through experimental works. In the present work an empirical relationship is used (cf. [39]) that associates n to the atmospheric state variables:

$$n = 1 + c_1 \frac{P_d}{T} \tag{2.28}$$

with $c_1 = 10^{-6} \cdot (77.604 \pm 0.014)\, K/mb$. A more refined expression is given by the same authors accounting for non ideal-gas behaviours:

$$n = 1 + c_1 \frac{P_d}{T} \cdot (Z_d)^{-1} \tag{2.29}$$

where Z_d is a compressibility coefficient for the real atmospheric gas mixture defined as:

$$(Z_d)^{-1} = 1 + P_d[57.97 \cdot 10^{-8}(1 + 0.52/T) - 9.4611 \cdot 10^{-4} T_c/T^2] \tag{2.30}$$

with T_c given in Celsius degrees.

2.2.2 Interaction with a neutral but polarised gas

.

The neutral but polarised part of the atmospheric gases is essentially what is called wet component of the atmosphere, in other words the water vapour, which is mainly concentrated in the low layers of the troposphere. In § 1.2.3 we have already introduced the polar nature of water, and thus of water vapour. The water molecule forms an angle, with hydrogen atoms at the tips and oxygen at the vertex. Since oxygen has a higher electronegativity than hydrogen, the side of the molecule with the oxygen atom has a partial negative charge. This charge difference forms a dipole with the moment pointing towards the oxygen (Fig. 2.5), that ends partially negative, while the hydrogen ends partially positive.

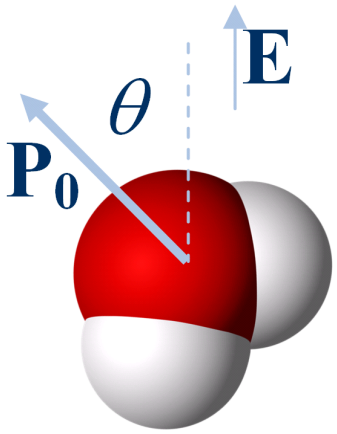


Figure 2.5: Dipole of the water vapour molecule in an electric field.

Under the effect of an electric field, a polar molecule is forced to orientate its moment parallel to the field. This phenomenon adds to the polarisation due to the charge centre displacement, described in § 2.2.1. The refractive index n for a polar gas is consequently expected to contain a part depending on P/T as for non polar gases, plus something else, depending on the moment reorientation due to $\mathbf{E}$.

A comprehensive analysis of such phenomenon is very complex, and outside of the scope of the present work. However it simplifies a lot if we consider a slow changing electric field. Slow means that the system is in equilibrium at any time. Departure from this assumption brings relevant changes in the final expression that can be obtained for n. Nevertheless the qualitative dependence on the gas

parameters remains explained even with this strong hypothesis, while for the final expression of n we will refer once again to experimental measurements.

If we immerse a gas of particles owing an electric moment $\mathbf{p_0}$ in an electric field $\mathbf{E}$, each particle will experiment a mechanical moment $\vec{\mathcal{M}}$ as:

$$\vec{\mathcal{M}} = \mathbf{p_0} \wedge \mathbf{E} \tag{2.31}$$

that tends to align the particle electric moment parallel to the field, against thermal energy that tends to randomise particle motions and thus orientations. After a time interval τ (we suppose infinitesimal) we will reach an equilibrium between these forcing terms. At the equilibrium the spatial distribution of $\mathbf{p_0}$ will form an average angle θ (see Fig. 2.5) such as we could write:

$$\overline{p} = p_0 \, \overline{\cos\theta} \tag{2.32}$$

where $p_0 = \|\mathbf{p_0}\|$ and overlined quantities are averaged on small gas volumes but containing a large number of particles (in thermodynamic equilibrium). Namely $\overline{p}$ is the module of the electric moment of the small gas volume, computed as the mean of the single molecular electric moment vectors.

We can imagine θ satisfying the following generic limits:

$$\lim_{T\to 0} \overline{\cos\theta} \;=\; \lim_{\mathbf{E}\to\infty} \overline{\cos\theta} \;=\; 1 \tag{2.33}$$

$$\lim_{T\to\infty} \overline{\cos\theta} \;=\; \lim_{\mathbf{E}\to 0} \overline{\cos\theta} \;=\; 0 \tag{2.34}$$

In other words we expect to have all particles oriented as $\mathbf{E}$ when the thermal motions becomes negligible with respect to the forcing effect due to the electric field, that is for temperatures T (expressed in K) approaching to 0 or very intense $\mathbf{E}$. The opposite happens when the thermal motions dominate, bringing a perfect stochastic orientation of $\mathbf{p_0}$, i.e. for very high T or negligible $\mathbf{E}$. Thermodynamics gives a way to solve the dependence of $\overline{\cos\theta}$ on T and $\mathbf{E}$. In fact at the equilibrium we know that the number of particles with potential energy $\mathcal{E}$ is distributed proportional to $\exp(-\mathcal{E}/kT)$, wth k Boltzmann constant. The potential energy of the dipole is:

$$\mathcal{E} = -\mathbf{p_0} \cdot \mathbf{E} = -p_0 \, E \cos\theta \tag{2.35}$$

As a consequence the probability for θ will be:

$$\mathcal{P}(\theta) \propto e^{\frac{p_0 E}{kT}} \cos\theta \tag{2.36}$$

and thus:

$$\overline{\cos\theta} = \frac{\int\int \cos\theta\, \mathcal{P}(\theta) \sin\theta\; d\theta\, d\phi}{\int\int \mathcal{P}(\theta) \sin\theta\; d\theta\, d\phi} \tag{2.37}$$

where the integration must be made over all directions, with $0 \leq \phi < 2\pi$ and $0 \leq \theta < \pi$. Solving Eq. (2.37) we obtain the Langevin function, $L(a)$:

$$\overline{\cos\theta} = L(a) = \frac{e^a + e^{-a}}{e^a - e^{-a}} - \frac{1}{2} = \coth a - \frac{1}{2} \tag{2.38}$$

being $a = \frac{p_0 E}{kT}$.

For very small values of a, as in our case, namely for $0 < a \ll 1$, $L(a)$ can be developed and at the first order we have $L(a) \simeq a/3$. Thus Eq. (2.38) finally gives:

$$\overline{p} = p_0 \overline{\cos\theta} = p_0 L(a) \simeq \frac{p_0^2}{3kT} E \tag{2.39}$$

Eq. (2.39) says that our gas particles behaves as having a dipole moment on average proportional to the inverse of temperature, in addition to the induced dipole moment as for the non polarised particles. What was done to obtain an expression for n for neutral non polarised particles can be applied to the dipole expression of Eq. (2.39). Thus for neutral and polarised ideal gases (whose density is proportional to P/T) we expect to have:

$$n = 1 + \overbrace{\mathcal{A} \cdot \frac{P}{T}}^{\textit{induced polarisation}} + \overbrace{\mathcal{B} \cdot \frac{P}{T^2}}^{\textit{polarisation due to orientation}} \tag{2.40}$$

Of course, Eq. (2.40) holds also for non polarised gases; in this case we have $\mathcal{B} = 0$. More generally, for a mixture of neutral gases, we will have:

$$n = 1 + \sum_m \mathcal{A}_m \cdot \frac{P_m}{T} + \sum_p \left(\mathcal{A}_p + \frac{\mathcal{B}_p}{T} \right) \cdot \frac{P_p}{T} \tag{2.41}$$

where the index m maps specific constants and partial pressures for the neutral gases and p for the polarised ones.

We have to stress that our derivation on how n depends on gas thermodynamic parameters is just valid on a qualitative point of view (thus we will not report quantification of n based on our modelling of polarised gases). Assumptions made for a polarised gas in an electric field are stronger than what done for the non polarised one. In fact our derivation assumed static $\mathbf{E}$ or eventually varying $\mathbf{E}$ but as slowly as necessary to have a negligible time lag for the dipole reorientation (i.e. instantaneous dipole reorientation). For electromagnetic waves thus we should expect to find in n, namely in the $\mathcal{B}_p$ coefficients, some dependencies on the wave frequency and on the molecule inertia. A complete analysis of the problem, even through a quasi-classical approach (see for instance [45]), would be too wide for the aim of the present work. In addition it is not necessary, as we will refer to sound experimental results (c.f. [39]), whose precisions is about 0.02% (cf. [13]), that for a generic atmospheric gas composition (including water vapour) give:

$$n = 1 + c_1 \frac{P_d}{T} \cdot (Z_d)^{-1} + \left[c_2 \cdot (Z_d)^{-1} + \frac{c_3}{T} \cdot (Z_w)^{-1} \right] \frac{e}{T} \tag{2.42}$$

with $P_d = P - e$, dry pressure, being P the total pressure and e the partial pressure of water vapour[2]. Constants are:

$$\begin{cases} c_1 & = 10^{-6} \cdot (77.604 \pm 0.014)\, K/mb \\ c_2 & = 10^{-6} \cdot (64.79 \pm 0.08)\, K/mb \\ c_3 & = 10^{-6} \cdot (3776000 \pm 4000)\, K^2/mb \end{cases} \tag{2.43}$$

Finally the compressibility coefficients (to account for non ideal behaviours of atmospheric gases) are the following, even if they give very fine corrections of a few parts per thousands (see Fig. 2.6), that are not so relevant for the aim of our work.

$$\begin{cases} (Z_d)^{-1} & = 1 + P_d[57.97 \cdot 10^{-8}(1 + 0.52/T) - 9.4611 \cdot 10^{-4} T_c/T^2] \\ (Z_w)^{-1} & = 1 + 1650(e/T^3)[1 - 0.01317\, T_c + 1.75 \cdot 10^{-4} {T_c}^2 + 1.44 \cdot 10^{-6} {T_c}^3] \end{cases} \tag{2.44}$$

T_c is the temperature expressed in Celsius degrees.

[2] In § 1 we have used e_w for the water vapour pressure because of non-ambiguity reasons. In the following however we won't have such problem, so we will used simply e for the partial pressure of water vapour.

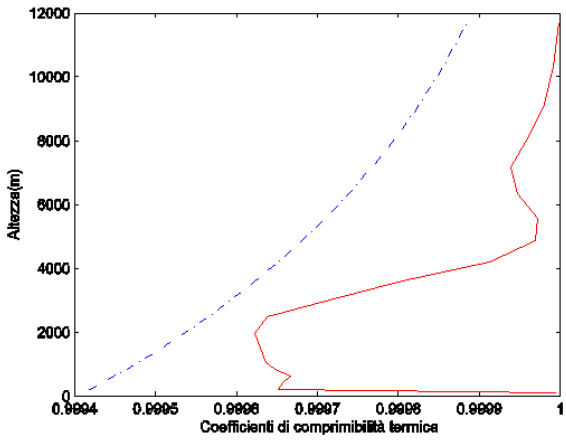


Figure 2.6: Values of the compressibility factors Z_w (solid line) and Z_d (dashed line) at different heights (after [30]).

2.2.3 Interaction with a ionised gas

The ionised part of the atmospheric gases is essentially the tenuous plasma characterising the ionosphere. Here free charges, in particular electrons that are the lightest ones, interacts with electromagnetic waves, according to the Maxwell equations, that we have to rewrite with respect to Eq. (2.1) for having free charges and relative currents.

The general form of the Maxwell equation in a medium is:

$$\begin{cases} \boldsymbol{\nabla} \cdot \mathbf{D} = \rho \\ \boldsymbol{\nabla} \cdot \mathbf{H} = 0 \\ \boldsymbol{\nabla} \wedge \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \\ \boldsymbol{\nabla} \wedge \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \end{cases} \tag{2.45}$$

with

$$\begin{cases} \mathbf{D} = \epsilon \mathbf{E} \\ \mathbf{B} = \mu \mathbf{H} \end{cases} \tag{2.46}$$

and with ρ electric charge density, $\mathbf{J}$ electric current density, ϵ dielectric constant and μ magnetic permittivity of the medium[3]. The electromagnetic wave propagation velocity is now given by $v = 1/\sqrt{\epsilon\mu}$ equal to c only in the vacuum (when $\epsilon = .\epsilon_0$ and $\mu = \mu_0$).

[3]In the general case ϵ and μ are tensors, in order to account for medium anisotropies.

Eq. (2.45) and (2.45) allow to treat electromagnetic field propagation in a medium knowing its macroscopic quantities. What we want to do is to analyse the problem also from a point of view closer to the medium microscopic properties.

The ionosphere is a very rarefied gas, whose major part is neutral and non polar, giving a negligible effect on GNSS signals, according to Eq. (2.29). On the contrary the effects of the charged part are to be quantified. Thus, for our scopes, we can approximate ionosphere as an extremely rarefied gas of charged particles, with a local charge balance between positive ions and electrons (i.e. $\rho = 0$), and also thermodynamic equilibrium among charged particle populations (i.e. with a unique defined temperature)[4].

Eq. (2.45) consequently simplifies in:

$$\begin{cases} \boldsymbol{\nabla} \cdot \mathbf{E} = 0 \\ \boldsymbol{\nabla} \cdot \mu\,\mathbf{H} = 0 \\ \boldsymbol{\nabla} \wedge \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \\ \boldsymbol{\nabla} \wedge \mathbf{H} = \sigma\mathbf{E} + \frac{\mu}{c^2}\frac{\partial \mathbf{E}}{\partial t} \end{cases} \tag{2.47}$$

where we ha used:

$$\mathbf{J} = \sigma\mathbf{E} \tag{2.48}$$

which is the Ohm law, with σ electric conductivity.

Solutions for Eq. (2.47) can be given in terms of transverse and longitudinal field components, with z propagation direction (c.f. [22]):

$$\begin{cases} \mathbf{E}(\mathbf{z},\mathbf{t}) = \mathbf{E_{tr}}(\mathbf{z},\mathbf{t}) + \mathbf{E_{long}}(\mathbf{z},\mathbf{t}) \\ \mathbf{H}(\mathbf{z},\mathbf{t}) = \mathbf{H_{tr}}(\mathbf{z},\mathbf{t}) + \mathbf{H_{long}}(\mathbf{z},\mathbf{t}) \end{cases} \tag{2.49}$$

In the ionosphere longitudinal components result negligible. In fact they give a static uniform magnetic field and a variable but dumped electric field $E_{z,t} = E_0 e^{-4\pi\sigma t/\epsilon}$, that for low conductivities gives negligible contributions.

[4]This last assumption will be not explicitly recalled in the following of this section, but it is one of the properties that is implicitly included in the further assumption of negligible contribution to the issue from positive ions with respect to electrons.

On the contrary transverse components give:

$$\begin{cases} \mathbf{H} = \frac{c}{\mu\omega}(\mathbf{k} \wedge \mathbf{E}) \\ \\ i(\mathbf{k} \wedge \mathbf{H}) + i\frac{\epsilon}{\omega}\mathbf{E} - \frac{4\pi\sigma}{c}\mathbf{E} = 0 \end{cases} \tag{2.50}$$

From Eq. (2.50) we obtain the dispersion relation for the propagation vector $\mathbf{k} = k\epsilon_{\mathbf{3}}$

$$k^2 = \mu\epsilon\frac{\omega^2}{c^2}\left(1 + i\frac{4\pi\sigma}{\omega\epsilon}\right) \tag{2.51}$$

Thus $k \in \mathbb{C}$ is a complex number, generating a damping term for the transverse field with a phase lag between the electric and magnetic components. In fact if we write $k = \alpha + i\beta$ and search for plane wave solutions we obtain:

$$\begin{cases} \mathbf{E}(\mathbf{x}, t) = \mathbf{E_0} e^{-\beta \mathbf{k}\epsilon_{\mathbf{3}}} e^{i\omega t - \mathbf{kx}} \\ \\ \mathbf{H}(\mathbf{x}, t) = \mathbf{H_0} e^{-\beta \mathbf{k}\epsilon_{\mathbf{3}}} e^{i\omega t - \mathbf{kx}} \end{cases} \tag{2.52}$$

with the amplitude reducing of a factor $1/e$ over a distance $\delta = 1/\beta$, with:

$$\frac{1}{\beta} \simeq c^2 \cdot \sqrt{\frac{\epsilon}{2\pi\omega\sigma}} \tag{2.53}$$

δ is (*skin depth*), that is so large in the ionosphere that signal attenuation is negligible..

The phase relation can be derived starting from the first equation of (2.52) (cf. [22]):

$$\frac{\mathbf{H_0}}{\mathbf{E_0}} = \sqrt{\frac{\epsilon}{\mu}}\left[1 + \left(\frac{4\pi\sigma}{\omega\epsilon}\right)^2\right]^{1/4} \tag{2.54}$$

In order to analyse the consequences of (2.54) in the ionosphere, we will consider the following simplified model.

Let us assume the equation of motion of a free charge in the ionosphere due to an electric field $E(\mathbf{x}, t) = E_0 e^{(i\omega t - \mathbf{kx})}$ (and negligible magnetic fields) to be:

$$m\dot{\mathbf{v}}(t) + m\gamma\mathbf{v}(t) = q\,\mathbf{E} \tag{2.55}$$

A stationary solution for the velocity $\mathbf{v}$ is:

$$\mathbf{v} = \frac{q}{m(\gamma - i\omega)}E(\mathbf{x}, t) \tag{2.56}$$

If n_0 is the density of charges (e.g. el/m^3), we obtain:

$$\sigma = \frac{n_0\, q^2}{m(\gamma - i\omega)} \tag{2.57}$$

For a very low density gas as the ionosphere we have $\sigma_p \simeq i\,\frac{n_0\, q^2}{m\omega}$, and the phase lag is essentially null.

As regards the dispersion relation in Eq. (2.51) we can introduce the plasma frequency $\omega_p = \frac{n_0\, q^2}{m}$. The refractive index consequently becames:

$$n^2 = 1 - \left(\frac{\omega_p}{\omega}\right)^2 \tag{2.58}$$

Electron densities in the ionosphere generally make $\omega_p < 20\, MHz$ (cf. [31]), and thus, at GNSS frequencies, we can simplify Eq. (2.58) in:

$$n = 1 - \frac{1}{2}\left(\frac{\omega_p}{\omega}\right)^2 \tag{2.59}$$

The approximation we have derived for $n = n(\omega)$ is very precise, but a more fine modelling for n in the ionosphere is the Appleton-Hartree formula, that includes geomagnetic effects on free moving charges (i.e. medium anisotropies) and a finite mean free path for electrons (i.e. a non null electron collision frequency). Such formula can be expressed as follows (cf. [31]):

$$n^2 = 1 - \frac{X}{W + X - \frac{Y^2 \sin^2\theta}{2W} \pm \frac{1}{W}\left(\frac{1}{4}Y^4 \sin^4\theta + Y^2 \cos^2\theta\, W^2\right)^{\frac{1}{2}}} \tag{2.60}$$

with:

$$\begin{cases} X &= \frac{f_p}{f} \\ Y &= \frac{f_H}{f} \\ Z &= \frac{\nu}{2\pi f} \\ W &= 1 - X - iZ \\ f_p &= \sqrt{\frac{q^2 n_0}{4\pi^2 \epsilon_0 m}} \qquad \text{plasma frequency} \\ f_H &= \frac{\mu_0 q H_0}{2\pi m} \qquad \text{gyro frequency} \end{cases}$$

In Eq. (2.60) the presence of the ± sign gives different solutions for the complex refractive index, depending on wave modes[5], and the different terms are:

$$\left[\begin{array}{lcl} \nu & = & \text{electron collision frequency} \\ f & = & \text{wave frequency} \\ \epsilon_0 & = & \text{permittivity of free space} \\ \mu_0 & = & \text{permeability of free space} \\ H_0 & = & \text{ambient geomagnetic field strength} \\ q & = & \text{electron charge} \\ m & = & \text{electron mass} \\ n_0 & = & \text{electron density} \\ \theta & = & \text{angle between ambient magnetic field vector and the wave one} \end{array}\right.$$

In our work however we will make use of the simplified formula for n we have previously derived (2.59), which is equivalent to Eq. (2.60), assuming $f_H \ll f$ and $\nu \ll f$ (i.e. X and Z negligible), then developing n at the first order in X, being $0 < X \ll 1$.

Nevertheless we can still use Eq. (2.60) to evaluate the accuracy of the approximations we have derivedF for n, depending on the varying ionospheric characteristics.

2.3 Signal delay

2.3.1 Delay components

We have already seen how the real part of n can be interpreted as a change in the electromagnetic wave velocity, for instance from that in vacuum (c), to that in the medium (v), writing $n = c/v$. When far from the resonant lines we have $n \simeq \mathrm{Re}(n) \geq 1$, that results in a delay in the wave propagation. An additional

[5]In an unmagnetised plasma an electromagnetic wave behaves simply as a light wave modified by the plasma medium. In a magnetised plasma on the contrary the situation is different and we can have two wave modes perpendicular to the field, the O and X modes, and two wave modes parallel to the field, the R and L ones. For propagation perpendicular to the magnetic field ($k \perp H_0$), the '+' sign is due to the "ordinary" mode and the '-' sign due to the "extraordinary" one. For propagation parallel to the magnetic field ($k \parallel H_0$), the '+' sign is due to a left-hand circularly polarized mode, and the '-' sign due to a right-hand circularly polarized mode.

delay originates when the ray path is not a straight line when locally n has a gradient perpendicular to the propagation direction. In geometrical optics we are used to visualise the problem as a electromagnetic ray intercepting a surface of discontinuity between two media with finite difference in n: if the ray path direction $\mathbf{k}$ is not orthogonal to such surface, we see an abrupt change in the propagation direction, more remarkable for increasing incidence angles and Δn, according to Snell law (see § 2.3.2). For smooth changes of n normal to $\mathbf{k}$, the result is a curved path, the so called *bending* phenomenon.

The final delay of a signal travelling through a given medium with respect to the path G in vacuum conditions, is thus generally given by two components, one due to the changed velocity $v = c/n$ and the other due to the changed path S, and can be written as:

$$\begin{aligned} \delta t &= \int_S \frac{1}{v}\, dl - \int_G \frac{1}{c}\, dl \\ &= \int_G \frac{n}{c}\, dl + \int_{S-G} \frac{n}{c}\, dl - \int_G \frac{1}{c}\, dl \\ &= \frac{1}{c} \int_G n - 1 \; dl + \int_{S-G} \frac{n}{c}\, dl \end{aligned} \tag{2.61}$$

δt is what we can measure if we precisely know the source time and the receiving time of a signal, and this is exactly the main thing that GNSS are born for.

Values of δt are so small in the atmosphere that generally we prefer to express the delay as an equivalent distance $\delta l = c \cdot \delta t$, that is as a lengthening of the signal path, as if were travelling at the same speed as in vacuum conditions. This practice sometimes generates a misunderstanding, as it is interpreted as essentially due to an increase path length, such as $S - G$, i.e. due to bending. On the contrary we will see in § 2.3.3 that this term is generally negligible, except for signal directions very low on the horizon[6].

[6]Signal low on the horizon are generally very noisy, mainly caused by multipath effects that artificially augment the signal delays, due to multiple reflections of signals on surface structures before reaching the receiver. As a consequence the error in the data processing rapidly increase, thus generally such data are discharged when not directly filtered out by the receiving hardware apparatus (see § 3.4.9).

We can consider *negligible* a contribution to δl at GNSS frequencies, when smaller than the expected minimum error in computing the wet delay, i.e. a few millimetres. This is always verified for the bending component in the troposphere and more generally in the non-charged atmosphere, while it is not necessarily true in the ionosphere. However the ionospheric contribution is computed and then subtracted from the total delay in a way that again minimise the bending effects. Let us see how.

The refractive index of Eq. (2.59) gives clearly a value of $n < 1$. This brings to the seeming contradiction with a basic principle of relativity, having a signal travelling with a superluminal velocity $v > c$. In this case however there is a difference between the phase velocity, $v_p = v > c$ and the group velocity of the wave packet $v_g < c$, due to the dispersive nature of the medium, and it is v_g the signal velocity, because information is transferred by wave packets, thus always at subluminal speed.

Reminding that $\omega = 2\pi f$ and $k = \frac{2\pi}{\lambda}$, we have:

$$v_p = \frac{\omega}{k} = \frac{c}{n} = \frac{1}{\sqrt{\epsilon\mu}} \tag{2.62}$$

$$v_g = \frac{\partial\omega}{\partial k} = \frac{c}{n} - \frac{kc}{n^2}\frac{\partial n}{\partial k} \tag{2.63}$$

Hence:

$$\begin{aligned} v_g &= v_p + \omega\frac{\partial v_p}{\partial\omega} \\ &= v_p + f\frac{\partial v_p}{\partial f} && (2.64) \\ &= v_p\left(1 - \frac{k}{n}\frac{\partial n}{\partial k}\right) && (2.65) \end{aligned}$$

We can introduce two useful indexes n_g and n_p defined as:

$$v_p\, n_p = c \tag{2.66}$$

$$v_g\, n_g = c \tag{2.67}$$

Using Eq. (2.63) we can then write:

$$n_g = \frac{n_p^2}{n_p - f\frac{\partial n_p}{\partial f}} = n_p^2\,\frac{1}{n_p}\,\frac{1}{1 - \frac{f}{n_p}\frac{\partial n_p}{\partial f}} \tag{2.68}$$

Assuming $\frac{f}{n_p}\frac{\partial n_p}{\partial f} \ll 1$ we have an expression that can be developed with Taylor, of the type $(1-x)^{-1} \simeq 1+x$, and thus we obtain:

$$n_g = n_p + f\frac{\partial n_p}{\partial f} \tag{2.69}$$

Introducing Eq. (2.59), with f instead of ω, we finally have:

$$n_p = 1 - \frac{n_0\,\alpha}{f^2} \tag{2.70}$$

$$n_g = 1 + \frac{n_0\,\alpha}{f^2} \tag{2.71}$$

in which $\alpha = \frac{e^2}{8\pi^2\epsilon_0 m}$. From Eq. (2.66) and Eq. (2.67) we can verify:

$$v_p = \frac{c}{n_p} \geq c \tag{2.72}$$

$$v_g = \frac{c}{n_g} \leq c \tag{2.73}$$

As a consequence for the delays in the ionosphere we have a phase advance and a group delay.

Commonly when dealing with the ionosphere, the TEC (Total Electron Content) parameter is introduced, which is defined as the columnar content of free electrons (i.e. number of electrons for unit surface):

$$\mathrm{TEC} = \int_L n_0 dl \tag{2.74}$$

Using Eq. (2.74) we can finally write:

$$\textit{Phase advance} = \frac{1}{c}\int_L (1-n_p)dl = \frac{1}{c}\int_L \frac{n_0\,\alpha}{f^2}dl = \frac{\alpha}{f^2}\,\mathrm{TEC} - \beta_f \tag{2.75}$$

$$\textit{Group delay} = \frac{1}{c}\int_L (n_g-1)dl = \frac{1}{c}\int_L \frac{n_0\,\alpha}{f^2}dl = \frac{\alpha}{f^2}\,\mathrm{TEC} + \beta_f \tag{2.76}$$

We can observe that the phase advance and the group delay are equivalent and they depend only on the TEC, except for β_f, the bending contribution at frequency f for a given signal path.

The availability of at least two different frequencies in the GNSS systems aims at computing this delay due to the ionospheric dispersive medium. In fact,

with two frequencies, the difference between the two delays can be obtained with great precision, as we will see in § 3. As the two signals origin at the same point, their paths are very close except for the difference due to f. Generally the bending difference $\beta_2 - \beta_1$ is negligible, even if β_f is not for observation directions different from the receiver zenith. If the differential bending is negligible and if we accurately know the source position, we can directly retrieve n from the delay measurements, assuming a straight path between the source and the receiver.

In the following section however we will quantify typical bending contributions, in order to assess the range of validity of such assumptions.

2.3.2 Bending

A way we can understand the deviation of ray paths due to varying n is to reanalyse the stationary phase approximation in Eq. (2.19). Stationary points cancel the first derivatives of the phase term, and in our case they lay along the straight $\mathbf{k}$ direction. This approximation is rigorous if the g function is constant or at least isotropic with respect to the stationary points. If not, we can imagine to develop g in Fourier series, whose components will be not all isotropic with respect to the stationary points (accordingly to g). This will introduce a slight phase displacement that can be transferred from g to f through the Fourier modes. The result will be a displacement of the stationary phase points due to the anisotropies of g. In our problem the anisotropies could be brought by the shape of η normal to $\mathbf{k}$.

In geometrical optics the deviation is assessed by the Fermat's principle. It is also known as the principle of least time as it says that the path taken between two points by a ray of light is the path that can be traversed in the least time. A modern definition states that the optical path length, S, must be stationary, which means that it can be either minimal, maximal or a point of inflection (a saddle point), being S defined as follows:

$$S = \int_L n(l)\, dl \tag{2.77}$$

From Eq. (2.61) we know that the associated phase delay is:

$$\delta\phi = k \int_L (n(l) - 1)\ dl \tag{2.78}$$

In this way, through the Eulero-Lagrange mathematical formulation we can determine the path from $n = n(x, y, z)$. More easily we can obtain a solution for Eq. (2.77) through geometrical analyses, assuming for the atmosphere a quasi plane parallel structure (that is equivalent to assume for n a pure radial dependence). Such assumption is not too strong at GNSS frequencies to prevent the comprehension of typical bending behaviours.

With reference to Fig. 2.7, being $\alpha(r)$ the local angle between the ray vector $\mathbf{r}$ and the tangent to the signal path, the Snell law for a medium with $n = n(r)$ is (cf. [7]):

$$n(r) \cdot r \sin \alpha(r) = a_0 \tag{2.79}$$

where a_0 is a constant to be determined knowing n and incidence angle $(\frac{\pi}{2} - \alpha)$ for a given $r = r_0$. As an example in Fig. 2.7 we have $r_0 = R_{max}$ and $\alpha(r_0) = \alpha_2$, that are values at the top of the troposphere[7].

Eq. (2.61) can be interpreted as the difference of the optical paths $S - S_0$, being S_0 the geometric distance between P_2 and P_1, or, in other words, the path from tropopause to receiver in the vacuum conditions.

For the computation of S it is easier to proceed in polar coordinates. We see in Fig. 2.8 that $dl = \sqrt{dr^2 + r d\theta^2}$ and $dr = dl\ \cos\alpha$, and thanks to the Snell law we can write:

$$dl = \frac{r\, n(r)}{\sqrt{r^2\, n(r)^2 - {a_0}^2}} \tag{2.80}$$

and consequently:

$$\delta\phi = \frac{2\pi}{\lambda} \int_{R_T}^{R} \frac{r\, n(r)^2}{\sqrt{r^2\, n(r)^2 - {a_0}^2}}\, dr \tag{2.81}$$

[7]We refer only to the troposphere as it is largely the main source of signal delay effects (bending included) due to neutral gases in the atmosphere.

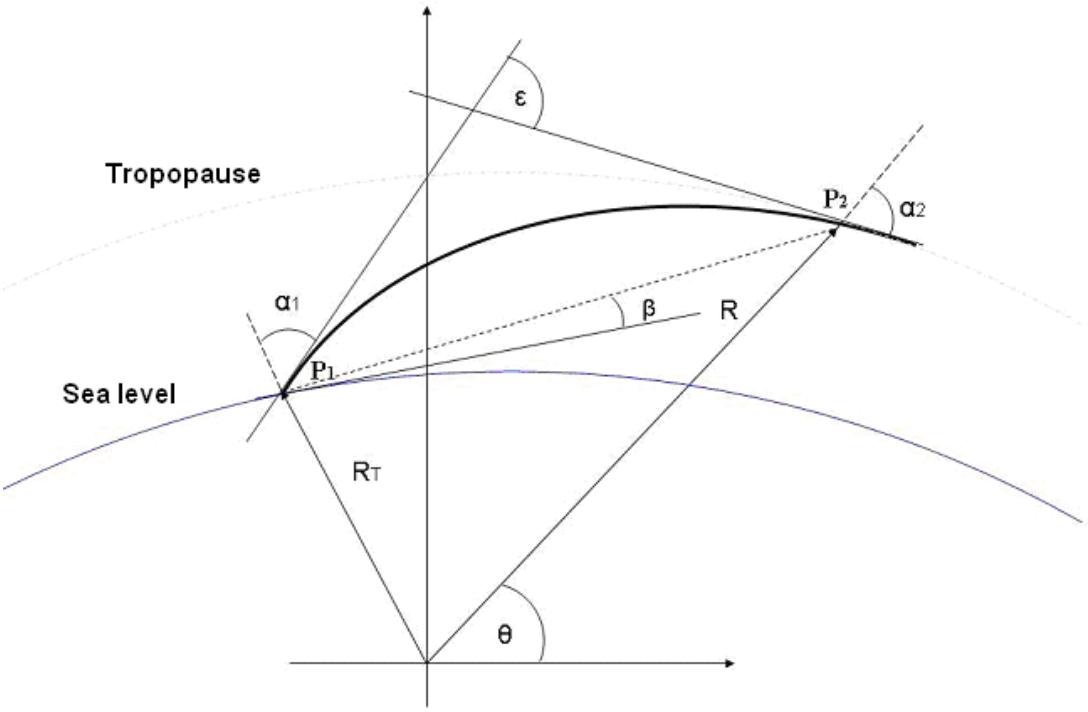


Figure 2.7: Bending geometry for a plane parallel atmosphere.

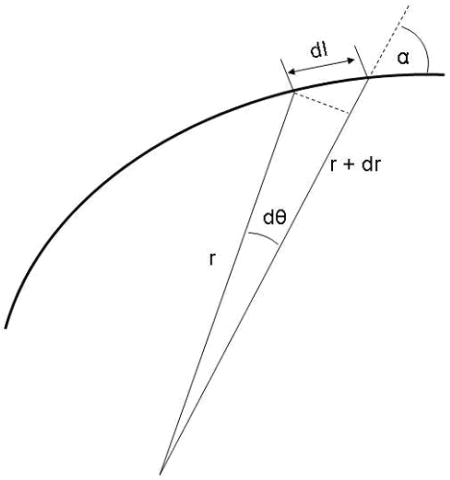


Figure 2.8: Bending geometry in polar coordinates.

The integration of Eq. (2.80) gives the total length of the path travelled by the signal, while Eq. (2.81) gives the associated phase variation.

The relationship between the θ and α angles is:

$$\theta = \alpha_1 - \alpha_2 + \epsilon \tag{2.82}$$

where:

$$\epsilon = \int_{P_1}^{P_2} dl/\rho = \int_{x_1}^{x_2} \frac{dn/dx}{n(x)\sqrt{x^2 - a_0^2}} \tag{2.83}$$

with dl given by Eq. (2.80), ρ local curvature radius and $x = rn(r)$ (cf. [28]).

We can observe that the second part of Eq. (2.83) is correct only if ϵ monotonically grows with l, which is always true except if ρ changes its sign (i.e. the path changes its convexity). This can happen only for path close to the horizon and under some specific conditions. However we have preferred to find a more general alternative, and computing θ as follows:

$$\theta = \int_{P_1}^{P_2} a_0 \frac{dl}{n(r) \cdot r^2} = a_0 \int_{R}^{R_T} \frac{1}{r \cdot \sqrt{n(r)^2 r^2 - a_0^2}}\, dr \tag{2.84}$$

Finally we give an expression for the β angle in Fig. 2.7:

$$\beta = \arctan\left(\frac{R\cos\theta - R_T}{R\sin\theta}\right) \tag{2.85}$$

2.3.3 Simulations

Equations in § 2.3.1 and in § 2.3.2 have been implemented in a simulator in order to compute path delays and trajectories, for given atmospheric profiles[8]. Through the simulator we have evaluated variations of the order of fractions of millimetres in ray paths of the order of tens of kilometres or more. Thus we have had to pay particular attention to the way we have implemented the equations and to the tolerance values for integrals, in order to avoid artificial results due to numeric truncations.

[8]The same simulator will be used to generate synthetic GNSS delays in § 4, to perform numerical experiments

Here we present only simulations performed with standard atmospheric values in order to fix some typical behaviours of the GNSS signal useful for the following of the present work. Namely we have adopted:

- for the ionosphere, diurnal and nocturnal standard profiles from [5] that is a sort of ionospheric equivalent of the standard atmosphere;
- for the wet neutral atmosphere, the standard atmosphere as described in § 1.2.2 and specifically through equations (1.14) and (1.15);
- for the wet atmosphere, an "extremely wet" standard tropospheric profile, i.e. a standard profile that have been saturated in the troposphere.

Fig. 2.9 in the top panel shows a typical range of zenith delay values, where the extremes are given by a lowest delay profile, due to a ray path crossing first a ionosphere with a minimum (i.e. nocturnal) TEC and then a completely dry atmosphere, and a highest delay, due to a ray path crossing first a ionosphere with a maximum (i.e. diurnal) TEC and then a standard profile but saturated in the troposphere. We can see how the main and more variable contribution comes from the ionosphere, leading to zenith delays from 5 to over $30\,m$. This is the reason why an accurate retrieval of such term, through a multi-frequencies signal processing (as introduced in § 2.3.1and that will be completed in § 3.4.2), is mandatory for precise positioning as well as for tropospheric sounding. The tropospheric delays are "only" around $2.3 \div 2.5\,m$ and from the bottom panel of Fig. 2.10 we can see how water vapour can contribute for about a 10% of such delays at most. From this we understand how, for tropospheric water vapour retrieval, it is necessary to process tropospheric delays with an accuracy of the order or better than the cm, i.e. it is necessary to set up a GNSS pre-processing chain according to such constrain.

Non zenith ray paths lead to greater delays according to the greater thickness of the atmospheric shell to be crossed. Such delay increments go as $1/\sin\alpha_{el}$, with α_{el} elevation angle (in Fig. 2.7) the complementary of angle α_1), apart from

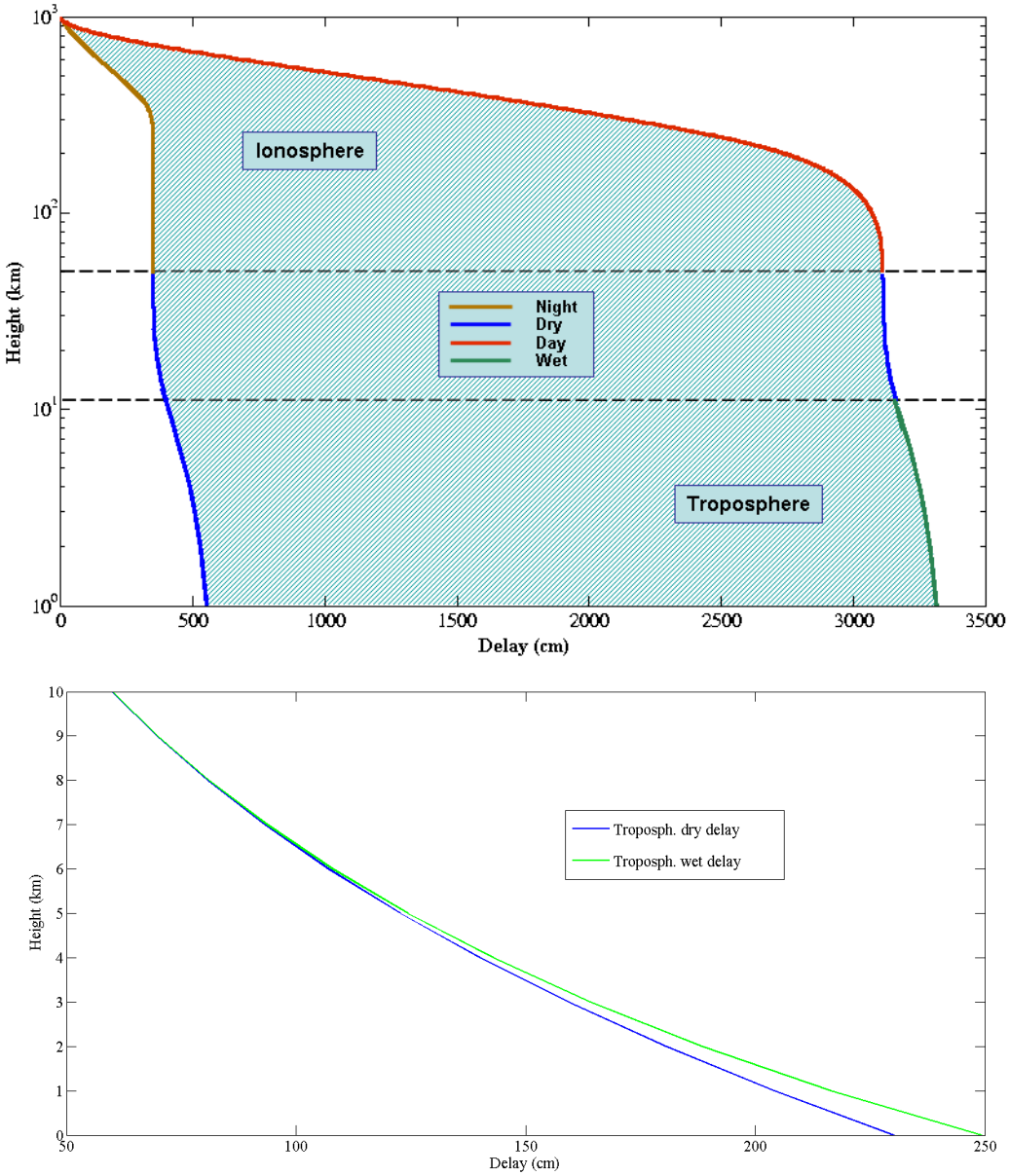


Figure 2.9: Top panel: increment of integrated delays for GNSS (L-band) signals along zenith paths. The lower (higher) delay is obtained for a combination of standard atmospheric conditions producing the lower (higher) refractive indexes, namely nocturnal (diurnal) ionosphere and dry (saturated wet) troposphere. Heights are from the sea level.
Bottom panel: as in the top panel but only for delays due to the neutral atmosphere at tropospheric levels. Note that the vertical scale here is linear.

Earth curvature effects, whose importance increases for small α_{el}. Practically small α_{el} are not of interest for us, as for elevation angles minor than about 20^o (i.e. ~ 0.35 rad) GNSS data (when received) are affected by critical noise effects, primarily due multipath, that normally make them to be rejected (see § 3.4.9). We already know that in addition to increased delays, non zenith paths are subject to bending. In Fig. 2.10 results are shown of a simulation aimed at evaluating ray path characteristics for different elevation angles, bending included. Simulations are limited to the troposphere for a standard (plane parallel) atmosphere. Various panels show the different parameters related to:

- the geometric characteristics of the path, namely the polar angle, the geometrical (i.e. Euclidean) distance, the total length;
- the difference between the total length and the geometrical distance, i.e. the path lengthening due to the bending;
- the optical path, i.e., neglecting the bending, the signal total delay multiplied by the speed of light;
- the difference between the optical path and the geometrical distance, i.e. the signal delay component due to the atmospheric medium.

Apart from paths very low on the horizon (not used in practice), the bending contribute for fractions of millimetres to the total signal delay that is of the orders of metres. In addition in Fig. 2.11 we have reported the vertical distance between a straight path and a curved path, evaluated for different path trajectories. We can observe how the distances are never larger than a few metres, that is several order of magnitude less than what we expect to be able to resolve atmospheric structures (and than what is of meteorological interest). This means that in the course of the present work we will always consider the bending contribution as negligible, approximating the ray trajectories as straight lines.

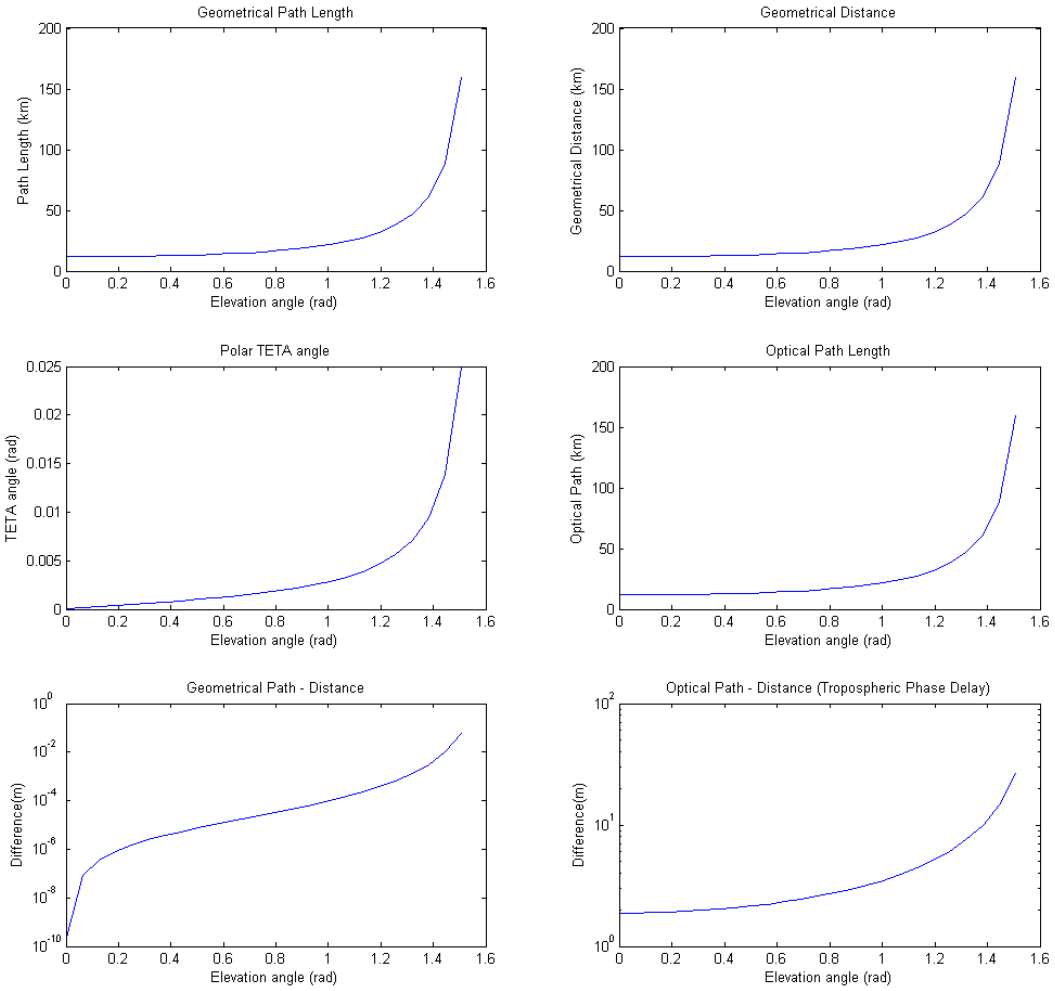


Figure 2.10: Simulation of path parameters for different elevation angles in a standard troposphere.

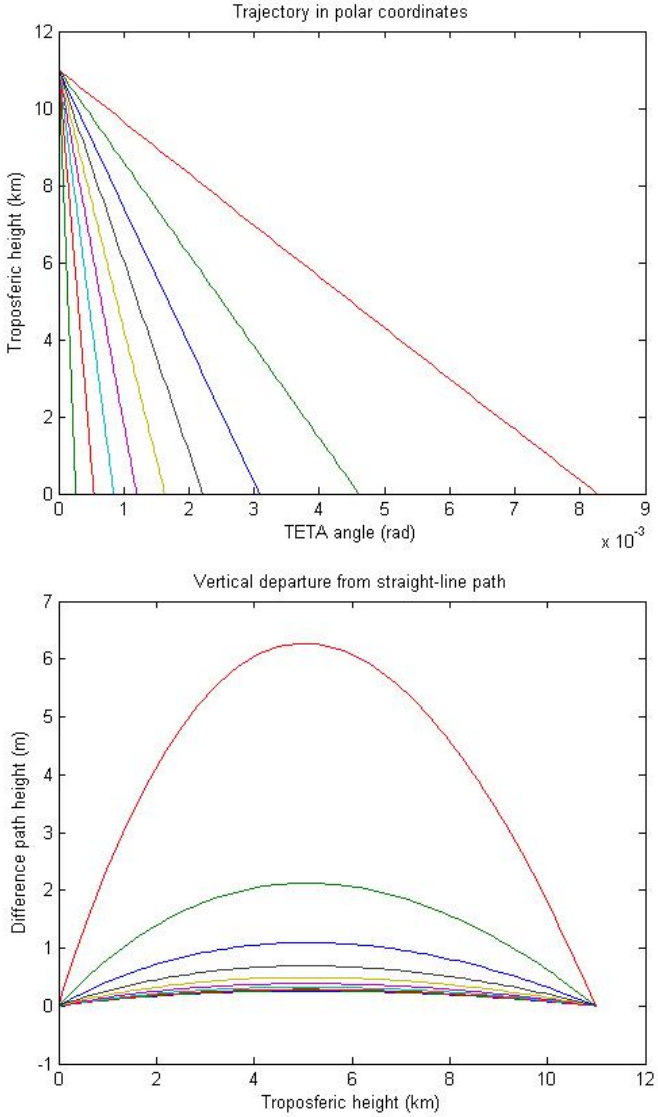


Figure 2.11: Different ray trajectories (top panel) and corresponding vertical distances between straight and curved paths due to the bending (bottom panel) in a standard troposphere.

Chapter 3

The GPS satellite navigation system

3.1 The basis of satellite navigation

Satellite navigation represents the last evolution in navigation, since on early 1960's an initiative of U.S. Navy started experiments to create a system for precise positioning. The system was called Navy Navigation Satellite System (NNSS) or, more commonly, TRANSIT, and was developed by Johns Hopkins Applied Physics Laboratory (APL). The first proof of concept was based on data coming from the first artificial satellite Sputnik I (Russian satellite) and it demonstrated the possibility to track satellite orbits using a Doppler signal coming from a satellite, detected from points of known position on the Earth surface. The inversion of the problem led to the first satellite radio navigation system able to determine the position of an user by knowing the Doppler shift of detected satellite signal and the satellite orbital parameters. The TRANSIT navigation system was developed for military purposes, mainly for submarine navigation. The experience of TRANSIT gave origin to the TIMATION (TIMe/navigATION), program of Naval Research Laboratory's (NRL's) Naval Center for Space Technology (NCST), which designed satellites equipped with accurate oscillators (quartz oscillator in the start project phase and atomic clocks in the final project phase) controlled and systematically re-synchronised by ground master stations. The program started in 1967 and proved that a system based on

a passive ranging technique, combined with highly accurate clocks, could lead to an unprecedented navigation system with three-dimensional coverage (longitude, latitude, and altitude) all over the world. The results of the TIMATION program gave the basis for the development of the first satellite based passive radio navigation system, the NAVSTAR (Navigation System with Time and Ranging) or Global Positioning System (GPS). GPS was the first born of a small number of analogous systems from different countries, all belonging to the family of Global Navigation Satellite Systems (GNSS) whose working principle is essentially the same: the possibility to measure the distance of a receiver from satellites and consequently to analytically solve a system of three equations with the receiver coordinates as unknowns, provided we have (at least) three non aligned satellites in view, that transmit a signal with the "exact" time of transmission and corresponding satellite position, and that are all synchronised with the receiver clock, that gives the exact time of signal arrival[1]. However for cost saving and practical advantages, GPS receivers are not equipped with precise atomic clocks, so they cannot be considered synchronised with the satellites: as a consequence, the receiver time becomes an additional unknown and at least four satellites are needed to analytically solve a system of four equations with four unknowns.

The GPS design started in 1973 and the system became fully operational in 1993. The main services provided by GPS are user position, speed vector determination and the time synchronisation at a global scale. The GPS system has been designed and developed from U.S. military departments for military and civilian users, but the performance of the system could be degraded for reason of security, using the so called "Selective Availability" (SA). Since May 1st 2000 the SA has been turned off, and the full precision of GPS system is available for all GPS users.

The architecture of the system consists of a space segment, a control segment and a users segment.

[1]It is apparent that such a problem admits two solutions, but only one of interest because on the Earth surface, the other being in a point of the exosphere, which results symmetric to the one of interest, with respect to the plane individuated by the three satellites in view.

The space segment is the constellation of satellites; that send a signal with a unique code for each satellite, giving information for time synchronisation and space vehicle position determination.

The Operational Control Segment (OCS) performs the tracking of space vehicles, computing, monitoring and adjusting their positions; the monitoring of space vehicle clock offset and drift to maintain synchronisation within the satellite clocks; the updating of the navigation signal sent by satellites. The control segment is composed by the Master Control Station (MCS) and five Monitor Stations (MS), some of them equipped with Ground Antennas (GA).

The User Segment consists of the final users of the GPS system. The hardware of user receivers detects the GPS satellites signals and gives the position by means of the installed software.

The GPS system is designed to contemporarily provide two different services: Standard Position Service (SPS) and Precise Positioning Service (PPS). At this purpose the system is equipped with two different type of signals sent by satellites, called respectively coarse acquisition (C/A) code and precision (P) code.

The system is designed to ensure the correct positioning in every place on the Earth: this is guaranteed by the possibility to receive simultaneously data from more than four satellite (up to ten in some cases), always and everywhere[2]. This redundancy is essential when the computation is not limited to the four unknowns which constitute the space-time position, but involves other uncertainties, e.g. biases and offsets given by noise sources.

3.2 GPS constellation

3.2.1 Space segment structure

The Space Segment consists on the GPS satellite constellation. At the moment there are 31 operational satellites. For redundancy reasons they are more than the minimum number ensuring the proper system coverage, which is 24. The

[2]Of course this is true except when signal reception is obstructed by local shields.

satellites are arranged on 6 orbital planes, each of them containing at least 4 slots where satellites can be equidistantly set. The orbital planes have an inclination of 55° with respect to the equatorial plane and are rotated in the equatorial plane by 60° against each other. This geometry ensures total global coverage so that at least four satellites are simultaneously visible anytime and anywhere, allowing to take advantage of the positioning service. The GPS orbits are Medium Earth Orbit (MEO) near circular with a radius of about 26560 km, resulting in an eight of 20200 km above the Earth surface. The revolution time is half a sidereal day, precisely corresponding to 11 hours and 58 minutes. The speed of GPS satellites is about 3.9 km/s.

3.2.2 Building up of the GPS constellation

The constellation of first generation (block I) GPS satellites, consisted of 10 satellites with one caesium and two rubidium atomic clocks, that were launched from 1978 to 1985. These satellites were experimental and are unused since years. Block I satellites constituted the GPS Demonstration system and reflected various stages of system development.

The second generation (block II) of 9 satellites constellation was launched from 1989 to 1990, with the main objective of testing 14 days of operations without contact with the Control Segment (CS). It was followed by un upgraded generation of satellites (block IIA) consisting of 19 satellite, launched between 1990 and 1997. These satellites were designed to provide 180 days of operation without contact from the control segment. During this autonomy period, accuracy degradation was verified in the navigation message. The block II and IIA satellites are equipped with two rubidium and two caesium atomic clocks. They have the Selective Availability (SA) and Anti-Spoof (A-S) capabilities.

When the GPS system became operational in 1993 the constellation of 24 satellites was composed by blocks I, II and IIA.

Later, 12 satellites of block IIR where successfully launched between 1997 and 2004. They were called “Replenishment” satellites. An important innovation was

the capability to autonomously navigate themselves (AUTONAV), by creating the 50 Hz navigation. This includes the ability to determine thier own position by performing inter-satellite ranging with other IIR space vehicles.

Following 8 satellites of block IIR-M, (Modernisation) were launched between September 2005 and August 2009. The IIR-M capabilities include developmental military-use-only M-code on the L1 and L2 signals and a civil code on the L2 signal, known as the L2C signal.

Both block IIR and block IIR-M are equipped with three rubidium atomic clocks. Their extreme precision of about 1 second in 1 million years is absolutely necessary for the functioning of the system, as we shall see later.

The last generation of on orbit GPS satellites is the block IIF. The first satellite of this block was launched in May 2010. These satellite are functionally equivalent to the IIR/IIR-M satellites and pave the way towards operational M-code and L2C signal. The improvement introduced by this satellite generation involves converting the GPS caesium clocks from analog to digital. Block IIF also adds a new separate signal for civilian use, designated L5 at the frequency of 1176.45 MHz. These two codes mark the transition to the GPS III era.

3.2.2.1 the future GPS III

GPS III block will give new navigation warfare (NAVWAR) capabilities to shut off GPS service for security reasons, but only on limited geographical locations. The new system will offer significant improvements in navigation capabilities by new navigation signals:

- The M-code, signal transmitted in the same L1 and L2 frequencies already in use by the previous military code, the P(Y) code. It is designed to be autonomous, meaning that users can calculate their positions using only the M-code signal. It will be broadcast from a high-gain directional antenna, in addition to a wide angle (full Earth) antenna.

- The civilian L2 (L2C) signal, providing improved accuracy of navigation,

and acting as a redundant signal in case of localised interference (in addition to the opportunity of correcting for ionospheric effects)

- The Safety of Life (L5) signal, a civilian-use signal, broadcast on the L5 frequency (1176.45 MHz), implemented since the first GPS IIF launch.

- The further L1 civilian-use (L1C) signal, to be broadcast on the same L1 frequency (1575.42 MHz) that currently contains the C/A signal used by all current GPS users. The L1C will be available with first Block III launch, currently scheduled for 2013. The L1 signal will be interoperable with Galileo satellite navigation system.

- A further frequency (L4) will be available at 1379.913 MHz and will be under study for improved ionospheric correction.

All these new signals will be transmitted with increased power and wider bandwidth and will ensure improvements on the services which they drive. The full operational constellation will consist of 32 satellites.

The GPS III system will provide also information about worldwide localisation of signals of nuclear detonations, through the Nuclear detonation Detection System payload (NDS).

3.2.3 Satellite instruments

All satellites are equipped with dual solar arrays supplying over 400 W. A S-Band communication link (2227.5 MHz) is used for control and telemetry. A UHF channel provides cross-links among spacecrafts. A propulsion system is used for orbital correction.

The core of the payload includes two L-Band navigation signals, L1 at 1575.42 MHz and L2 at 1227.60 MHz, locally generated by means ofprecise atomic oscillators. The antenna for the navigation signals sends Right-Hand Circularly Polarised (RHCP) waves.

3.2.3.1 Transmitter and receiver clocks

Precise measurements of time are crucial in GNSS positioning. Satellite clock reference and clock offset parameters are also included in the navigation message, because normal receiver clocks are not synchronised with satellite ones.

The core of the satellite payload is the atomic oscillator. Currently all satellites are provided by two Rubidium and two Caesium atomic clocks. One of these atomic standards is designated as primary and serves as timing reference on board of the space vehicle for navigation signal generation and transmission. These extremely accurate GPS atomic clocks must keep time to within a few nanoseconds a day and they are synchronised. Despite the extreme precision, if the satellite clock drifts and offsets are not taken into account, the resulting error in the positioning can become quite large. Furthermore the correction cannot be applied to the satellite clock, therefore drifts and clocks add up. The Master Control Station (in Colorado Springs), collects all satellite data received by all the Control Segment stations and systematically they update parameters for orbits and clock synchronisation. Parameters for satellite clock correction are provided in the navigation message by means of coefficients of a second order polynomial interpolation:

$$t_e = a_0 + a_1(t - t_r) + a_2(t - t_r)^2 \tag{3.1}$$

where: a_0, a_1, a_2 are the polynomial coefficients; t_e is the estimated (corrected) time; t_r is the reference time for the parameters computation; t is the clock uncorrected time.

These parameters are uploaded to the satellites for realtime broadcast, providing the real time positioning and timing services.

The stability and accuracy of a clock is often presented in terms of Allan variance:

$$\sigma_A^2 = \left\langle \frac{[\phi(t + 2\tau) - 2\phi(t + \tau) + \phi(t)]^2}{2\tau\omega_0} \right\rangle \tag{3.2}$$

where: $\sigma_A^2(\tau)$ is the Allan variance; τ is the averaging time; $\phi(t)$ is the clock

signal phase at time t; ω_0 is the natural frequency of the source; $\langle\rangle$ indicate an average over a very long time.

Practically the Allan variance is approximated through a series of samples. Using the definition:

$$\bar{y_k} = \frac{\phi(t_k + \tau) - \phi(t_k)}{2\tau\omega_0} \tag{3.3}$$

we can write the Allan variance based on N samples, as:

$$\sigma_A^2(\tau) = \frac{1}{2(N-1)} \sum_{1}^{N-1} (\bar{y_{k+1}} - \bar{y_k})^2 \tag{3.4}$$

Clock stability is then usually expressed in terms of the deviation, $\sigma_A(\tau)$. The most common oscillators available are quartz, rubidium cell, caesium beam, and hydrogen maser. Figures 3.1 and 3.2 shows the typical stability and accuracy $\sigma_A(\tau)$ that can be expected from each oscillator. Quartz oscillators show an

	τ = 1 sec	τ = 1 day	τ = 1 month
Quartz	10^{-12}	10^{-9}	10^{-8}
Rubidium	10^{-11}	$10^{-12}/10^{-13}$	$10^{-11}/10^{-12}$
Cesium Beam	$10^{-10}/10^{-11}$	$10^{-13}/10^{-14}$	$10^{-13}/10^{-14}$
Hydrogen Maser	10^{-13}	$10^{-14}/10^{-15}$	10^{-13}

Figure 3.1: Typical Oscillator Stabilities expressed as Allan variance. (cf. [24]).

Allan variance of 10^{-12} over short periods of about one second to one minute. This is comparable to a caesium beam and better than a rubidium cell over short periods. In the longer term, for instance a day or a month, quartz oscillators perform much worse than atomic standards. The stability of a hydrogen maser is about an order of magnitude better than the caesium beam for periods of up to one day.

The very high short term accuracy and stability of quartz oscillators make them well usable in GPS receivers for positioning purposes, requiring high receiver clock precision during the signal travel from satellite to ground station, and without need of synchronisation with GPS reference time.

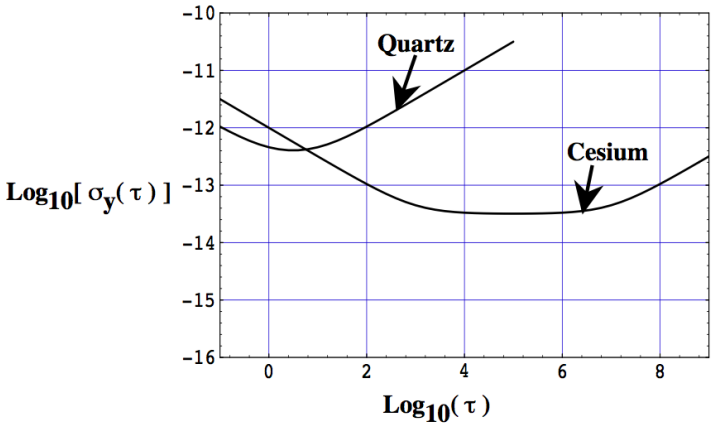


Figure 3.2: Typical Allan deviations of caesium clocks and quartz oscillators, plotted as a function of averaging time τ (cf. [2]).

3.3 GPS signal

The GPS signal is created on board of the space vehicle. As we already know two signals at different frequencies are produced to be used for ionospheric corrections. The signal carriers are in the L IEEE (Institute for Electrical and Electronic Engineers) band (1 to 2 GHz) and they originate from a common local oscillator with fundamental frequency f_0, nominally at 10.23 MHz. The two carriers derive from the fundamental frequency by using Phase Locked Loop (PLL) frequency multipliers. The carrier signals are thus coherent with the same frequency clock:

$L_1 = f_0 \times 154 = 1575.42\ MHz$

$L_2 = f_0 \times 120 = 1227.60\ MHz$

The GPS carriers are modulated by signals which identify univocally the transmitting satellite and provide information about its position (cf. [33]). Three code types modulate the carrier signals (Fig. 3.3):

- The C/A code is the basis of the SPS. It is a bi-phase $(+1, -1)$ signal 1023 chips long. The chip rate is 1.023 $Mchip/s$, resulting in about 1 μs time length of each chip. The entire C/A code has the length of 1 ms. The

spectrum of the signal (a square wave) is a *sinc* function, with a null to null bandwidth of 2 MHz.

- The P-code, or Y code when the encryption of Anti-Spoofing (AS) system is activated, is a bi-phase $(+1, -1)$ very long code, with a repetition time of 266 days (8 weeks). The chip rate is 10.23 $Mchip/s$. The whole P-code is divided in 38 segments, each of them is 7 day long. Each segment is assigned to a satellite. Also this signal is a square wave and the spectrum bandwidth is 20 MHz.

- Finally the navigation data message (D) code with a bit rate of 50 bps, each bit is 20 ms long. It's a bi-phase $(+1, -1)$ signal. The navigation code includes satellite ephemeris, time information, clock synchronisation parameters, i.e. all these parameters that are required during the process of position determination.

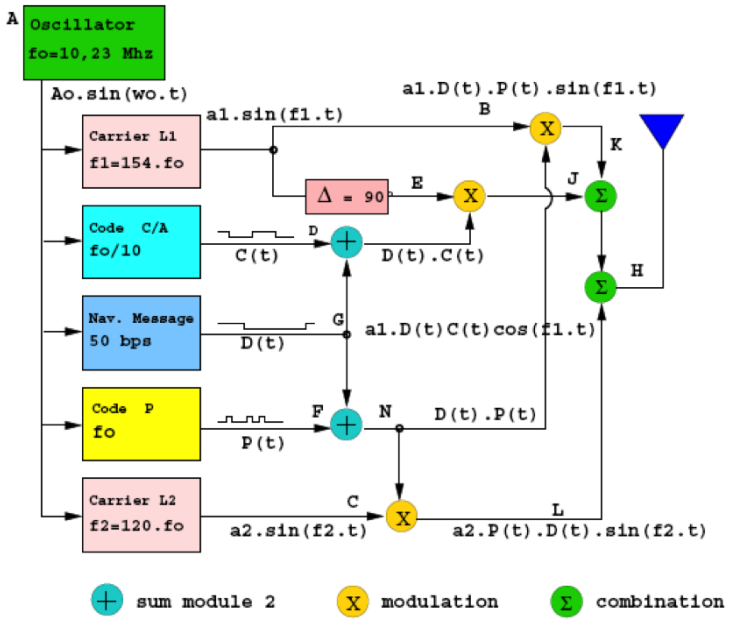


Figure 3.3: Scheme of the GPS satellite signal structure.

Signals S_{L_1} and S_{L_2} at two frequencies, can be expressed as:

$$\begin{aligned} S_{L_1} &= A_P \mathrm{P}^k(t)\mathrm{D}^k(t)cos(2\pi f_1 t + \phi) + A_C\, \mathrm{C/A}^k(t)\, D^k(t)sin(2\pi f_1 t + \phi) \\ S_{L_2} &= A_P \mathrm{P}^k(t)\mathrm{D}^k(t)cos(2\pi f_2 t + \phi) \end{aligned} \tag{3.5}$$

where:
A_P is the amplitude of the P-code
$\mathrm{P}^k(t)$ is the precision code ($\pm$ 1) for the k_{th} satellite
$\mathrm{D}^k(t)$ is the navigation code ($\pm$ 1) for the k_{th} satellite
A_C is the amplitude of the C/A code
$\mathrm{C/A}^k(t)$ is the coarse acquisition code ($\pm$ 1) for the k_{th} satellite
ϕ is the initial phase of the signal

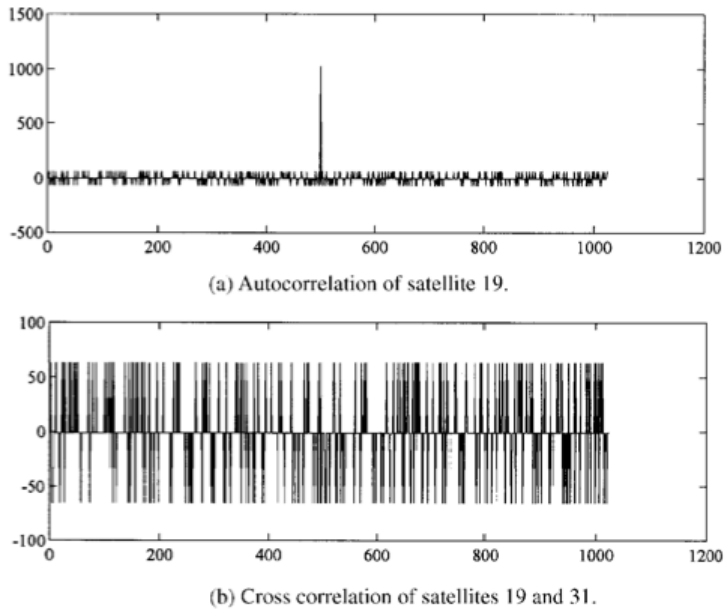
The GPS signal is a phase-modulated signal with phase $\phi = (0, \pi)$; this type of phase modulation is referred to as Binary Phase Shift Keying (BPSK).

Coarse Acquisition (C/A) and Precision (P) codes are in quadrature to allow a best detection of the signal.

The code signals can reach a minimum power level of $-130\,dBm$ and the spectrum is spread; so they cannot be detected from a spectrum analyser. The Spread Spectrum Modulation (SSM) is used to transmit the signal and the use of the Code Division Multiple Access (CDMA) allows to use the same band portion and the same centre frequency for all the signals transmission.

The codes (C/A and P) relative to different satellites are orthogonal: their cross correlation and their autocorrelation, computed for a non-zero shift delay, are very low. The autocorrelation function has a strong peak when the shift delay is zero, so that the receiver can accurately reconstruct and decode the signal. Often these codes are referred as Pseudo Random Noise (PRN), because the white noise has an autocorrelation function essentially with the same properties.

At the receiver level, codes of each signals must be generated and used to correlate with the received satellite signals, for decoding the information content.



(a) Autocorrelation of satellite 19.

(b) Cross correlation of satellites 19 and 31.

Figure 3.4: PRNs crosscorrelation and autocorrelation examples.

3.3.1 Pseudorange measurements

The signal travel time is the time employed by the signal transmitted at time T_j^i from the *i-th* satellite[3], to reach the *j-th* receiver at time T_j. A receiver provides the travel time by computing the time shift needed to align the code receipt from satellite at time T_j with a replica generated at the same time in the receiver. The quantity obtained by multiplying this measure by the GPS reference speed of light (299792458 m/s) is called pseudorange. "Pseudo" means that such measurement does not give the "range" (i.e. the distance) between satellites and receivers, due to a number of effects generically referred as signal errors. In § 2 we have already seen the effects of atmosphere on the signal delay, and now we know that they can contribute for at least some tens of metres to mis-positioning, and much more for satellites far from the receiver zenith. In addition to the atmospheric effects there are a number of further error sources: for our task the

[3]In a receiver the transmission time is normally computed subtracting the pseudorange (as travel time of the signal) from the instant of signal reception. This is the reason why we prefer to indicate the time of satellite signal transmission not simply as T^i, but as T_j^i, reminding that it is known thanks to a procedure of signal decoding and processing made by a given receiver.

neutral atmospheric effects, and the wet component, are a source of information, but all the other error components are something to be eliminated as precisely as possible. The present work however, and specifically the analyses of § 4, we will not make a systematic use of real GPS data, thus the aim of this section, and more generally of the following of this chapter, will be to analytically identify the various terms composing the pseudorange and the error associated in their evaluation, using some samples of GPS observations analysed for the purpose.

The main error sources in GPS signals can be expressed through the following formula:

$$\begin{aligned}\rho_j^i(T_j, f) &= R_j^i(T_j) + c[\epsilon_{rj}(T_j) - \epsilon_s^i(T_j^i) + \Delta T_{Sat}(T_j^i, f) + \Delta T_{Trop} + \Delta T_{Ion}(f) \\ &\quad + \Delta T_{Rel}(T_j^i) + \Delta T_{Orbit}(T_j^i) + \Delta T_{Ric}(T_j, f)] \end{aligned} \tag{3.6}$$

where:

$\rho_j^i(T_j, f)$ is the pseudorange measured by the *j-th* receiver at time T_j relative to the *i-th* satellite at the frequency f (L_1 or L_2);

$R_j^i(T_j)$ is the geometric distance between the *i-th* satellite position at time T_j^i of signal transmission and the *j-th* receiver position at time T_j;

$\epsilon_{rj}(T_j)$ is the *j-th* receiver clock shift due to the synchronisation error at time T_j;

$\epsilon_s^i(T_j^i)$ is the *i-th* satellite clock shift due to the synchronisation error at time T_j^i;

$\Delta T_{Sat}(T_j^i, f)$ is the satellite instrumental delay introduced by the *i-th* satellite at time T_j^i of signal transmission. It is dependent on the frequency of the signal. Different frequencies have different channel in the transmitter. It is commonly referred to as T_{GD} (Group Delay);

ΔT_{Trop} is the delay introduced by the signal propagation into the troposphere, with respect to the free space propagation;

$\Delta T_{Ion}(f)$ is the delay introduced by the signal propagation into the ionosphere, with respect to the free space propagation. It is frequency dependent;

$\Delta T_{Rel}(T_j^i)$ is the relativistic effect on the *i-th* satellite at time T_j^i.

$\Delta T_{Orbit}^i(T_j^i)$ is the delay caused by the error in the computation of the *i-th* satellite position at time T_j^i.

$\Delta T_{Ric}(T_j, f)$ is a further delay introduced by the receiver. It is primarily caused by the processing and propagation signal delay, but also by poor measurement accuracy. It is dependent on the frequency of the signal, as different frequencies have different channels in the receiver.

3.3.2 Carrier phase measurements

The signal carrier phase, expressed in terms of carrier cycles, consists of an integer part and a fractional part, the latter being the only one measurable by a receiver. Proper algorithms in post-processing can estimate the integer part of carrier phase (i.e. resolving the ambiguity on the number of cycles). The *j-th* receiver measure the phase difference between the signal transmitted by the *i-th* satellite at time T_j^i, detected by the receiver at time T_j, and the signal replica generated in the receiver. This phase measure is ambiguous due to the unknown integer number of carrier cycles during the signal travel between satellite transmitter and ground receiver.

Also carrier phase is affected by the same error sources as pseudorange:

$$\begin{aligned} \Phi_j^i(T_j, f) &= \delta\phi_j^i(T_j, f) + N_j^i(T_j, f) \\ &= \frac{f}{c} R_j^i(T_j) + f[\epsilon_{rj}(T_j) - \epsilon_s^i(T_j^i) + \Delta T_{Sat}(T_j^i, f) + \Delta T_{Trop} \\ &\quad -\Delta T_{Ion}(f) + \Delta T_{Rel}(T_j^i) + \Delta T_{Orbit}(T_j^i) + \Delta T_{Ric}(T_j, f)] \end{aligned} \tag{3.7}$$

where the additional terms, with reference to (3.6), are:

$\Phi_j^i(T_j, f)$ is the total phase comprehensive of both integer and fractional part;

$\delta\phi_j^i(T_j, f)$ is the receiver measurement of the fractional part of carrier phase (hereafter referred as carrier phase measurement);

$N_j^i(T_j, f)$ is the integer number of carrier cycles during the signal propagation (i.e. the ambiguity of the measurement);

Note that in (3.7) the term caused by the ionospheric effect is the opposite than the analogous in (3.6), according to what seen in § 2.3.1.

Next sections will go through the main the error sources included in (3.6) and in (3.7).

3.3.2.1 Observables of common use

In this section the main types of "observables" used in the GPS community are introduced, and their meaning shortly explained.

1. $P1$, $L1$ pseudorange and phase measurements respectively in the first frequency: $f_1 = 1575.42 MHz$

2. $P2$, $L2$ pseudorange and phase measurements respectively in the second frequency: $f_2 = 1227.6\, MHz$

3. $P3$, $L3$ pseudorange and phase linear combinations respectively, called *Ionosphere free LC* because this linear combination nearly completely eliminates the ionospheric refraction effects:

$$P_3 = \frac{1}{f_1^2 - f_2^2}(f_1^2 P_1 - f_2^2 P_2) \tag{3.8}$$

$$L_3 = \frac{1}{f_1^2 - f_2^2}(f_1^2 L_1 - f_2^2 L_2) \tag{3.9}$$

4. $P4$, $L4$ pseudorange and phase linear combinations respectively, called *Geometry free LC* because their linear combination cancel the frequency independent part of the measurement, leaving only the ionospheric effects and the instrumental constants (multipath if present, instrumental biases, and other observational noises, see § 3.4):

$$P_4 = P_1 - P_2 \tag{3.10}$$

$$L_4 = L_1 - L_2 \tag{3.11}$$

5. $P5$ is a pseudorange linear combination called *Narrow Lane (PW) LC*:

$$PW = P_5 = \frac{f_1 P1 + f_2 P_2}{f_1 + f_2} \tag{3.12}$$

6. $L5$ is a phase linear combinations, called *Wide-lane (LW) LC* that gives an observable with a wavelength ($\lambda_W = 86.2cm$) four times bigger than $L1$ or $L2$:

$$LW = L_5 = \frac{f_1 L1 - f_2 L_2}{f_1 - f_2} \tag{3.13}$$

7. $L6$ is phase linear combinations, called *Melbourne-Wubbena (MW) LC*, that is exactly the difference between LW and PW:

$$MW = L_6 = \frac{f_1 L1 - f_2 L_2}{f_1 - f_2} - \frac{f_1 P1 + f_2 P_2}{f_1 + f_2} \tag{3.14}$$

3.4 Components of the apparent signal travel-time

We already know that the signal travel time from a satellite to a receiver is given by much more than the simple satellite-receiver distance times the light speed. At this purpose this section is focused to the analyses of the various terms that compose such travel time, and to the errors they may introduce into the delay measurements.

3.4.1 Tropospheric delay

For tropospheric delay it is normally meant the phase delay due to the dry and wet neutral atmosphere. As already told this is an error source for positioning but a source of information for us, that we want to evaluate as accurately as possible, in order to estimates the tropospheric parameters of interest in meteorology. As a consequence we will not treat the problem of how to eliminate the effects we have described in § 2, and more precisely in § 2.2.1 and § 2.2.2. On the contrary we will analyse, but in the following sections, all the other sources of errors, while

the discussion on how to retrieve the searched information from the tropospheric delay will be dealt in § 4.

3.4.2 Ionospheric delay

In § 2.2.3and § 2.3.1we have already seen the origin of a refraction index in ionosphere that results smaller than 1, namely on average $n:\ n-1 \sim 10^{-5}$. We have also commented that this implies a phase advance and a group delay that are equal in magnitude, apart from bending effects.

The difference $\Delta\tau$ between group delays, as in Eq. (2.76), and phase advances, as in Eq. (2.75), obtained in the two different frequencies L_1 and L_2, can be written as:

$$\Delta\tau = B\left[\frac{1}{f_{L2}^2} - \frac{1}{f_{L1}^2}\right] = \frac{B}{f_{L2}^2}\left[\frac{f_{L1}^2 - f_{L2}^2}{f_{L1}^2}\right] = \Delta T_{Ion}(f_{L2})\left[1 - \frac{f_{L2}^2}{f_{L1}^2}\right]$$
$$... = \Delta T_{Ion}(f_{L1})\left[\frac{f_{L1}^2}{f_{L2}^2} - 1\right] \qquad (3.15)$$

where $B = \alpha\, TEC$.

Considering P_1 e P_2 the pseudorange measurements with the *precision-code*, by assuming negligible the variation with the frequency of all terms present in Eq. (3.6):

$$\Delta T_{Ion}^{P_2}(f_{L2}) = \frac{[P_2 - P_1]}{c}\frac{f_{L1}^2}{f_{L1}^2 - f_{L2}^2}$$
$$\Delta T_{Ion}^{P_1}(f_{L1}) = \frac{[P_2 - P_1]}{c}\frac{f_{L2}^2}{f_{L1}^2 - f_{L2}^2} \qquad (3.16)$$

Unfortunately the frequency dependent terms, introduced by the satellite and receiver hardware, are not negligible and must be carefully modelled in order to correctly retrieve the ionosphere effect, as we will see later. On the contrary the precision in the approximation of Eq. (2.59) are generally very accurate, but, under some conditions, they can result precise at the 1 % or less: for high TEC values this means to have errors of several centimetres or more (for zenith observations). This can be too big an error for atmospheric sounding; in any case from Eq. (2.60) the value of such approximations can be evaluated.

Finally it is worth to note that it is not possible to process only carrier phase measurements in two different frequencies for computing the ionospheric delay, because of the ambiguities. Some ancillary information must be used for solving these ambiguities, such as pseudorange measurements.

3.4.3 Relativistic effects

In order to determine the orbits of satellites, the GPS system uses an inertial Cartesian coordinate system, the Earth Centred Inertial (ECI) reference system in which the origin is at the centre of mass of the Earth, the x-y plane is coincident with the Earth's equatorial plane and the x and y axes are oriented along determined positions over the celestial sphere; the z axis is normal to the x-y plane in the north direction.

Non inertial motions (i.e. satellite orbits and Earth rotation) and the Earth gravitational field make relativistic effects non negligible with respect to signal accuracy constrains.

A first effect regards the change in the signal frequency. The nominal frequency of $10,23\,MHz$ at a ground receiver is however guaranteed by a pre-launch proper shift in the atomic oscillator frequency, that accounts for the in-orbit average relativistic effect, as follows:

$$-\frac{f_0 - f'}{f'} = \frac{1}{2}\left(\frac{v}{c}\right)^2 + \frac{\Delta U}{c^2} \tag{3.17}$$

Where f' is the working frequency, f_0 is the nominal frequency ($10.23\,MHz$), v is the mean satellite speed, c is the light speed, and ΔU is the mean gravitational potential difference between satellite and receiver. It is found $f' = 10.23 \times 10^6 - 4.567 \times 10^{-3}\,MHz$.

Eq. (3.17) relationship considers a circular orbit for GPS satellites. To take into account for the orbit eccentricity (that has consequences in the gravitational potential differences) a correction must be applied to the pseudorange and carrier phase relationships (3.6) and (3.7):

$$\Delta T_{Rel} = \frac{2\sqrt{\mu a}}{c^2}\, e \sin E \tag{3.18}$$

where $\mu = 3986005 \times 10^8 \frac{m^3}{s^2}$ is the product of the universal gravitational constant and the Earth mass, a is the semimajor axis, e is the orbit eccentricity and E is the eccentric anomaly (to be computed using the Kepler equation). a and e are part of the broadcast ephemeris.

In any case some other relativistic effects caused by the orbit oscillations, perturbations on the Earth gravitational field etc., persist. Furthermore the satellite and receiver relative motions can produce a Doppler effect approximately within $\pm 5\ KHz$.

These effects can be modelled for precise estimations in order to give negligible contribution to the overall error in positioning or, in the case of our interest, in the tropospheric delay evaluation. Here follows some major terms.

The Holdridge model (cf. [48]) accounts for general relativity effects on the signal delay:

$$\delta\rho_{Rel} = \frac{2\mu}{c^2} \ln \frac{\rho_j + \rho^i + \rho^i_j}{\rho_j + \rho^i - \rho^i_j} \tag{3.19}$$

where: ρ_j is the geocentric position of the *j-th* receiver, ρ^i is the geocentric position of the *i-th* satellite, ρ^i_j is the geocentric distance between the *j-th* receiver and the *i-th* satellite.

$\delta\rho_{Rel}$ can reach a maximum value of about $2\,cm$.

The correction for the acceleration of the Earth satellite, following the model of McCarthy (cf. [48]):

$$\Delta\vec{a} = \frac{\mu}{c^2 r^3} \left\{ \left[4\frac{\mu}{r} - v^2 \right] \vec{r} + 4(\vec{r} \cdot \vec{v})\vec{v} \right\} \tag{3.20}$$

where $\vec{r}$ is the geocentric satellite position vector, $\vec{v}$ is the geocentric satellite velocity vector and $\vec{a}$ is the geocentric satellite acceleration vector.

In addition to all the previous terms, there is a further factor due to relative motions of satellites and receivers. A receiver can be generally in motion on the Earth's surface. For our purposes however only receivers with fixed position over the Earth surface will be taken into account. Thus the only additional effect that we have to consider is known as Sagnac effect, which is due to the Earth rotation during the signal travel time. This effect will be treated in § 3.4.6.

3.4.4 Instrumental delays and differential code biases

There are different instrumental biases for different codes and different frequencies. The available pseudorange observables are C1 (C/A code pseudorange on the f_1 frequency), P1 (P-code pseudorange on the f_1 frequency), P2 (P-code pseudorange on the f_2 frequency). For each one we can identify the following biases:

- B_{C1} : bias on the C1 observable
- B_{P1} : bias on the P1 observable
- B_{P2} : bias on the P2 observable

As these biases are from instruments there are two of them for each observable, one produced by the satellite and the other from the receiver. Separately for the satellite and for the receiver we have to consider the discrepancies between "True" (corrected from the biases) and "Measured" (affected by biases) values:

$$\begin{aligned} C_1^M &= C_1^T + B_{C1} \\ P_1^M &= P_1^T + B_{P1} \\ P_2^M &= P_2^T + B_{P2} \end{aligned} \tag{3.21}$$

The single biases are generally not retrievable in their absolute values, because they are generated by fluctuations in the performances of the various instruments caused by changes of work conditions (e.g. temperature) and status (wear) and because the instrumental biases of receiver and satellite are adds up into measurements, so it is common the use of biases combinations, usuallt referred to as Differential Code Biases (DCB):

$$\begin{aligned} B_{P1-P2} &= B_{P1} - B_{P2} \\ B_{P1-C1} &= B_{P1} - B_{C1} \end{aligned} \tag{3.22}$$

These combination cannot directly be applied for a single observable correction, but to a combination of observables. For example considering the availability

of observables C1 and P2 from a receiver, the known approximated model of pseudoranges (see 3.6), by neglecting the time (transmission and reception) dependence, can be written :

$$\begin{aligned} C1_j^i &= R_j^i + c[\epsilon_{rj} - \epsilon_s^i + \Delta T_{Trop} + \Delta T_{Ion}(f_1) + \Delta T_{Rel}^i + \Delta T_{Orbit}^i + B_{C1}] \\ P1_j^i &= R_j^i + c[\epsilon_{rj} - \epsilon_s^i + \Delta T_{Trop} + \Delta T_{Ion}(f_1) + \Delta T_{Rel}^i + \Delta T_{Orbit}^i + B_{P1}] \\ P2_j^i &= R_j^i + c[\epsilon_{rj} - \epsilon_s^i + \Delta T_{Trop} + \Delta T_{Ion}(f_1) + \Delta T_{Rel}^i + \Delta T_{Orbit}^i + B_{P2}] \end{aligned} \tag{3.23}$$

where

$B_{C1} = \Delta T_{Sat}(T_j^i, C_1) + \Delta T_{Ric}(T_j, C_1)$ is the bias introduced on the C1 observable by the instrumental delays of the satellite and the receiver;

$B_{P1} = \Delta T_{Sat}(T_j^i, P_1) + \Delta T_{Ric}(T_j, P_1)$ is the bias introduced on the P1 observable by the instrumental delays of the satellite and the receiver;

$B_{P2} = \Delta T_{Sat}(T_j^i, P_2) + \Delta T_{Ric}(T_j, P_2)$ is the bias introduced on the P2 observable by the instrumental delays of the satellite and the receiver.

The computation of the delay introduced by the ionosphere, as previously mentioned, involves the combination of the dual frequency observations:

$$\begin{aligned} \Delta T_{Ion}(f2) &= \frac{[P_2 - C_1]}{c} \frac{f_1^2}{f_1^2 - f_2^2} \\ \Delta T_{Ion}(f1) &= \frac{[P_2 - C_1]}{c} \frac{f_2^2}{f_1^2 - f_2^2} \end{aligned} \tag{3.24}$$

Therefore, in the computation of the ionospheric delay, we have to compute the quantity:

$$\begin{aligned} P_2 - C_1 &= \Delta T_{Ion}(f2) - \Delta T_{Ion}(f1) + B_{P2} - B_{C1} \\ &= \Delta T_{Ion}(f2) - \Delta T_{Ion}(f1) - B_{P1} + B_{P2} + B_{P1} - B_{C1} \\ &= \Delta T_{Ion}(f2) - \Delta T_{Ion}(f1) - B_{P1-P2} + B_{P1-C1} \end{aligned} \tag{3.25}$$

Differential Code Biases have thus to be known for the complete correction of the ionospheric delay.

Differential code biases are different for satellites and receivers. Methods for estimating with acceptable accuracy the satellite based ones have been investigated from several years from the GPS community and from several institution. The values of B^i_{P1-P2} are computed (referred to as τ^i_{GD} in the Interface Control Document (cf. [21]) by ground control stations and transmitted by satellites in the navigation message. The values of B^i_{P1-C1} can be corrected by applying tables that are constantly updated. The relationship between B^i_{P1-P2} and τ^i_{GD}, is:

$$\tau^i_{GD} \simeq \frac{1}{c(1-\gamma)} B^i_{P1-P2} \simeq -1.55\, B^i_{P1-P2} \tag{3.26}$$

being $\gamma = \frac{f_1^2}{f_2^2} \simeq 1.65$. Now we can write:

$$\tau^i_{GD} = \frac{1}{c(1-\gamma)} B^i_{P1-P2} \tag{3.27}$$

Eq. (3.25) will become:

$$\begin{aligned} P_2 - C_1 = \Delta T_I(f2) - \Delta T_I(f1) + c(\gamma-1)\tau^i_{GD} + \\ + B^i_{P1-C1} - B_{jP1-P2} + B_{jP1-C1} \end{aligned} \tag{3.28}$$

Thus in order to correctly compute the ionospheric delay, we have to evaluate all the differential code biases not given in the broadcast message (B^i_{P1-C1}; B_{jP1-P2}; B_{jP1-C1}). When all these contribution are known the ionospheric delay is:

$$\begin{aligned} \Delta T_I(f2) &= \frac{\gamma}{\gamma-1}\left[P_2 - C_1 - c(\gamma-1)\tau^i_{GD} - B^i_{P1-C1} + B_{jP1-P2} - B_{jP1-C1}\right] \\ \Delta T_I(f1) &= \frac{1}{\gamma-1}\left[P_2 - C_1 - c(\gamma-1)\tau^i_{GD} - B^i_{P1-C1} + B_{jP1-P2} - B_{jP1-C1}\right] \end{aligned} \tag{3.29}$$

As mentioned before, the quantity B^i_{P1-C1} can be computed using methods of proven accuracy and well known in the GPS community. Some software codes are freely available for the correction of this satellite based biases, and the results gives errors negligible with respect to other sources.

On the contrary for the computation of the receiver based DCBs there are no standard methods or tables to refer to. In order to make an approximate estimation for DCBs, we have compared measures from some GPS stations in Tuscany with theoretical delays obtained from an established model of ionosphere. Such model is the Klobuchar model (cf. [25]), which can estimate ionospheric effects, and that is used as reference model by the GPS community, generally to remove ionospheric effects from one band signals.

The Klobuchar model allows to compute the ionospheric TEC by means of some algorithms that parameterise the sun activity. In order to correct for seasonal and sun spot number changes, the algorithm uses different sets of eight coefficients depending upon the period of the year and the average solar flux. Better estimations are obtained during night periods, when part of the ionisation relaxes towards more stable (and known) values (see § sez:iono).

Fig. 3.5 shows in the same plot the ionospheric delay computed by using Eq. (3.29), including correction for satellite instrumental delays, and the ionospheric delay as evaluated through the Klobuchar model for the same period. The ionospheric delay computed using two frequencies is affected by receiver DCBs. The comparison with the Klobuchar data during the minimum value of TEC relative to minimum values of sun effects (i.e. night time close to the sunset and for high satellite elevation angles) can be used to estimate a receiver DCB (comprehensive of B_{jP1-P2} and B_{jP1-C1}), as a free parameter whose value minimises the distance of the two curve minima.

In this case a delay of around $-3\,m$ seems to be imputable to the receiver DCB: it is clear that the capability of accurately estimating DCBs and their time variability is critical for a successful retrieval of atmospheric parameters from fixed stations.

3.4.5 Orbit parameters errors

Ephemeris data, broadcast in the satellite navigation message, give the satellite positions as a function of time. They are predicted starting from satellite po-

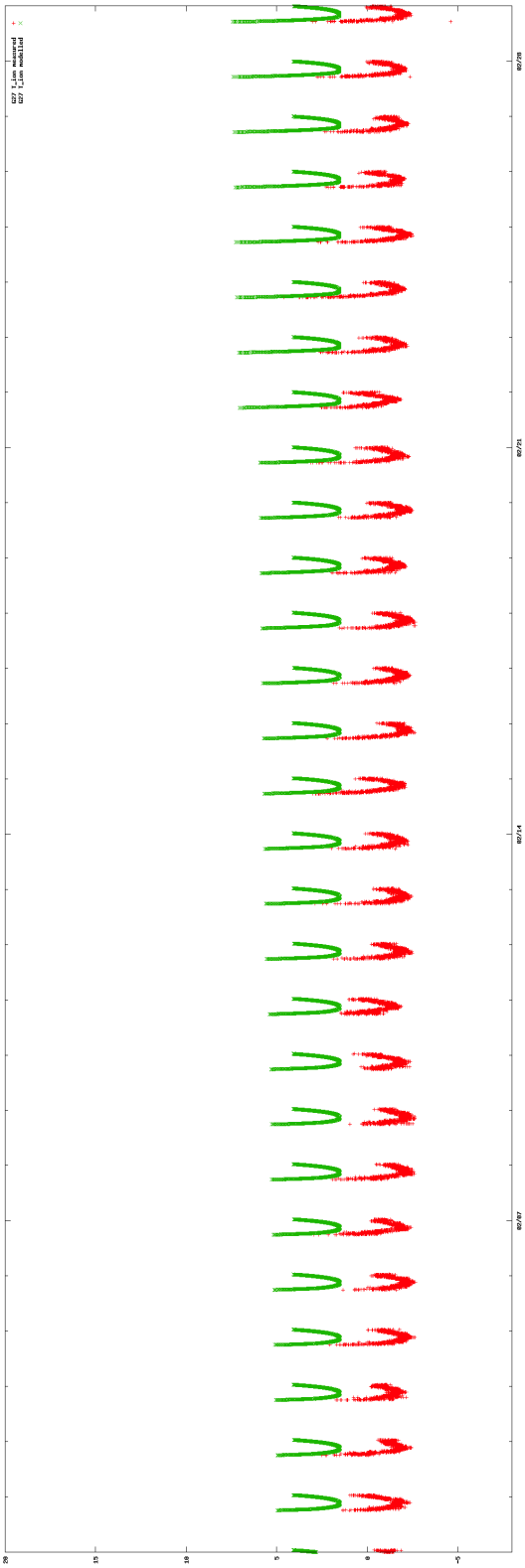


Figure 3.5: Comparison between dual frequencies ionospheric delays affected by receiver biases (red) and corresponding ionospheric delays from Klobuchar model data (green).The period is 2010/01/02-2010/28/02. Data are from the station of Massa (Tuscany).

sitions regularly verified at the ground control stations. Typically, overlapping intervals of 4 hours of GPS data are used by the operational control system to update satellite orbital elements for a following period of 1 hour.

Ephemeris errors can be of 2 m to 5 m, and can reach up to 50 m under selective availability. The ephemeris error is usually decomposed into components along three orthogonal directions defined for the satellite orbit: radial, along-track and cross-track.

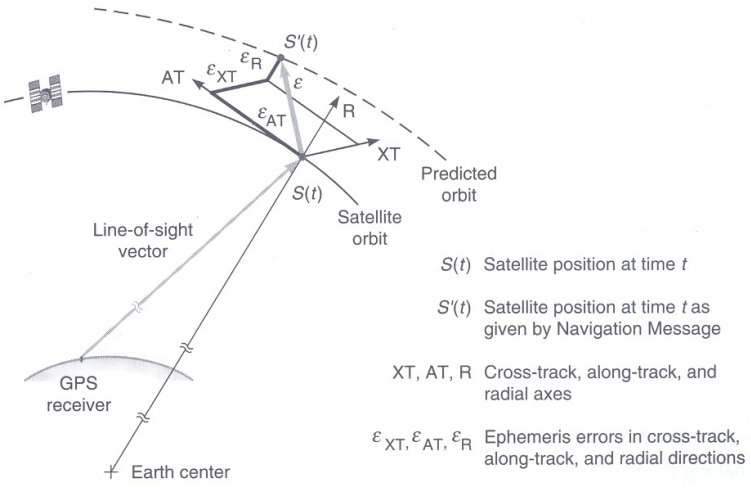


Figure 3.6: Ephemeris error components (cf [29]).

The radial component in the estimation and prediction of satellite position is smaller than the other by one order of magnitude. These are good news as we know that geometric errors impacts on atmospheric sounding only on their along the receiver-satellite (line-of-sight) vector direction, and this projection depends primarily on the radial component and secondarily on the cross-track and along-track components.

Due to the fact that this errors originate from a prediction, they increase with the time interval between observations and updating of the orbital parameters.

The increasing need of precise ephemeris in quasi-real time for an increasing number of applications has concentrated the efforts of the user community to develop products more and more precise with respect to ephemeris broadcast in

the navigation message. Several institutions, e.g., the International GPS Service for Geodynamics (*IGS*), the U.S. National Geodetic Survey (*NGS*), and Geomatics Canada, have developed postmission precise orbital services. Precise ephemeris data is based on GPS data collected at a global GPS network coordinated by the Internation GNSS Service (*IGS*). Precise ephemeris data contain very accurate parameters for the correction of space vehicle clock offset and drift and for retrieving satellite position at transmission time. At present, precise ephemeris data are also near real-time available through the Ultra-Rapid ephemeris service[4], with an accuracy of 5 *cm*. The best accuracy can be gained with the delayed service, after $12 \div 18$ days, that is 2.5 *cm*. There are other two services with intermediate delivery-time, and consequent intermediate parameter accuracies.

3.4.5.1 Retrieval of satellite positions

For computing the user position it is more convenient to refer to an Earth rotating reference frame, the Earth Centred Earth Fixed (ECEF) Cartesian coordinate system. This simplifies the computation of geographical coordinates (latitude, longitude, height) for the receiver position. The ECEF reference system has the origin in the Earth centre of mass, the x-y plane coincident with the equatorial plane, the x axis oriented along 0° longitude, and the z axis normal to the x-y plane and oriented towards the north pole.

Satellite position is univocally determined from the knowledge of instant of time (epoch) we want to refer to, by means of orbital parameters (ephemeris) contained into the navigation message of the satellite signal, that are systematically updated about every two hours. This parameters will be described in § 3.4.5.2. Once decoded, ephemeris must be processed through a (large) set of established equations. The steps start from the computation of the correct instant of the signal transmission and the resolution of the Kepler equation for the satellite orbit, in a loop fashion up to obtain the convergence to the satellite

[4]See http://igscb.jpl.nasa.gov/components/prods.html.

position coordinates. A detailed explanations of such steps is not relevant for this work, however an exhaustive analysis of the process can be found in (cf. [3] and [11]).

In order to completely describe the observation geometry and its effect on signals, it is necessary to compute the Earth rotation during the signal travel time: this is known as the Sagnac effect and will be described in § 3.4.6.

3.4.5.2 Ephemeris

Orbital parameters are "written" into the navigation message. They are referred to a determined time (epoch). Their meaning is clear only knowing the orbital equation to be solved for retrieving satellite positions. However these parameters will be listed to give an idea of the number of features that need to be taken into account, according to the Interface Control Document (cf. [21]):

t_{oc}: clock reference time (in seconds) used for clock offset computation;

a_0 , a_1, a_2: polynomial coefficient for clock offset computation;

t_{GD}: (Group Delay), instrumental delay delay differential;

t_{oe}: reference time (in seconds) since the GPS week start (at the Saturday/Sunday transition) of ephemeris values;

M_0: mean anomaly;

Δn: mean motion difference from computed value;

e: satellite orbit eccentricity;

$\sqrt{a}$: square root of orbital semi-major axis

Ω_0: latitude of ascending node of the orbit plane at the weekly epoch;

i_0: inclination angle of the orbit plane with respect to the equatorial plane;

ω: argument of perigee;

$\dot{\Omega}_0$ rate of right ascension;

$\dot{i}$ rate of inclination angle;

C_{rc}: amplitude of the cosine harmonic correction term to the orbit radius

C_{rs}: amplitude of the sine harmonic correction term to the orbit radius

C_{uc}: amplitude of the cosine harmonic correction term to the argument of latitude

C_{us}: amplitude of the sine harmonic correction term to the argument of latitude

C_{ic}: amplitude of the cosine harmonic correction term to the angle of inclination

C_{is}: amplitude of the sine harmonic correction term to the angle of inclination

3.4.6 Earth rotation: the Sagnac effect

The ECEF reference system rotates with the Earth, so it is not inertial. In this coordinate system a fixed receiver position is constant in time. The satellite position, computed through the broadcast ephemeris, is the one at the time of the signal transmission. Since the reference system rotates during the signal travel time, receiver position and satellite position refer to different reference systems, the ECEF at reception time and the ECEF at transmission time respectively. This effect is known as the Sagnac effect, and it belongs to the family of phenomena that happen on electromagnetic signals in moving (non necessarily non-inertial) frames due to the finite speed of light. A way to account for this effect is to recompute the satellite position for the ECEF at the reception time.

Form Fig. 3.7 we note that that the geometric distance changes during the signal travel time (Δt), for the Earth rotation. If with ω_e we indicate the rotation angular speed, $\dot{\omega_e}\Delta t$ gives the rotation angle during the signal travel time. So we can write:

$$\begin{cases} X^R &= X^T cos(\dot{\omega_e}\Delta t) + Y^T sin(\dot{\omega_e}\Delta t) \\ Y^R &= -X^T sin(\dot{\omega_e}\Delta t) + Y^T cos(\dot{\omega_e}\Delta t) \\ Z^R &= Z^T \end{cases} \tag{3.30}$$

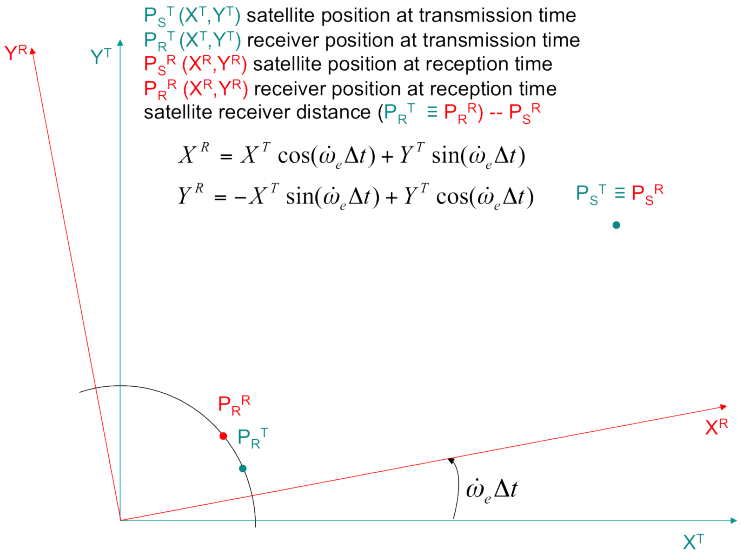


Figure 3.7: ECEF reference system rotation due to Earth rotation.

This simply solves the Signac effect for GNSS signal processing. The Sagnac effect however has also to be included in the procedures for synchronizing clocks all over the globe, satellite ones included.

3.4.7 Satellite clock offsets and drifts

The GPS system uses atomic clocks with Caesium or Rubidium oscillators. They have a nominal precision of about 1 part on 10^{13} (see the Allan Variance in Fig. 3.2). In one day ($86400\,s$) the satellite clock offset can reach the order of 10^{-7} seconds, that multiplied by the speed of light give an equivalent distance error of about 2.5 m. If the prediction of the clock behaviour can be done just with this accuracy, the resulting precision is of this order. Using updates of clock correction prediction every 12 hours, we can assume an error of $1 \div 2$ m.

The clock predicted parameters are phase bias, frequency bias, and frequency drift rate. They are uploaded to the satellites and broadcast into the navigation message. As can be easily inferred the ephemeris and satellite clock errors are closely related. As mentioned for ephemeris, the effort of scientific community is in the direction of the implementation of models for clock correction parameters

prediction with increasing accuracy. In real time Ultra-Rapid ephemeris there are also clock correction parameters, predicted with a nominal accuracy of 1.5 ns, i.e. 0.45 m. The best accuracy can be gained with the delayed service, after $12 \div 18$ days, that is $20\,ps$ equivalent to $6\,mm$

Residual satellite clock errors are the same in all measurement involving the same satellite, including P-code, C/A code and carrier phase, and can be removed when differential techniques are applicable.

3.4.8 Receiver clock errors

GPS receiver commonly use cheap crystal clocks, which are much less accurate than satellite clocks (see the Allan Variance in Fig. 3.2). The quartz receiver clock is however more stable than satellite atomic ones, for instantaneous observations. Therefore the quartz clocks are very suited for positioning purposes that require high stability for very short periods (tens of nanoseconds), that go from the transmission instant to the reception one.

The receiver clock offset, including receiver and satellite instrumental biases, should be considered as a constant for all observations regarding different satellites on view at the same instant. If atmospheric effects are neglected the problem has a simple solution. In the positioning problem, in fact, it becomes an unknown in addition to the three receiver coordinates and can be computed by solving a system of equations for the pseudorange, ρ_j^i, that, neglecting atmospheric effects and all other error sources (orbital, instrumental etc.) can be written as:

$$\rho_j^i(T_j, f) = \sqrt{[X^i(T_j^i) - X_j(T_j)]^2 + [Y^i(T_j^i) - Y_j(T_j)]^2 + [Z^i(T_j^i) - Z_j(T_j)]^2} + c\,[\epsilon_{rj}(T_j) - \epsilon_s^i(T_j^i)] \tag{3.31}$$

with X, Y, Z, coordinates and ϵ_r clock offset of satellite, if with apex i, and receiver, if with subscript j. Clearly four (non aligned) observations guarantee to solve the problem. A receiver clock error thus can be much larger than satellite clock errors: however theoretically it can be removed by making differences between observations relative to different satellites.

Of course in atmospheric sounding the problem is more complex, as Eq. (3.31) must be completed with atmospheric effects, that are unknown in addition to receiver clock errors, and the valuation through differential observations is no more straightforward.

Clock stability depends critically on the type of oscillator in use. A crystal oscillator is an electronic oscillator circuit that uses the mechanical resonance of a vibrating crystal of piezoelectric material to create an electrical signal with a very precise frequency. Iit is thus possible to obtain stabilised frequencies, that, however are influenced by temperature. In order to cope with temperature instabilities various forms of compensation are used, from analog compensation (TCXO) to stabilisation of the temperature through a crystal oven (OCXO)[5]. OCXO are the most stable oscillators, as they are actively stabilised in temperature. TCXO on the contrary has a sort of feedback mechanism that realigns frequencies when, due to temperature variations, they jump beyond predetermined thresholds. In Fig. 3.8 there are two example measurements of GPS signals we have processed from two receivers of the GPS network of the Regione Toscana, mounting these two different kinds of oscillators. We have used differential measurements with a time lag of $\Delta t = 1\,s$ of pseudorange ρ and satellite distance from the receiver R, in order to eliminate (at a first order) slow varying contributes (due to atmosphere, antennas, biases, etc.). We have also computed a ionospheric-free signal and from this a "total" bias B less the tropospheric delay, $c\,\tau_{Trop}$, that we expected to possess a short term stability. B is the sum of three terms, the first consists of known parameters, namely the interfrequency bias B_{2-1} and the ratio of the square frequencies, γ; the second is the clock bias, b_c, and the third is the signal processing bias, $b(f_2)$, that are unknown. We can easily observe the dramatic difference in stability of the two oscillators, and the large realignment jumps of $100\,ns$ (equivalent to about $30\,m$) for TCXO measurements, that brings to a very unstable bias, much greater than precision

[5]TCXO stays for Temperature Controlled Crystal Oscillator; OCXO stays for Oven Controlled Crystal Oscillator (being "XO" an old acronym for "crystal oscillator").

constraints for many applications (our included). This means that single TCXO measures are essentially unusable, and it is necessary to process reasonable ensembles of measurements in order to obtain enough precise results at least for positioning.

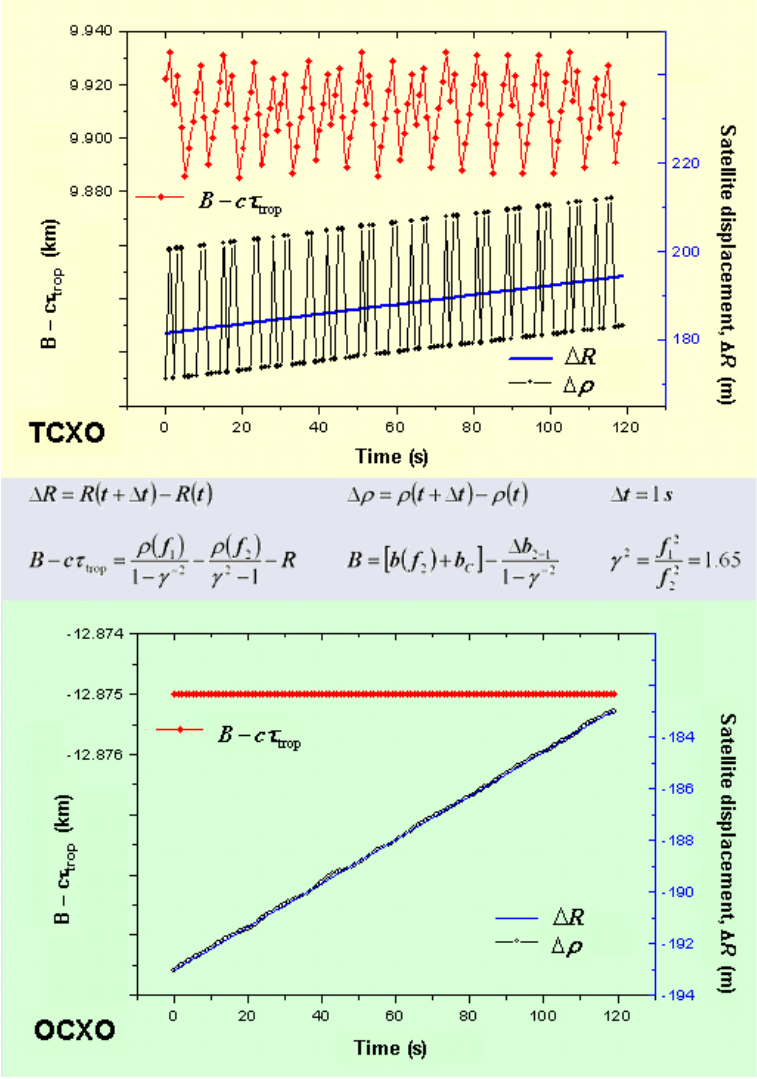


Figure 3.8: Stability of TCXO and OCXO clock crystals (top and bottom panels respectively). R is the satellite distance from the receiver, ρ the pseudorange, B is the "total" bias, containing the interfrequency bias B_{2-1}, the clock bias b_c, and the signal processing bias $b(f_2)$; $c\,\tau_{Trop}$ is the tropospheric delay. Figures are from two different receivers of the GPS network in Tuscany.

3.4.9 Multipath

If the GPS signal reaches the receiver antenna after reflections from obstacles, or from the Earth surface, we have the "multipath" phenomenon. The presence of indirect signals, that arrive later than direct ones and interfere with them, causes waveform distortion but also phase distortion. Multipath errors affect both carrier phase and pseudorange measurements, but the effects in the carrier phase measurement is limited to a quarter of a cycle and in the pseudorange measurements can reach several tens of metres (in C/A code measurements). The problem is that the error on pseudorange measurements can degrade the ambiguity resolution process required for using carrier phase measurements. Limiting the analysis to the effects on carrier phase the error derived from multipath is maximum when vectors of the direct and reflected signals are perpendicular. It does not exist a completely safe mitigation technique for the multipath effects, but a series of efficient and commonly used methods, primarily based on receiver hardware solutions. The use of differential GPS cannot reduce the multipath effects. Multipath effects can be detected if dual frequency observation are available (after compensation for ionospheric effects). The best ways to mitigate multipath effects are:

- to choose an installation site with a limited number of near potential reflecting objects;
- to use a groundplane antenna (i.e. chock ring) able to attenuate signal reflected from the ground;
- to use a directive antenna array with many directive patterns simultaneously operative and able to adapting to changing satellites geometry;
- to use an hardware (antenna gain) or software (signal processing) selected "cutoff" angle (i.e. minimum desired elevation angle) to eliminate measurements with low satellite elevation angles which have higher probability to generate multipath;

- to use a polarised antenna able to match the polarisation of the GPS signal (i.e., right hand) but mitigating in this case only single reflection or odd number of them;

- to observe the signal for a longer time than strictly needed, in order to detect sudden changes caused by multipath (only for fixed stations);

- to use decoding techniques based on narrow correlation of the receiver that allow to suppress automatically signals delayed by more than 1.5 chips (a chip correspond to 1 pulse: about $1\,\mu s$ for the C/A-code and about $100\,ns$ for the P-code);

- to analyse the shape of the correlation function.

Other methods exist but they will not be analysed in this work. Is is sufficient to mention that receivers with *chock ring* antenna have been chosen for all the used fixed station and they have been installed over roofs of buildings and far from sources of reflections.

3.4.10 Additional error sources

In addition to the errors discussed in the previous sections, there are a number of additional error sources whose effects are generally of secondary importance. The main contributions of these secondary sources can be however corrected in order to minimise the further uncertainties they can introduce. The variation of the antenna centres is one of these. This effect is well known and it has a straightforward impact on the variation of the signal travel length. Another well known source is due to tidal effects, namely the Earth tide and the ocean loading effects. Inland water loading is another one (e.g. cf. [44]). If not corrected they can contribute up to a few mm of errors or even more[6], that for atmospheric applications are still relevant. When corrected, thanks to well calibrated public models, their contribution to error is essentially negligible.

[6]In some regions of the planet tidal effects can reach tens of centimetres.

From what shown up to now in this chapter, we can understand how single quasi real-time measurements of GPS delays have errors at least of a few centimetres. However approaches based on simultaneous processing of several observations, better if involving one or more reference fixed stations with precise receivers, can improve the accuracy up to an order of magnitude.

3.5 Other GNSS systems

Satellite navigation has become a matter of main interest for several countries. A number of systems with global or local navigation capability are already operational or on experimental stage. In the following we give some details on two systems of main interest for their characteristics and capability of integration with GPS: the Russian GLONASS and the European GALILEO.

3.5.1 GLONASS

GLONASS (GLObal NAvigation Satellite System) is an alternative to the US GPS system. As GPS, it is composed by two main services: the military one and the civil one, with civil services and applications quality-degraded. The full constellation is composed by 24 satellites (excluding spare ones) equally spaced over 3 orbital planes. Satellites orbit are near circular at altitude of 19100 km over the Earth surface. Orbital planes are inclined 64.8° on the equatorial plane. Revolution time is 11 hours and 15 minutes.

The ground control stations of the GLONASS are maintained only in the territory of the former Soviet Union, thus limiting the global coverage capability of the system. The satellite coverage of GLONASS is greater at northern latitudes, where GPS system has minimum coverage. In the signal transmission the GLONASS system uses a spread spectrum technology. The signal are right-hand circularly polarized and the modulation type is the BPSK (Binary Phase Shift Keying). The code is the same for all the satellites, but each satellite transmit in a different frequency using a total of 25-channels to realize a Frequency Division

Multiple Access (FDMA)[7].

Two main frequencies are used for navigation: L1 (1602.0 MHz) and L2 (1246 MHz). The central frequency relative to each satellite carrier can be obtained as follow:

$$\begin{aligned} L1 &= 1602 + n \times 0.5625 MHz \\ L2 &= 1246 + n \times 0.4375 MHz \end{aligned} \tag{3.32}$$

with $n = -12, ...0, ..., 12$.

GLONASS satellites transmit two types of signals: a standard precision (SP) signal and a so called "obfuscated" high precision (HP) signal. The chipping rates for the HP and SP codes are 5.11 and 0.511 $Mbps$, respectively. The HP signal is broadcast in phase quadrature with the SP signal, has a ten time larger bandwidth and is available only for authorised users. As with GPS also with GLONASS it is possible to make phase measurements through carrier signals.

The navigation message (50 bps) contains the parameters for computing spatial and temporal coordinates of each satellite. The combination of measurements and navigation messages allows to determine the position coordinates, the speed vectors and the time of the receiver.

GLONASS has been operational since 1982. Its operational history is linked the URSS political and economic events. The first modernisation plan is the GLONASS-M from 2003; the second modernization plan, GLONASS-K, is being implemented from 2011: the first GLONASS-K satellite has already been launched, on the 26$^{\text{th}}$ of February 2011. At present civilian GLONASS is a bit less accurate than GPS, but the GLONASS-K system will double the accuracy of the previous one.

3.5.2 Galileo

GALILEO is a joint initiative of ESA and EC. It is a global satellite navigation system designed to provide a multimodal service, in different domains. It is

[7]Satellites placed at the antipodes of an orbit use the same frequency: as they are never simultaneously visible, there is no possibility of signal source ambiguity.

conceived to be completely independent and autonomous, but consistent and interoperable with the American GPS system. The Galileo system is mainly a civil initiative and will develop civil applications at different level:

Open Service (OS)

Commercial Service (CS)

Safety Of Life (SOL)

Public Regulated Service (PRS)

Search & Rescue (SAR)

The Space segment consists of a constellation of 30 Medium Earth Orbit (MEO) satellites with orbit at height of 23616 km. Satellite are planned to be placed in 3 orbital plane with an inclination of 56° over the terrestrial equatorial plane. The orbital period is 14 hours and 22 minutes.

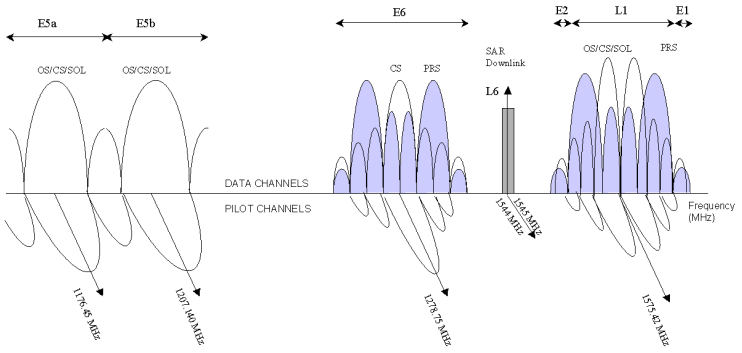


Figure 3.9: Structure of the Galileo Navigation Signal (image from http://www.esa.int/esaNA/SEM86CSMD6E_galileo_1.html).

Similarly to the GPS case the Galileo system will be provided of Ranging Codes (RC). In order to allow the completion of all services each Galileo satellite will broadcast 10 different navigation signals and 1 Search And Rescue signal.

Open access RC: E5a (I), E5a (Q), E5b (I), E5b (Q),E2-L1-E1 (B), E2-L1-E1(C)

RC with commercial encryption: E6 (B), E6 (C)

RC with governmental encryption: E6 (A), E2-L1-E1(A)

The Galileo signal bands will be:

$1164 - 1215\ MHz$

$1260 - 1300\ MHz$

$1559 - 1591\ MHz$

Different codes for different satellite reception is possible using CDMA. Different signals are broadcast on the in-phase (I) and in quadrature (Q) channels and, in the case of the $1164 \div 1215\ MHz$, different signals are provided in the upper (E5b) and lower (E5a) part of the band.

The first testing platforms, GIOVE-A and GIOVE-B have been launched in 2005 and in 2008 respectively. In a second moment four IOV (In Orbit Validation) Galileo satellites will be launched starting from April 2011. These satellites will be very similar to the final satellite system configuration. The Full Operational Capability will be completed with the launch of 30 operational satellites.

3.5.3 Complementarity and interoperability

The fundamental frequency of both GPS and GLONASS and Galileo system is the same: $f_0 = 10.23\ MHz$. This simplifies the design and construction of receivers that can detect all signals coming from all satellite systems. An advantage of simultaneous processing of several observations, especially for retrieving atmospheric parameters, is that such data redundancy contributes to better remove oscillation, drift, biases etc., and generally a large number of satellite on view improve the measurement accuracy. As a consequence the perspectives for GNSS "non-native" applications, such as the atmospheric one, so demanding in terms of accuracy and number of observations, become more and more interesting. The future development of GLONASS, GPS and Galileo are on line with these expectations, not only for the upgrading on single satellite signals, but also

because they are planned to ensure the interoperability among the various systems, for example foreseeing signals sharing by satellites belonging to different systems.

This trend involves also other global navigation systems, extending also to systems under development, as the Chinese Beidou and the Indian IRNSS systems, conceived as regional systems but on the way to be extended towards a global coverage.

Chapter 4

Retrieval of atmospheric profiles

4.1 Classical approaches for GNSS Water Vapour retrieval

In the present chapter we will assume that we have a GPS (or more generally a GNSS) precision signal delay already "cleaned" from ionospheric effects, various biases and the distance between the receiver and the satellite. The delay $\delta l = c\delta t$, thus measured as a distance, will be only due to the neutral atmospheric effects, that is mainly (around 90%) due to the troposphere.

In § 2.2 we have seen that the refractive index for the neutral component of the atmosphere can be expressed as:

$$n = n(z) = n\left(P_d(z), T(z), e(z)\right) \tag{4.1}$$

It is useful to introduce $\mathcal{N}$ defined as $\mathcal{N} = n - 1$, that, from Eq. (2.42), and neglecting the contribution for non ideal-gas behaviours (i.e. assuming $Z_d \simeq Z_w \simeq 1$), can be written as:

$$\mathcal{N} = c_1 \frac{P_d}{T} + \left[c_2 + \frac{c_3}{T}\right] \frac{e}{T} \tag{4.2}$$

From what discussed in § 2.3, the signal delay, δl, will be:

$$\delta l = \int_L \mathcal{N} ds \tag{4.3}$$

being L the ray path, that we assume coincident with the straight line connecting the satellite to the receiver, committing a negligible error, as shown in § 2.3.3.

Thanks to the integral mean value theorem[1], from δl we can immediately obtain an average measure of $\mathcal{N}$ along L. If we want to get more than this from GNSS delays, we have to include additional information and/or assumptions.

A method that can be considered a sort of "classical way" to proceed in order to extract at least an average estimation of water vapour (i.e. of e) or related quantities, will be illustrated in the following of this section. The aim of such method is to retrieve values of precipitable water at the receiver coordinates, using GNSS delays and local *in-situ* pressure and temperature measurements. Precipitable water is the depth of the amount of water in a column of the atmosphere, as if all the water in that column were precipitated as rain[2]. The relevance of the method lays on the fact that it is rather simple, that in addition to GNSS measurements it needs just basic atmospheric parameters as given by common meteorological stations (as normally installed close to GNSS fixed stations) and that it produces a parameter of main meteorological interest, as precipitable water is. For this reasons it is at the basis of a number of works on GNSS atmospheric applications, whose differences are more on preprocessing methods to obtain δl, than on the core of the precipitable water retrieval. The main limitations are in some assumptions that are adopted and that will be commented step by step, and in the information that is virtually wasted and that has been the primary reason from which the present work has been originated.

The basic idea is to partition the total atmospheric delay into a large quantity which depends only on the total ground pressure, called the "hydrostatic delay", and a smaller quantity which is a function of water vapour distribution and that

[1]The integral mean value theorem states that, if $f : [a, b] \to \mathbb{R}$ is a continuous function on the closed interval $[a, b]$, and differentiable on the open interval $]a, b[$, where $a < b$, then there exists some $c \in]a, b[$ such that:

$$f(c) = \frac{1}{b-a} \int_a^b f(x)dx \, .$$

[2]Generally water in an atmospheric column can be present as water vapour, liquid water and ice crystals. The GNSS signal delay however depends on vapour and few on liquid water or ice. Thus we will assume contributions to the total water other than vapour to be negligible, knowing that this is not always verified.

is called the "wet delay" ([35]; [13]).

Let us consider only zenith observations of the delays, using ZHD for the zenith hydrostatic delay and ZWD for the zenith wet delay.

A semi-empirical relationship between ZHD and the total pressure, P_s at the Earth's surface, can be assumed, as in [17]:

$$\Delta L_h^0 = \text{ZHD} = (2.2779 \pm 0.0024)\, \frac{P_s}{f(\lambda, H)} \tag{4.4}$$

where ZHD is in millimetres, P_s in millibars, and:

$$f(\lambda, H) = 1 - 0.00266 \cos(2\lambda) - 0.00028\, H \tag{4.5}$$

accounts for the variation in gravitational acceleration with the latitude λ and the height H, expressed in kilometres. The hydrostatic delay in Eq. (4.4) is not exactly the "dry delay", as it is given in terms of the total pressure and not on the partial pressure of dry air. We could say that ZHD is made by all the parts of $\mathcal{N}$ that depend on P/T, included the non-polar contribution from water vapour, the largest contribution to the hydrostatic delay remaining however that of the dry air. Wet delay consequently refers to the component produced mainly by the atmospheric water vapour, due to the dipole component of its refractivity.

Then ZWD is given by:

$$\Delta L_w^0 = \text{ZWD} = c_2' \int \frac{e}{T}\, dz \, + c_3 \int \frac{e}{T^2}\, dz \tag{4.6}$$

where $c_2' = (17 \pm 10) \cdot 10^{-6}\, K mbar^{-1}$ is different from c_2 in Eq. (2.42) (because of the part included in the ZHD), while c_3 is the same (cf. [13]. Eq. (4.6) is usually approximated to:

$$\Delta L_w^0 = \mathcal{C} \int \frac{e}{T^2}\, dz \tag{4.7}$$

with $\mathcal{C} = 0.382 \pm 0.004\, K^2 mbar^{-1}$

Now we have a relationship between ZHD and P_s that is considered fairly good, while GNSS data are necessary to evaluate ZWD, for which no accurate relationship with solely ground parameters is achievable.

In order to be able to link a GNSS observation of tropospheric delay to ZHD and ZWD, we need to map delays along paths with arbitrary elevation angles into zenith delays. Mapping functions are summoned at this purpose. The total delay for a path with an elevation angle α [3], is computed from the hydrostatic and wet zenith delays via:

$$\Delta L = \Delta L_h^0 M_h(\alpha) + \Delta L_w^0 M_w(\alpha) \tag{4.8}$$

where $M_h(\alpha)$ and $M_w(\alpha)$ are the hydrostatic and wet mapping functions respectively. Various forms have been proposed for the wet and hydrostatic mapping functions (e.g. Chao; [13]). These functions differ in the number of meteorological parameters that are incorporated, some simply referring to geographical features, other including meteo/climatic ones. Because hydrostatic and wet mapping functions are similar above elevation angles of 15° and GPS observations at lower elevations are rarely used, they are jointly estimated, grouping the two delays together ([41]). Nearly all of the mapping functions that have been suggested in the literature assume no azimuth variation in path delay, but it is known that the assumption of azimuth symmetry may cause significant errors (some centimetres) when the local troposphere has large horizontal temperature, pressure, or humidity gradients, as usually come with strong frontal weather systems ([12]).

Now, if we assume to have found an expression for the mapping functions that satisfy our needs, we still need to derive an approximate relationship between the vertically integrated water vapour (IWV) and an observed zenith wet delay. We can then introduce the weighted "mean temperature" of the atmosphere, T_m, defined by (cf. [13]):

$$T_m = \frac{\int \frac{e}{T}\, dz}{\int \frac{e}{T^2}\, dz} \tag{4.9}$$

Combining Eq. (4.6) with Eq. (4.9), and the equation of state for water vapour we obtain:

$$\mathrm{IWV} = \int \rho_w dz \simeq k \Delta L_w^0 \tag{4.10}$$

[3]Note that hereafter we use α for indicating what in Fig. 2.7 was the complementary of α_1.

where ΔL_w^0 is the zenith wet delay, and k is given by (cf. [4]):

$$1/k = \left(\frac{c_3}{T_m} + c_2' \right) R_w \tag{4.11}$$

where R_w is the specific gas constant for water vapour (see § 1.2.1). It is IWV that sometimes is stated as the height of an equivalent column of liquid water, i.e. the precipitable water (PW). Numerically, the IWV is just the product of PW and ρ_l, the latter being the density of liquid water. PW as ZWD have units of length and we have:

$$\mathrm{PW} = \frac{k}{\rho_l} \mathrm{ZWD} \tag{4.12}$$

From Eq. (4.11) we observe that an estimation of T_m is necessary for retrieving PW. This could be done, for example, by statistical analysis of a large number of radiosonde profiles, in order to obtain a model of T_m "tuned" to a specific area and season. This gives very rough estimation of T_m and a more reliable approach can be to use operational meteorological models to predict its actual value. Alternatively T_m can be estimated on the basis of the observed ground temperature, using suitable atmospheric profiles calibrated to the measurement point(s). A number of different solutions have been tried for this issue (e.g. cf. [4]), and for some of them the error on the final estimation of IWV is claimed to be a few percents. However these error estimates are generally the result of posterior validations, made on more or less averaged quantities, while the number of assumptions on the atmospheric structure, and the equation simplifications we have previously listed (plus all the uncertainties in the tropospheric delay estimation), makes error on single measurements realistically much higher and also very variable and little predictable.

The lack of knowledge of reliable error on single measurements, related to the specific measurement conditions, is one of the weak point of all approaches belonging to the "family" of the described one. Another weak point is in the forcing of the problem solution to a zenith integrated estimation of average atmospheric parameters. On one side this implies assuming symmetries and homogeneities for the atmosphere that generally are not verified. On the other side it prevents

the exploitation of the horizontal and vertical information contained in the numerous simultaneous GNSS observations from different satellites and different stations, whose interest will increase more and more in the future thanks to the increasing number of operational platforms, satellite constellations and ground stations, most of the latter installed for different purposes , but that can become a precious resource also in atmospheric science.

This high number of different profiles could in principle open the possibility to a tomographic retrieval of the main atmospheric parameters (the one $\mathcal{N}$ depends on). In the following section we will analyse which are the limits of applicability of the classical tomographic approach with respect to our issue, in order to propose and test an alternative probabilistic approach, that will be described in the second part of the present chapter.

4.2 Classical tomographic recontruction problem

The intensity I of a signal crossing a layer dl of absorbent material is reduced by an amount dI proportional to the product Idl [4], that is:

$$dI = -\kappa I dl \tag{4.13}$$

where κ is the linear absorption coefficient of the medium. By integrating Eq. (4.13) along the path L, followed by the signal inside the absorbent medium, we obtain:

$$I_L = I_0 \exp\left(-\int_L \kappa \, dl\right) \tag{4.14}$$

which reduces to the law of Beer-Lambert, and simplifies in an exponential decay for the intensity, if the absorption coefficient remains constant along the entire path. Therefore, if a signal of known intensity I_0 is sent through an absorbent medium, the measure of the emerging intensity I_L, allows to immediately retrieve

[4] This is true within the limit of validity of linear absorption.

the integrated absorption coefficient:

$$\int_L \kappa \, dl = \ln \frac{I_0}{I_L} \tag{4.15}$$

The basic data for a tomographic reconstruction are just a set of values expressing line integrals of the type (4.15), characterising the features of a medium.

Signals can be radio waves, visible light, beam of X-rays or neutrons, acoustic waves etc., and the basic integrated feature has not to be necessarily the absorption coefficient, thus it can be something not necessarily derived from measurements of absorption. In this work, for instance, we deal with integrated path delays of MW L-band radiation, and the feature of interest are "local" delays, i.e. locally averaged refractive indexes n, from which retrieving basic thermodynamic parameters for the troposphere, namely partial pressure of the dry gas mixture, partial pressure of water vapour, and temperature (or some derived quantities).

In the following we will denote by n instead of κ a general feature to be determined. This is not only for similarity to our problem, but also and primarily to avoid confusion with the conjugate variable of the spatial coordinates (x, y, z), typically denoted by (k_x, k_y, k_z).

With these notation, the general problem of tomography is to reconstruct the spatial distribution of the feature $n(x, y, z)$ characterising a medium in a certain volume V from a set of line integrals of the form (4.15) along different paths L. For simplicity we will consider a two-dimensional problem (but results can be directly generalised to the three dimensional case). The medium thus extends in the plane identified by the Cartesian reference system (x, z). We suppose to be able to measures integrated quantities along straight lines $L_\alpha(s)$, with α indicating the elevation angle with respect to the x axis and s a coordinate along the direction perpendicular to the lines themselves, as shown in Fig. 4.1.

The following function:

$$f(\alpha, s) = \int_{L_\alpha(s)} n(x, z) dl \tag{4.16}$$

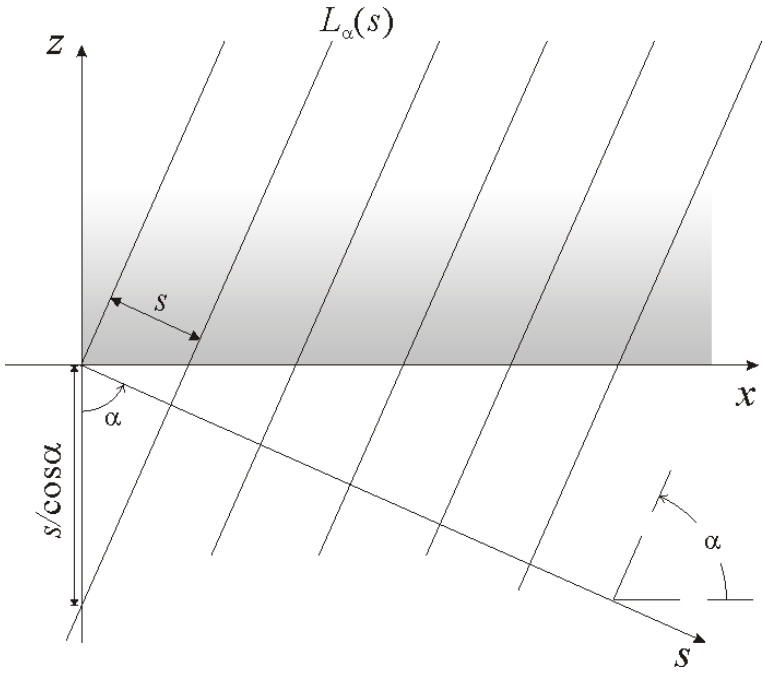


Figure 4.1: Geometry of data acquisition in a tomography problem.

is called Radon transform of the function $n(x, z)$, after the Austrian mathematician Johann Radon, who first introduced it. It is essentially the basic term of tomography. In fact, the equation of the general line $L_\alpha(s)$ is:

$$z = x \tan\alpha - \frac{s}{\cos\alpha} \equiv \frac{x \sin\alpha - s}{\cos\alpha} \tag{4.17}$$

and as $dx = dl \cos\alpha$, Eq. (4.16) can be written as:

$$f(\alpha, s) = \int_{-\infty}^{+\infty} n\left(x, \frac{x \sin\alpha - s}{\cos\alpha}\right) \frac{dx}{\cos\alpha} \tag{4.18}$$

Introducing the generalized Dirac δ function, the former can be rewritten as:

$$f(\alpha, s) = \int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} n(x, z) \frac{1}{\cos\alpha} \delta\left(\frac{x \sin\alpha - z \cos\alpha - s}{\cos\alpha}\right) dz \tag{4.19}$$

that, for the scale properties of Dirac δ functions, i.e. $\delta(x/a) = a\delta(x)$, becomes:

$$f(\alpha, s) = \int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} n(x, z)\, \delta(x \sin\alpha - z \cos\alpha - s)\, dz \tag{4.20}$$

If we now calculate the Fourier transform of $f(\alpha, s)$ with respect to the vari-

able s, we obtain for the main property of the Dirac δ:

$$\begin{aligned}\tilde{f}(\alpha, k_s) &\equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(\alpha, s) \exp(-ik_s s) ds \\ &= \frac{1}{\sqrt{2\pi}} \iint_{-\infty}^{+\infty} n(x, z) \exp(-ixk_s \sin\alpha + izk_s \cos\alpha) dx dz\end{aligned} \tag{4.21}$$

This expression, apart for a multiplication factor, is clearly the Fourier transform of $n(x, z)$ evaluated at the point of the conjugate space $(k_x = k_s \sin\alpha\,,\ k_z = -k_s \cos\alpha)$:

$$\tilde{f}(\alpha, k_s) = \sqrt{2\pi}\, \tilde{n}(k_s \sin\alpha, -k_s \cos\alpha) \tag{4.22}$$

For a given α angle and varying k_s, these points lay on the line $k_x = -k_z \tan\alpha$ of the conjugate space (k_x, k_z). This definitely solves the problem of reconstructing $n(x, z)$, just anti-transforming $\tilde{n}$ computed from $\tilde{f}$ through Eq. (4.22).

In real problems continuous features give place to discrete sampling, and consequently Fourier integrals become discrete Fourier transforms. The Nyquist-Shannon theorem (cf. [37]) assesses that a signal $s(t)$ with a limited band f_M can be univocally reconstructed through samples $s(n\Delta t)$, with $(n \in \mathbb{Z})$, taken at frequency $f_s = \frac{1}{\Delta t}$ if $f_s > 2f_M$. We can also interpret such theorem as establishing the relationship between sampling frequencies and domain dimensions in Fourier conjugate spaces. Coming back to our problem, thanks to the formal symmetry between Fourier transforms and anti-transforms, the Nyquist-Shannon theorem gives us the relationship between spatial resolutions $(\Delta x\,,\ \Delta z)$ and the amplitude of the frequency domains (k_x, k_z), that in turn is linked to the range dimensions for k_s and for the angle α [5].

There are two things that consequently comes out from the previous analysis, that says that classical tomography is feasible if:

[5]For the same symmetry reasons the intervals $(\Delta k_x, \Delta k_z)$ are linked to the dimension in $(x\,,\ z)$ of the medium to be sounded.

1. discrete Fourier transforms can be computed, that means that regular sampling (i.e. measurements) are necessary at several points for several angles;

2. the sample point density (i.e. the range of k_s) and the range of the α angle have to be large enough to resolve the target of interest, in other words they must define integration paths crossing in some points inside the area of the medium to be reconstructed with a desired resolution.

In practice a classical approach to tomographic reconstruction makes sense if we are in a condition similar to the one illustrated in the top panel of Fig. 4.2. Unfortunately a much more realistic view of what happens with GNSS paths in troposphere is given by the bottom panel of Fig. 4.2. The sketch wants to show that normally we have sparse and irregularly distributed measurements, made along paths that nearly never cross each other, except in few cases at limit tropospheric altitudes. In fact we already know that tropopause is around $12\,km$, that observations lower than 20° over the horizon are not reliable, and that typical distances of stations are grater that a few tens of kilometres.

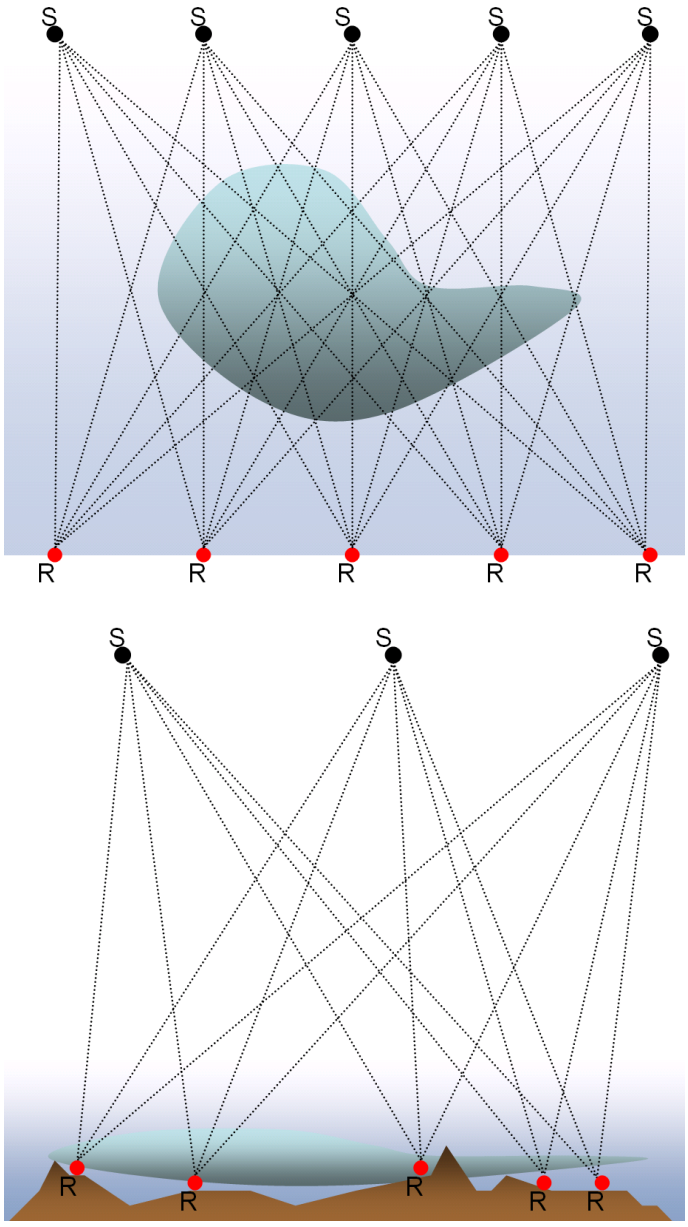


Figure 4.2: Sketches of GNSS measurement geometries. A theoretically ideal case for atmospheric tomography (top panel) and a much more realistic one (bottom panel), apart from the low number of satellites (S) in view (much lower than what already available), and the distance of satellites (not in scale). Note, in the latter case, the irregular distribution of GNSS receivers (R) and the thinness of troposphere (blue layer) with respect to the path distances and crosses.

We could argue that the irregular distribution of receivers could be overcome, assuming a regular network of virtual receivers whose measurements are given by some interpolation techniques of data from real receivers. If Fourier series are the target, they could own some properties that facilitate the issue: some suggestions could be searched in approaches like, for example, the one known as the Lomb method for spectral analyses of irregularly sampled data (cf. [26]). However interpolation methods bring always additional errors, that in our case should be carefully managed. Furthermore this would not solve the problem of scarcity of crossing profiles: since we are particularly interested in water vapour, it is the low troposphere that must be resolved with highest accuracy, exactly where we have no crossing paths. At the end we need more resolution (both vertical and horizontal) where we have less information from a classical tomographic point of view. This is equivalent to saying that the problem appears strongly underconstrained, and the approach that seems more natural is the probabilistic (Bayesian) one.

4.3 A probabilistic approach to atmospheric tomography

Assume that we have a set of state vectors for the profiles of the target parameters, and assume that this set is complete, meaning that it contains all possible states of the target parameters. This means also that such set forms an ideal *a priori* knowledge of the target parameters. it will be very dense to be easily expressed in terms of a probability density function for each target parameter[6].

Now let us imagine to have o set of measurements of observables depending on some of the target parameters in a known functional form, which can be assumed deterministic[7]. Each measurement can be considered as a constraint

[6] A more realistic view is that we will have a large number of possible values for the target parameters, such that this set of values can be considered representative, meaning that it reasonably represents our *a priori* knowledge, and that possibly is dense enough to easily allow its expression in terms of probability density functions.

[7] This equivalently means that the models that link the target parameters to the observables have negligible errors. If rigorously speaking this is not the case, we can imagine to cope with

with a tolerance given by its uncertainty[8] and it allows us to select, from our a priori dataset, only "state vectors" that satisfy all the available constraints, within to their tolerance.

So doing we obtain a new dataset of all possible state vectors, given the available measurements. The best set of target parameters is thus defined by the most probable state vector, and uncertainties are defined by the new probability density function for each target parameter (around its best value).

We now want to formalise these concepts according to our specific goals and we will do it gradually, starting from the simplified case of parameter retrieval in a one dimensional (1D) problem[9], and extending the process to three dimensional (3D) case.

4.3.1 1D retrievals

The 1D problem can be imagined as the retrieval of some zenith atmospheric parameters from measurements in a given point of the Earth surface. What described in § 4.1 is a special case of the more general 1D problem, where the target parameter is PW (obtained from the mean value of the wet part of the refractive index along the zenith path) and the observables are ground total pressure P_s, ground temperature T_s and the integrated zenith wet delay ZWD, measured from a fixed GNSS station equipped with meteorological sensors. As the target parameter is an integrated value along the vertical path, the 1D problem "collapses" into a 0D one (i.e. a point measurement).

Here instead, we want to address a more general problem of measuring $P(z)$, $T(z)$ and $e(z)$ from the same measurement configuration (a GNSS station with meteorological sensors). The observables for us are still P_s and T_s, but instead of

such problem "transferring" a model error (i.e. the probabilistic dependence of an observable on one or more target parameters) into the observable uncertainty.

[8] If uncertainties are negligible, i.e. the probability density functions for measurements are Dirac δ's, the tolerant constraint "collapses" in a rigid constraint.

[9] In the following of the present work, the notation 'nD' will be always used to identify a dimensionality in the real physical space and not in generic geometrical spaces, with $n \in \mathbb{N}$ and $n \in [0, 3]$.

ZWD we directly use the total tropospheric delay δl. Practically the continuous coordinate z is discretised in a number L of vertical levels z_i, according to the vertical resolution of the starting dataset (i.e. the *a priori* information). The target parameters become the finite set P_i, T_i, e_i, on each i-th layer.

We can indicate with $\mathbf{P}$, $\mathbf{T}$ and $\mathbf{e}$ the target parameter vectors of dimension L, and assume to have N realisations of such vectors, i.e. N atmospheric state vectors, constituting our *a priori*. The N vectors can be represented with a matrix of states, $\mathbf{S}$, as follows:

$$\mathbf{S} = \begin{pmatrix} P_{11} & T_{11} & e_{11} & \cdots & P_{1L} & T_{1L} & e_{1L} \\ P_{21} & T_{21} & e_{21} & \cdots & P_{2L} & T_{2L} & e_{2L} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ P_{N1} & T_{N1} & e_{N1} & \cdots & P_{NL} & T_{NL} & e_{NL} \end{pmatrix} \tag{4.23}$$

where each line defines a specific atmospheric state, in this case through a vertical profile for P, T and e.

The *a priori* information can then be expressed as a probability density function generated from Eq. (4.23):

$$\mathcal{P}(P_1\, T_1\, e_1 \cdots P_L\, T_L\, e_L \,|\, \Im) \tag{4.24}$$

where in this case $\Im$ contains the conditions related to environmental characteristics of the measurement place (i.e. the local climate), plus all other characteristics the dataset refers to (e.g. a specific season).

Let's use the vector notation as follows:

$$\begin{aligned} \mathbf{X} &\equiv \left[P_1\,,\, T_1\,,\, e_1\,,\, \cdots\,,\, P_L\,,\, T_L\,,\, e_L\right] \\ \mathbf{O} &\equiv \left[O_1\,,\, \cdots\,,\, O_N\right] \\ \tilde{\mathbf{O}} &\equiv \left[\tilde{O}_1\,,\, \cdots\,,\, \tilde{O}_N\right] \end{aligned} \tag{4.25}$$

$\mathbf{X}$ has dimension $3L$ and its components are all the target parameters. $\mathbf{O}$ and $\tilde{\mathbf{O}}$ have dimension M, the former components being the observable "true values" O_k (that are unknown), and the latter components being the value $\tilde{O}_k$ measured for the observable, i.e. the known measurement results, that in our specific case could be some P_s, T_s, or δl.

What we assume to know are $\mathcal{P}(\mathbf{X} \,|\, \Im)$, $\mathcal{P}(O_k \,|\, \mathbf{X}\, \Im)$, and $\mathcal{P}(O_k \,|\, \tilde{O}_k\, \Im)$. $\mathcal{P}(\mathbf{X} \,|\, \Im)$, is Eq. (4.24), i.e. the *a priori* information on the target parameters. $\mathcal{P}(O_k \,|\, \mathbf{X}\, \Im)$, is given by the model that links each observable to the target parameters, and that for us is deterministic through specific functional forms f_k, so that:

$$\mathcal{P}(O_k \,|\, \mathbf{X}\, \Im) = \delta\left(O_k - f_k(\mathbf{X})\, \Im\right) \tag{4.26}$$

$\mathcal{P}(O_k \,|\, \tilde{O}_k\, \Im)$, is the measurement errors. Our goal is to determine $\mathcal{P}(\mathbf{X} \,|\, \tilde{\mathbf{O}}\, \Im)$.

Thanks to the Bayes' theorem [10]we can write:

$$\begin{aligned} \mathcal{P}(\mathbf{X} \,|\, \tilde{\mathbf{O}}\, \Im) &= \\ &\stackrel{\mathcal{B}}{=} \mathcal{P}(\tilde{\mathbf{O}} \,|\, \mathbf{X}\, \Im) \cdot \frac{\mathcal{P}(\mathbf{X} \,|\, \Im)}{\mathcal{P}(\tilde{\mathbf{O}} \,|\, \Im)} \\ &= \left[\prod_k \mathcal{P}(\tilde{O}_k \,|\, \mathbf{X}\, \Im)\right] \cdot \frac{\mathcal{P}(\mathbf{X} \,|\, \Im)}{\mathcal{P}(\tilde{\mathbf{O}} \,|\, \Im)} \end{aligned} \tag{4.27}$$

With '$\stackrel{\mathcal{B}}{=}$' we have marked exactly where we have used the Bayes' theorem. The last (bottom right) term of Eq. (4.27) has been obtained considering that the result of each measurement is independent on the other measurements, and that this holds also for measurement prior probabilities.

In order to evaluate the terms into the square roots of Eq. (4.27), we can observe that:

$$\mathcal{P}(\tilde{O}_k \,|\, \mathbf{X}\, \Im) \stackrel{\mathcal{B}}{=} \mathcal{P}(\mathbf{X} \,|\, \tilde{O}_k\, \Im)\, \frac{\mathcal{P}(\tilde{O}_k \,|\, \Im)}{\mathcal{P}(\mathbf{X} \,|\, \Im)} \tag{4.28}$$

and that:

$$\begin{aligned} \mathcal{P}(\mathbf{X} \,|\, \tilde{O}_k\, \Im) &= \int \mathcal{P}(\mathbf{X} O_k \,|\, \tilde{O}_k\, \Im)\, dO \\ &= \int \mathcal{P}(O_k \,|\, \tilde{O}_k\, \Im) \mathcal{P}(\mathbf{X} \,|\, O_k \tilde{O}_k\, \Im)\, dO \end{aligned} \tag{4.29}$$

[10]The Bayes' theorem assesses that:

$$\mathcal{P}(A \,|\, B\, \Im) = \frac{\mathcal{P}(B \,|\, A\, \Im)\, \mathcal{P}(A \,|\, \Im)}{\mathcal{P}(B \,|\, \Im)}$$

where $P(A \,|\, \Im)$ is the prior probability of A (in the sense that it does not take into account any information about B); $P(A \,|\, B\, \Im)$ is the conditional probability of A, given B (sometimes referred as posterior probability because it depends upon the specified value of B); $P(B \,|\, A\, \Im)$ is the conditional probability of B, given A, also called the likelihood; $P(B \,|\, \Im)$ is the prior or marginal probability of B, that can be considered a simple normalising term.

Now we can further observe that $\mathbf{X}$ depends on $\tilde{\mathbf{O}}$ only through the true values $\mathbf{O}$ [11]. This means that we can write:

$$\begin{aligned} \mathcal{P}(\mathbf{X} \,|\, O_k \tilde{O}_k \,\Im) &= \mathcal{P}(\mathbf{X} \,|\, O_k \,\Im) && (4.30) \\ &\overset{\mathcal{B}}{=} \mathcal{P}(O_k \,|\, \mathbf{X} \,\Im)\, \frac{\mathcal{P}(\mathbf{X} \,|\, \Im)}{\mathcal{P}(O_k \,|\, \Im)} && (4.31) \end{aligned}$$

Substituting Eq. (4.31) into Eq. (4.29) we obtain:

$$\begin{aligned} \mathcal{P}(\mathbf{X} \,|\, \tilde{O}_k \,\Im) &= \mathcal{P}(\mathbf{X} \,|\, \Im) \int \frac{\mathcal{P}(O_k \,|\, \tilde{O}_k \,\Im)}{\mathcal{P}(O_k \,|\, \Im)}\, \mathcal{P}(O_k \,|\, \mathbf{X} \,\Im)\, dO \\ &= \mathcal{P}(\mathbf{X} \,|\, \Im) \int \frac{\mathcal{P}(O_k \,|\, \tilde{O}_k \,\Im)}{\mathcal{P}(O_k \,|\, \Im)}\, \delta\left(O_k - f_k(\mathbf{X})\right)\, dO \\ &= \mathcal{P}(\mathbf{X} \,|\, \Im)\, \frac{\mathcal{P}\left(O_k = f_k(\mathbf{X}) \,|\, \tilde{O}_k \,\Im\right)}{\mathcal{P}\left(O_k = f_k(\mathbf{X}) \,|\, \Im\right)} \qquad (4.32) \end{aligned}$$

Substituting then Eq. (4.32) into Eq. (4.28) we have:

$$\mathcal{P}(\tilde{O}_k \,|\, \mathbf{X} \,\Im) = \frac{\mathcal{P}\left(O_k = f_k(\mathbf{X}) \,|\, \tilde{O}_k \,\Im\right)}{\mathcal{P}\left(O_k = f_k(\mathbf{X}) \,|\, \Im\right)}\, \mathcal{P}(\tilde{O}_k \,|\, \Im) \qquad (4.33)$$

Finally using Eq. (4.33) into Eq. (4.27) we obtain:

$$\mathcal{P}(\mathbf{X} \,|\, \tilde{\mathbf{O}} \,\Im) = \frac{1}{\mathcal{K}} \left[\prod_k \mathcal{P}\left(O_k = f_k(\mathbf{X}) \,|\, \tilde{O}_k \,\Im\right) \cdot \mathcal{P}(\tilde{O}_k \,|\, \Im)\right] \mathcal{P}(\mathbf{X} \,|\, \Im) \qquad (4.34)$$

where $\mathcal{K}$ is a normalisation term which is given by the expression $\mathcal{K} = \left(\mathcal{P}(\tilde{\mathbf{O}} \,|\, \Im) \cdot \prod_k \mathcal{P}\left(O_k = f_k(\mathbf{X}) \,|\, \Im\right)\right)$, but that can be more easily computed imposing $\int \mathcal{P}(\mathbf{X} \,|\, \tilde{\mathbf{O}} \,\Im)\, d\mathbf{X} = 1$.

In order to reduce computing time, we have assumed for $\mathcal{P}(O_k \,|\, \tilde{O}_k \,\Im)$ a square shape, that means a constant value for a tolerance range[12] and 0 elsewhere:

$$\mathcal{P}(O_k \,|\, \tilde{O}_k) = \begin{cases} 1/\Lambda & \text{if } |O_k - \tilde{O}_k| \leq \Lambda/2 \\ 0 & \text{otherwise} \end{cases} \qquad (4.35)$$

[11] This allows to write Eq. (4.30) but of course in general we *cannot* write $\mathcal{P}(\mathbf{X} \,|\, \tilde{O}_k \,\Im) = \mathcal{P}(\mathbf{X} \,|\, O_k \,\Im)$.

[12] Given the range, the constant value for the square probability function remains defined by the probability normalisation condition, that imposes the integral (or sum over all states) equal to 1.

This simplify the implementation of the right part of Eq. (4.34), that has been done as a direct selection of all and only the states $\mathbf{X}$ whose observable values $\mathbf{O}$ fall within the $\mathcal{P}(\mathbf{O} \,|\, \tilde{\mathbf{O}}\, \Im)$ non-zero domain (i.e. the tolerance range)[13]. The relevance of the results that we will show later, is not impacted by this choice, that however, in real cases, could result too rough (a short discussion on this issue is postponed to § 5.2).

Compatible values for each target parameter are organised in histograms, in order to obtain a discrete probability density function. Histogram bins are sized in order to be about $\sqrt{N}$, ranging from the variables extremes[14]. At void bins (within the range) it is associated the minimum possible probability greater than 0, that is $1/N$, for all the void interval. This means that if the same variable has n_0 void bins, at each bin is associated the value $1/(N\, n_0)$.

Once we have the probability function for each target parameter we have still not finished the work, unless all parameters can be assumed independent. In fact if they are independent, the best state vector can be chosen as the vector of the most probable parameters, and errors are obtained from each probability functions, around their maxima[15]. Of course P, T and e cannot be considered independent variables, nor their values at different layers (especially for adjacent or near layers). It is thus necessary to move to independent variables.

Generally speaking this can be a complex problem, and it is not assured that a solution exists that is valid not only locally[16], but all over the parameter domains. A way to address this issue is however the method at the basis of the computation of the Empirical Orthogonal Functions (EOF, cf. [34]). It starts from the computation of the covariance matrix of the variables, whose diagonal

[13]From Eq. (4.34) we understand that we have arbitrarily assumed the term $\mathcal{P}(\tilde{O}_k \,|\, \Im)$ non influent (as it would be flat) in the range of measurement we deal with.

[14]This resembles the Poisson noise threshold, in the sense that we want our bin to contain $\sqrt{N}$ "events" on average, with bins more populated than this threshold to be considered very significant.

[15]This is rigorously true if we assume a unimodal distribution for probability. Bimodal or multimodal distributions would need specific analyses, but in effect regarding the parameters of our interest it is reasonably assume that multimodal distribution should come only as artefacts due to wrong assumptions in the retrieval process.

[16]The notion of "locality" here refers to a finite small interval of parameter values.

elements are then cancelled through a matrix diagonalisation, that corresponds to change (rotating) the starting reference system to a new one, defined by the eigenvectors of the covariance matrix.

A general definition of the covariance matrix $\mathbf{\Sigma} = \left[\sigma_{ij}\right]$ can be given as follows. Consider the vector:

$$\mathbf{X} = \left[X_1, \cdots, X_{3L}\right] \tag{4.36}$$

If its entries are random variables, each with finite variance, then the covariance matrix $\mathbf{\Sigma} = \left[\sigma_{ij}\right]$ is the matrix whose (i, j) element is the covariance:

$$\sigma_{ij} = \mathrm{cov}(X_i, X_j) = \mathrm{E}\left[(X_i - \mu_i)(X_j - \mu_j)\right] \tag{4.37}$$

where $\mu_i = \mathrm{E}(X_i)$ is the expected value of the ith entry in the vector $\mathbf{X}$.

In our case $\mathbf{X}$ is defined by Eq. (4.25), and the expected values are simply the mean values of (P_i, T_i, e_i).

Regarding the diagonalisation issue, the spectral theorem of linear algebra assesses that: $\exists\ \mathbf{V} \in \mathbb{M}\left(\mathbb{R},\, 3\,\mathbf{L} \times 3\,\mathbf{L}\right)$, which is orthogonal and satisfies:

$$\mathbf{\Sigma_d} = \mathbf{V^T} \cdot \mathbf{\Sigma} \cdot \mathbf{V} \tag{4.38}$$

with $\mathbf{\Sigma_d}$ diagonal. The eigenvalues gives the variances of the new generalised variables, with respect to the new reference axes.

By means of a linear transformation Eq. 4.38 allows to transform the $\mathbf{S}$, into the new matrix $\mathbf{S_{gen}}$, such that:

$$\begin{cases} \mathbf{S_{gen}} & = \mathbf{V^T} \cdot \mathbf{S^T} \\ \mathbf{S_{gen}} & \in \mathbb{M}\left(\mathbb{R},\, \mathbf{N} \times 3\mathbf{L}\right) \end{cases} \tag{4.39}$$

We obtain new state vectors in the phase space generated by the (orthogonal) eigenvectors of Σ. We will name these new state vectors *generalised state vectors* (see Fig. 4.3), meaning that they have no more a direct physical sense, because they do not express physical quantities at given levels (as it is in $\mathbf{S}$), but however they are linked to them through the simple linear transformation in Eq. (4.39).

The new matrix $\mathbf{S_{gen}}$ is as follows:

$$\mathbf{S_{gen}} = \begin{pmatrix} S_{11} & S_{12} & S_{13} & \cdots & S_{1\,(3L)} \\ S_{21} & S_{22} & S_{23} & \cdots & S_{2\,(3L)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ S_{N1} & S_{N2} & S_{N3} & \cdots & S_{N\,(3L)} \end{pmatrix} \tag{4.40}$$

If the distribution of the new target parameters S[17], that lead to a perfectly diagonal covariance matrix, is Gaussian, then their independence is assured. Thus we can compute their probability distribution as described before, and the best state vector is now correctly given by S's best values, at which corresponds a given set of target parameters

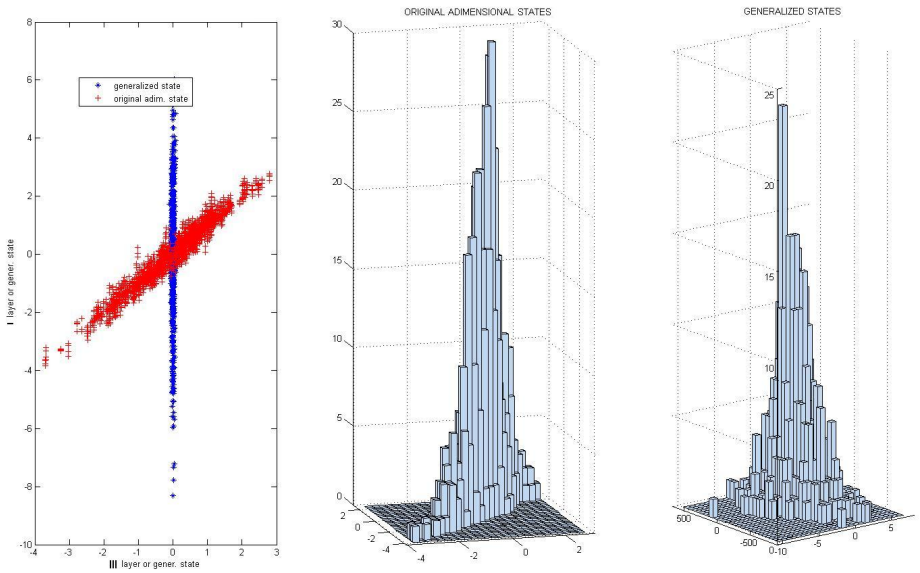


Figure 4.3: Example of correlated adimensional variables for two levels before and after orthogonalisation.

Covariance between non homogeneous variables does not make sense, because its value depends on the arbitrary choice of the units of measurements. Thus we have decided to proceed using adimensional variables obtained from the original ones dived by their standard deviation of the whole variable sample independent

[17]Hereafter with S free from subscripts, we indicate the generalised $3L$ variables linearly dependent on P, T, and e.

on the level. This means, for instance, that each P_i has been divided by the standard deviation computed on the ensemble $(P_1, \cdots, P_L)$, and the same for T_i and e_i. According to the structure of the matrix in Eq. (4.23), but following the notation of Eq. (4.37), we have:

$$X_{ij} \rightarrow X_{ij} \left(\frac{1}{L} \sqrt{\sum_{l=0}^{L} \sigma_{\{3l+[1+(j-1) \bmod 3]\}\{3l+[1+(j-1) \bmod 3]\}}} \right)^{-1} \tag{4.41}$$

where $[(j-1) \bmod 3] = 0, 1, 2$ flags P's, T's and e's respectively.

We have to note that the method chosen for the adimensionalisation could have critical impact one some results[18], if we used the EOF (orthogonalisation) process also to reduce the dimensionality of our problem, neglecting the components of the new S variables on the part of phase space mapped through eigenvectors corresponding to the lowest eigenvalues (i.e. forcing this phase space region to the null space). This however has no impact in this work as we have not reduced the dimensionality of our problem.

4.3.2 Extension to 3D retrievals

The extension to a 3D problem of what discussed in § 4.3.1 is theoretically straightforward. In fact if we consider to identify the horizontal spatial domain with a regular mesh, we will have a state vector with the same target parameters for each node. Our target parameters from $3L$ will become $3L \cdot M$, with M number of horizontal nodes. The matrix $\mathbf{S}$ of Eq. (4.23) will consequently increase to contain $3L \cdot M$ columns, but the process to obtain the matrix $\mathbf{S_{gen}}$ of Eq. (4.40) will obviously remain a matrix diagonalisation. Of course the processing time will grow up very rapidly with the number of nodes, and an analysis of the true

[18] Adimensional variables are scale invariant, but the same can be said for any analogous adimensionalisation process with an arbitrary multiplicative factor. Such factors can change the relative weight of the variable variances in the correlation matrix and so they can change eigenvalues. If we need EOF also to reduce the dimensionality of the problem, we must be aware of this arbitrariness. Defining variables that are physically homogeneous can solve the problem. For instance we could express all state vector variables in pressure units. It is straightforward for the total pressure and for the water vapour partial pressure. Temperature instead could be expressed by means of the saturation pressure of water vapour as given in Eq. (1.17).

dimensionality of the problem will be necessary in order to maintain the number of significant orthogonal variables into an acceptable limit.

Regarding the constraints in the retrieval process, no difference has to be introduced relatively to measurements of ground parameters. The same is also for zenith delay. Instead non zenith delay could be processed in order to include the effects of boundary or crossed cells, in other words the contributions of horizontal atmospheric gradients. This can be achieved through a proper spatial interpolation for the target parameters, which a given delay measurement depends on. A simple linear interpolation has been used in our numerical experiment will show in § 4.4.4, that can be considered more than reasonable along distances of a few kilometres, due to the "smoothness" of our target atmospheric parameters.

Of course non zenith delays should be compensated through a geometrical factor for the lengthening of their path due to the inclination. In our numerical experiments we have worked in a non curved tropospheric geometry and thus this term is just $1/\sin\alpha$, being α the elevation angle .

4.4 Numerical experiments

We have already stated that the principal aim of this work is to assess the feasibility of a probabilistic approach for retrieving basic atmospheric parameters, and the advantages in term of information that can be gained by such an approach. In order to address this issue we have decided to proceed through numerical experiments based on synthetic data, for working in a fully controlled environment. The reference area is Tuscany and its GNSS network of ground GPS stations[19] for precision positioning, mutually connected through the internet network of Regione Toscana, with an average distant between stations around $25\,km$.

[19]Three GPS stations are located and maintained by the LAMMA Consortium (www.lamma.rete.toscana.it), where this work has been developed.

4.4.1 Basic data

Our synthetic environment has been built starting from simulations of the Global Forecast System (GFS), the global numerical weather prediction computer model run by NOAA[20]. Such data has been used as basic *a priori* knowledge of the portion of the troposphere above a virtual receiver. We have imagined to differentiates the dataset due to the season and night and day time and we have accumulated values for a 4 year period. Data are each 6 hr; the horizontal resolution is about 50 km; and the vertical levels we have used are the first 20, beyond which no value of water vapour is given (as above the tropopause). Vertical levels are in pressure units and not in length one: we thus have $z = z(P)$ as free variable for given constant values of P. More precisely the model provides the geopotential height, $HGT(P)$ instead of $z(P)$. Among the various parameters available from GFS, we have worked with HGT, T and e, the last one computed from the relative humidity, thanks to Eq. (1.20) and to Eq. (1.17) (Fig 4.4).

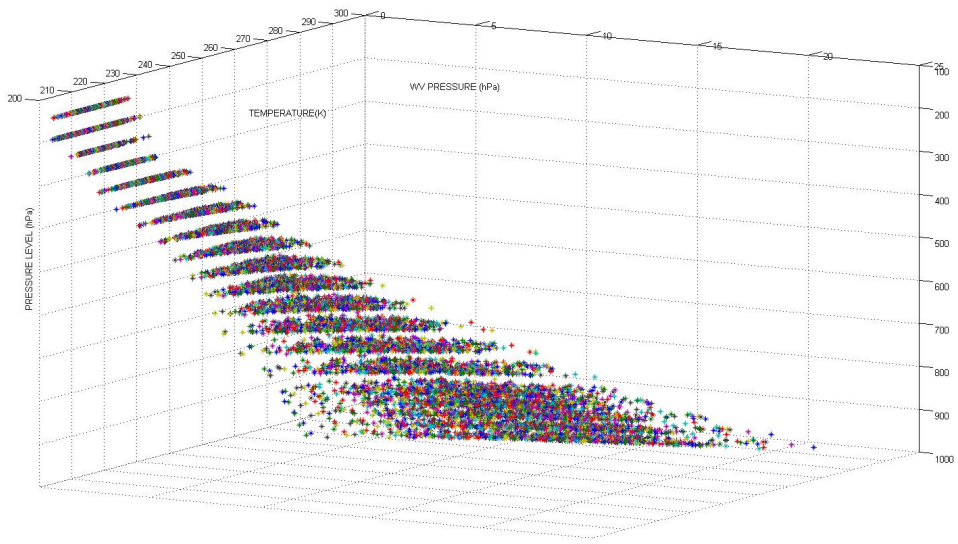


Figure 4.4: Distribution of temperature and water vapour pressure at different pressure levels.

[20]Data from http://nomad1.ncep.noaa.gov/ncep_data/index.html.

Through these data we have obtained a basic statistics to obtain the *a priori* $\mathcal{P}(\mathbf{X} \mid \Im)$, for the period of interest, relative to the target state variables, namely the isobaric levels of atmospheric geopotential height, temperature and water-vapour partial pressure (see Fig. 4.5), i.e. with:

$$\mathbf{X} \equiv \left[HGT_1\,,\, T_1\,,\, e_1\,,\, \cdots\,,\, HGT_L\,,\, T_L\,,\, e_L\right] \tag{4.42}$$

4.4.2 Generation of a synthetic database

The number of data from GFS simulations we accumulated for a given period (e.g. spring-daytime) was a bit more than 700 (namely 726) for each node. As prior information we would like to have something larger, especially if we imagine to drastically reduce this number due to the measurement constraints. As an indicative rule (but not always true as we will see in § 4.4.3) more efficient is the profile selection more informative is the set of measurements we have. Thus on one side this is desirable, but on the other side a few populated set of final states compliant with the constraints, makes an accurate definition of the retrieval error very difficult, as it prevents to build a reliable histogram[21]. What we need is thus a prior set of state vectors which is densely populated[22].

A way to overcome the limits of our "poor" prior dataset is to synthetically populate it . In fact if we look to the $\mathbf{S_{gen}}$ [23] of Eq. (4.40), we observe that being the $\left[S_{i1}\,,\, \cdots\,,\, S_{i\,(3L)}\right]$ variables that are independent each other ($\forall i$), any state vector made by $S_{p(i)\,j}$, with $p(i)$ random permutation of $i \in [1, N]$, is acceptable and has the same probability of any other. With such permutations in theory we can populate an *a priori* set of N initial state vectors of dimension $3L$, up to N^{3L} different ones (Fig. 4.7).

Now we need to clarify two points. The first one is that such process does not increment the prior information. No numeric technique in fact can do this

[21] If we want to be more precise about samples in a histogram, it is not so much the small number that prevent a reliable definition of the error, as the density.

[22] We have to note that in an operational scenario a high number of prior states is one of the factor that contributes to increment the processing time.

[23] Hereafter we will consider all state vectors and related matrixes generated with *HGT* instead of *P*.

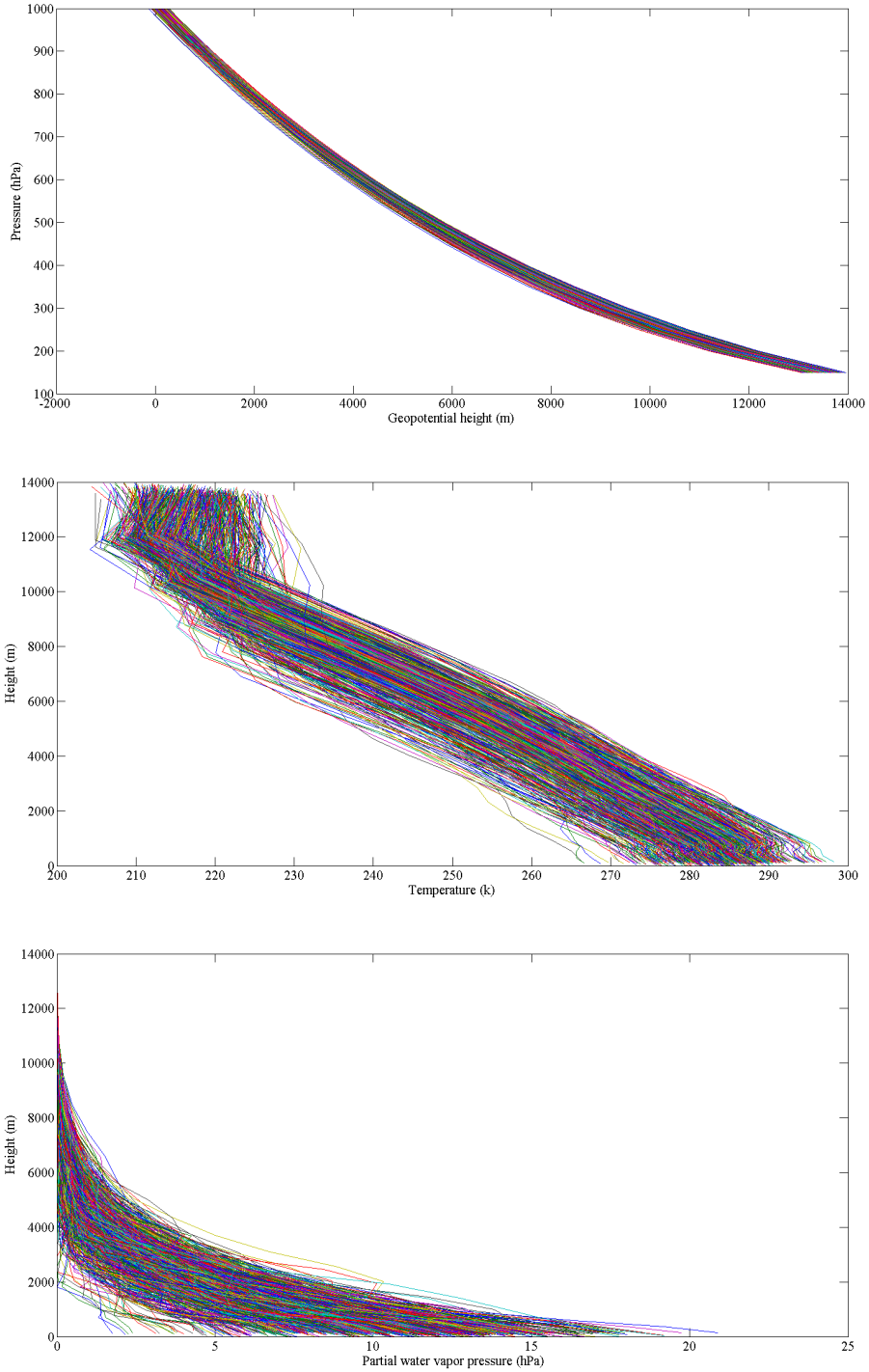


Figure 4.5: GFS profiles. The different colours refer to different profiles.

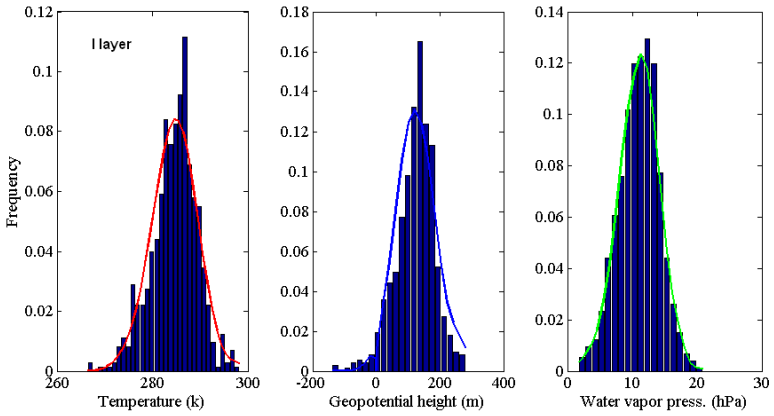


Figure 4.6: Distributions of original data (histograms) and synthetic ones (solid line). Note the smoother shape of the latter, due to the higher state density.

job and this growing of the prior dataset must be interpreted as a better way to compute the final probability distribution function for the target parameter than, for example, interpolating the bin values of a histogram, built with few bins and few states per bin[24]. The second one is that if the starting line variable S are not perfectly orthogonal (as discussed in § 4.3.1), their random combination can generate possible non physical states. This can be (partially) verified when the new $\mathbf{S_{gen}}$ is mapped back into the new $\mathbf{S}$, and if it happens the non physical states can be either rejected or eventually corrected. In our dataset for example a few slightly supersaturated states (see Fig. 4.8) have been generated, that have been corrected to saturated ones, but number and "violation" of such states was always marginal, providing a proof in favour of the applicability of the EOF technique to our problem.

[24]We should observe that an excessive growing of the dataset could artificially hide the problem of obtaining final histograms with few samples, that instead would clearly emerge using only the real prior information or a limited increment of states. In practice we should never exceed in filling the prior dataset with new synthetic states, bearing in mind that the technique is more effective when is less necessary (as any interpolation technique)!

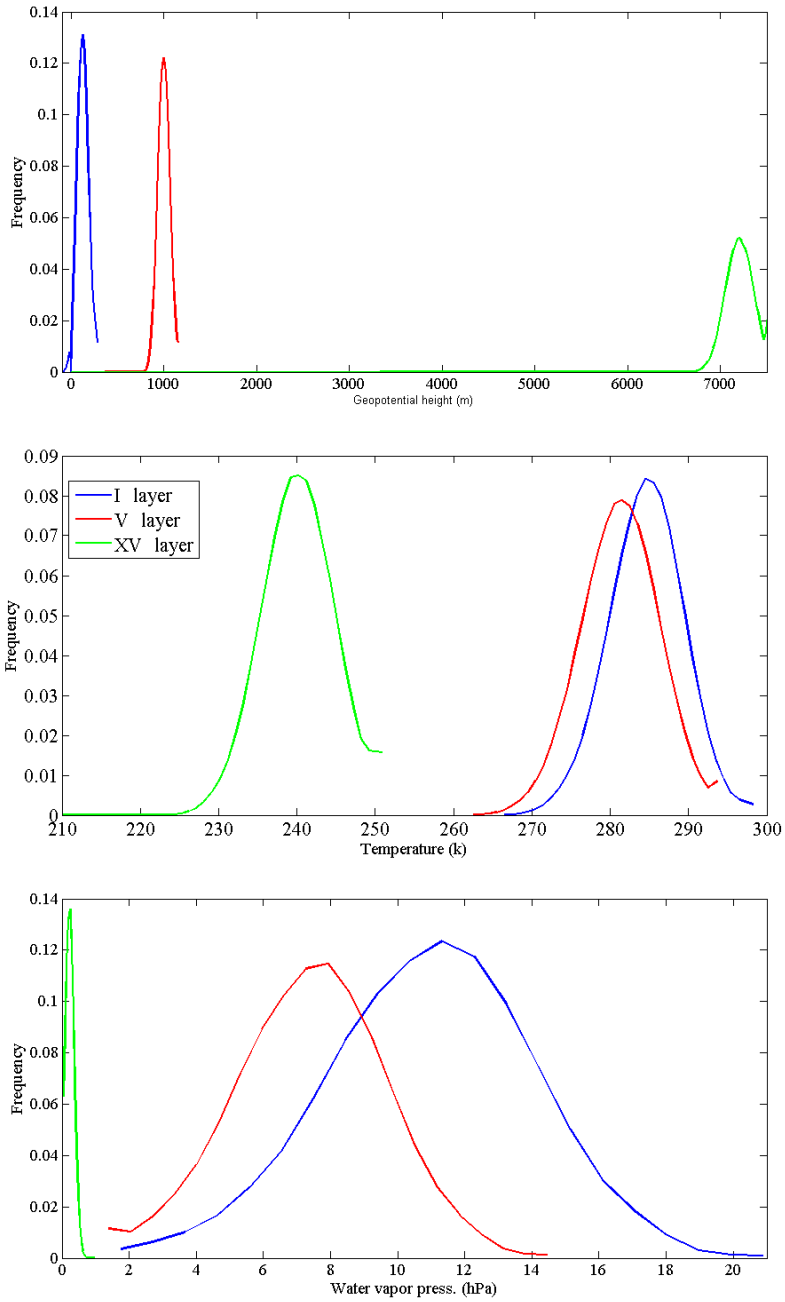


Figure 4.7: Distributions of synthetic data at different layers. Different colours mark different pressure layers, namely the 1st, the 5th and the 15th for the target parameters.

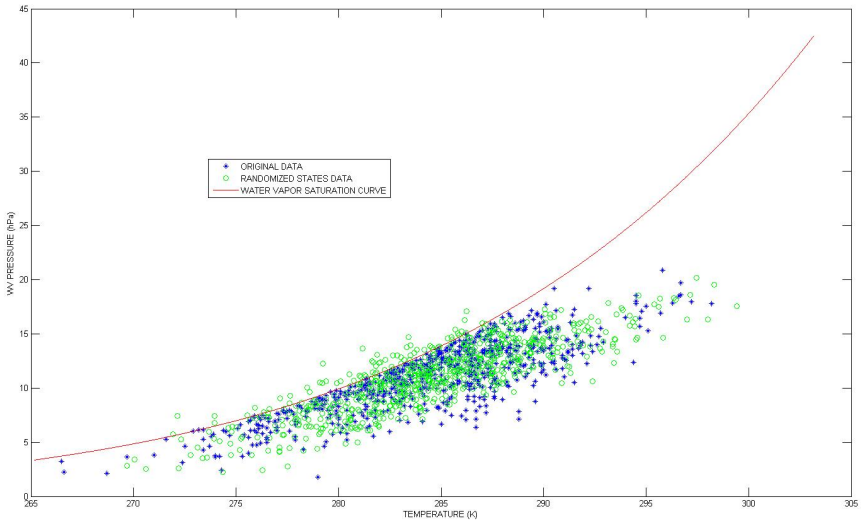


Figure 4.8: Artificial supersaturated states (above the red line) resulting from the generation of synthetic state vectors. The red threshold line has been plotted from Eq. (1.17).

4.4.3 Entropy and information

In a thermodynamic system entropy is defined as (cf. [49]):

$$dH = \frac{đQ}{T} \tag{4.43}$$

where dH is an infinitesimal variation of the entropy H, T the temperature and $đQ$ is the heat variation[25].

The interest in entropy is essentially on its variation, and it is known that in isolated systems entropy can only grows and that this can be interpreted as a tendency of the disorder to increase with time for such systems (our universe included). However locally, or equivalently in non isolated systems, entropy can decrease and can be used to measure the growth of order as a measure of the increment of information on the system. As an example, if we consider an

[25]We remind that $đ$ indicate a non exact differential for Q, that means that it is not the differential of any function of the thermodynamic variables. On the contrary dH is an exact differential, provided that Q is transmitted in a reversible way (e.g. through quasi-static transformations).

isothermal compression of an ideal gas of N molecules from the volume V_0 to the volume V_1, the variation of entropy is:

$$\Delta H = -Nk \ln \frac{V_0}{V_1} \tag{4.44}$$

being k the Boltzmann constant. Microscopically this isothermal decrease in the volume corresponds to a decrease in the number Ω of possible modes that can give the macroscopic state. In fact we can consider an arbitrarily small volume ΔV that contains a gas molecule, thus defining its position: so doing the number of "possible positions", that gives the number of the possible coordinate modes for a molecule, is given by $V/\Delta V$. The decrease in the number of possible modes corresponds to an increase in the information we have on the molecules, i.e. on the microscopic state of the system. Then for a molecule we could express the information $\mathcal{I}$ as:

$$\mathcal{I} = k \ln \frac{\Omega_0}{\Omega_1} = k \ln \frac{V_0/\Delta V}{V_1/\Delta V} = k \ln \frac{V_0}{V_1} \tag{4.45}$$

and for the whole gas of molecule we would have:

$$\mathcal{I} = Nk \ln \frac{\Omega_0}{\Omega_1} \tag{4.46}$$

that is, from Eq. (4.44), $\mathcal{I} = -\Delta H$

We can easily see an analogy between possible microscopic modes and possible state vectors in our problem, i.e. between the decrease of modes due to the thermodynamic transformation, bringing to the new volume constraint, and the selection of possible state vectors due to our measurement constraints (whose application represents our transformation). Such analogy has not to be literally assumed, as an increment of information is correctly assessed by changes in the shape of the probability density functions and not simply by the reduction of the number of compatible states. In fact, if we imagine a measurement that rejects a number states in a random way, it must not bring any information: accordingly the shape of the probability density function remains unchanged while the number of states diminish[26].

[26]This difference between an ideal simple thermodynamic system and our problem can be

With this in mind, we can then use the above variation of H to derive a measure of the information. The Gibbs theorem asses that in a system the whole entropy is the sum of partial entropies (e.g. of different gases). In this way Gibbs managed to formulate an expression for entropy, which is (cf. [34]):

$$H = -k \sum_i p_i \ln p_i \tag{4.47}$$

that is defined apart from additive constants (irrelevant in the estimation of entropy variations).

In the theory of information Shannon demonstrated that there is only a parameter that univocally respects all the requirements of logic consistency for the measurement of the information content for a distribution, and that parameter has the same expression of H, apart from a multiplicative factor, that have not a specific meaning as in thermodynamics, and that consequently can be arbitrarily chosen, as well as the basis of the logarithm operator (cf. [36]) .

Entropy and its variation are consequently the basic features we have evaluated to test the gain of information in our numerical experiments. Namely for any measurement, we have evaluated the gain of information both through the changes in the entropy defined as follows:

$$H = -\sum_i p_i \log p_i \tag{4.48}$$

and through the *relative entropy* H_{rel}, defined as:

$$H_{rel} = \sum_i p_i^C \log \frac{p_i^C}{p_i^A} \tag{4.49}$$

H_{rel} compares the difference between two distribution, A and C. As an example of application A could be the *a priori* probability density function and C the resulting one after the application of the measurement constraints. We can see $(-\Delta H)$ as a measure of the change in width between two distributions, i.e.

approximately explained by the fact that our knowledge of the possible states (before and after the constraint application) is given by the *a priori* dataset which: even in the most optimistic case, it is always much far than being "complete", and as a consequence it can be used only to infer probabilities and not to directly evaluate the number of possible states and its variation.

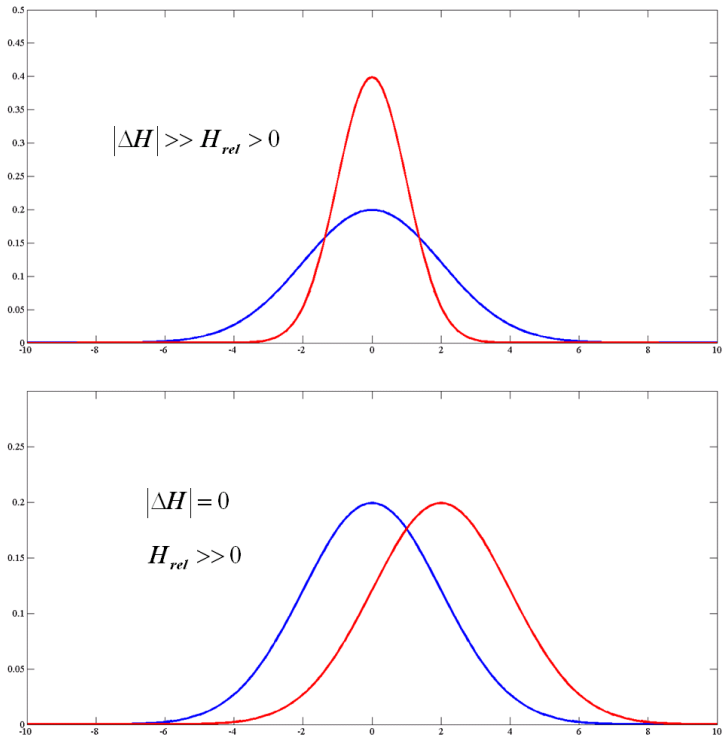


Figure 4.9: Effectiveness of absolute and relative entropy in measuring variations in distribution functions.

a measures of the gain in the *precision* of a target parameter. H_{rel}, instead, increases mainly due to differences in the two probability curves, e.g. due to centre displacements: in other words it is more a measure of a potential gain in the *accuracy* of a parameter, whose best value can be better retrieved thanks to a given measurement, even if eventually the measurement does not improve the parameter precision, i.e. the width of its probability distribution function (see Fig. 4.9).

4.4.4 Results of numerical experiments

The objective of this section is to show a selected number of results among the high number of the performed experiments, that we believe more representative for evaluating the gain of information we can obtain from given observation configurations.

Data are from the spring night-time dataset (but this is not so relevant), with 20000 synthetic state vectors. Experiments up to 500000 state vectors have been tried, obtaining coherent but slightly better results; however, from what discussed in § 4.4.2, we have opted for showing only results obtained with a lower number of prior states. In any case the difference of the entropy between synthetic datasets and the original one was found always negligible.

4.4.4.1 The 1D case

We will begin from a 1D problem, for which we have to imagine of taking measurements from a ground station, equipped with a meteo station providing P_s and T_s, and located approximately in the city of Florence (in Tuscany). Before discussing results, we want to observe from Fig. 4.10 the eigenvalues of the parameters covariance matrix for the *a priori* state vectors: we can see that the most of the variance is explained by the orthogonal functions (i.e. the eigenvectors) corresponding to the last few eigenvalues. From this we know that the dimensionality of our problem is much less than $3L$. This aspect should be considered in possible operational applications of the method, in order to reduce the computation time. However this result is very likely conditioned by the use of model data as prior information, moreover at low resolution[27], thus we imagine the dimensionality of the real case to be larger than this[28].

In order to achieve a significant evaluation of the obtainable information, the results that we are going to show are not of single (presumed representative) retrievals. In fact we have preferred to evaluate entropy and relative entropy as average values, resulting from a number of 25 numerical experiments, i.e. retrievals performed starting from 25 different state vectors. Such state vectors, assumed as "true" atmospheric states, have been chosen spanning all the range of T_1 (i.e. of the layer closer to the ground level) in a regular way. Experiments were

[27] In this case the only resolution is the vertical one, but for the 3D case also the low horizontal resolution can contribute to this aspect in a similar way.

[28] A correct estimation of the dimensionality could be performed with a large number of real measurements as prior dataset. In any case a careful estimation of the error associated to the prior dataset should be performed to evaluate the impact on eigenvalues.

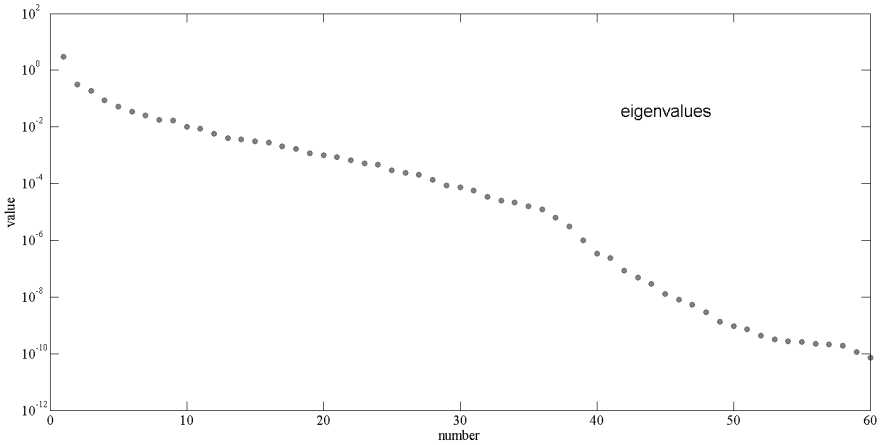


Figure 4.10: Eigenvalues of the covariance matrix.

done retrieving profiles only from ground P_s and T_s measurements, then adding a single measurement of tropospheric delay, DL. Errors (i.e. the tolerances) that we have used in the retrieval we present are $5.0\,hpa$, $2.0\,K$ and $1.0\,cm$ for P_s, T_s and DL respectively. Errors on P_s and T_s are not evaluated as instrumental errors, that would be much lower, but as representativeness uncertainties. These values in fact are point measurements of parameters that instead should refer to a finite area, defined by our grid representation of the atmosphere. Even if this is a 1D problem we are still working in a virtual cell, that in our case is given by the horizontal resolution of the model, which has been the source of our prior state distributions Errors have been thus estimated according to such cell dimension. The error on DL have been then estimated on the basis of the evaluations made in § 3, assuming to average a number of observations from the same satellites for few minutes[29].

In Fig. 4.11 we can observe the different decrease of entropy (i.e. the gain of information) from the starting *a priori* distribution due to the ground measurements and the addition of the GNSS delay, for all the target parameters at

[29]We remind that GPS standard ground stations provide delay measurements each second. The error we adopted is thus smaller than the error to be associated to a single observation, but at the same time larger than what could be obtained by averaging all possible observations for a relatively long time, that in some cases can be of few millimetres.

different levels. The range of variation of the entropy is from the curve of the prior distribution to 0. The prior distribution in fact is what the procedure returns if no measurement is available or if measurement errors are larger than the prior distribution itself. On the contrary $H = 0$ means that the distribution is a Dirac δ, i.e. that is the most informative distribution[30]. We find that:

- ground data provide a lot of information that constrains also values at higher levels, and this is not surprising once we have verified a large correlation among layers (probably in our dataset larger than what is realistic);
- GNSS delays bring relevant additional information at all levels, with major effects on water vapour e (as expected), but with a great contribution also on of T and HGT, the latter especially at high levels where the entropy of the prior distribution is larger[31].

In addition, from the relative entropies of Fig. 4.12, reminding that information increases for increasing H_{rel}, we find that:

- GNSS delays can bring a strong variation in distribution centres, changing the retrieved best values. This happens at all levels, but particularly at the lower levels of HGT and e, where the main components of the delay come from, due to higher gas density and average water vapour concentrations.

From Eq. (4.49) we see that the relative entropy has a minimum value of 0, that corresponds to no difference between the two compared distributions. Again this is the case of the procedure returning the prior distribution (i.e. when H is

[30]Reminding that the limit of the ratio of two functions is equal to the ratio of their derivatives, such that:

$$\lim_{p\to 0^+} p\log p = \lim_{p\to 0^+} \frac{\log p}{\frac{1}{p}} = \lim_{p\to 0^+} -\frac{\frac{1}{p}}{\frac{1}{p^2}} = \lim_{p\to 0^+} -p = 0^-$$

and referring to Eq. (4.48), if we have $p_i = 0\ \forall i \neq j$ and $p_j = 1$, we obtain $H = 0$. The Dirac δ is the limit of the former case for infinitesimal bin width.

[31]We remind that reducing the distribution width of HGT defined with respect to fixed pressure levels, is essentially equivalent to improve the precision with which pressure is known, in a scheme where the height levels are fixed and pressure is a variable depending on height.

maximum). In turn, we have a maximum value for the relative entropy when the distribution after measurements has all states just in one bin, and this bin corresponds to the one that was the least probable in the prior distribution[32]. Such threshold values are reported in Fig. 4.12 too.

An example of retrieved profiles (i.e. of a state vector) is given in Fig. 4.13, where we have assumed to have the ground measurements of pressure and temperature and a measurement of the zenith GNSS delay. In the example in figure the synthetic measurements have been corrupted with an artificial error, shifting their values of half of the full tolerance range, with respect to the chosen "true" observable ones. It results a good capability of the procedure to retrieve the target parameters, even at high levels. However to correctly quantify this aspect we should have a more realistic prior dataset: once we had it, it would make sense to try retrievals also with a larger number of observations (meteo and GNSS), as we should realistically have in many situations.

4.4.4.2 The 3D case

Among the infinite number of numerical experiments we could imagine to test the 3D retrieval process, we have decided to set up the most simple configuration in order to have a restricted set of results but of clear meaning. In our theoretical experimental set up, we assume to have a second ground station, equipped with meteorological sensors, in the town of Siena (in Tuscany), about 50 km distant from Florence[33]. We imagine also to have the Florence GNSS station switched-off . Our target parameters are still the atmospheric profiles over Florence as in § 4.4.4.1, but here we want to evaluate the variation of information we have from the prior distribution due to simply the ground meteorological measurements in Florence (same as the 1D case), than adding the ground meteorological measurements performed in Siena, and finally integrating the set of measurements with a

[32]We remind that to bins with 0 states have been assigned a minimum value of $1/N$ as explained in § 4.3.1. If there are n_0 void bins, the assigned minimum number is $1/(N\,n_0$, and any of these bins can be chosen with no impact on the result.

[33]This distance is the minimum one we could adopt in this work, according to the resolution of the source dataset.

GNSS delay measured in Siena for different slant path angles, towards Florence (i.e. in the vertical plane including Florence and Siena). This is the minimum set up we can imagine to evaluate the information content that a slanting signal path can bring to neighbouring cells, not owing other ground GNSS stations. We must say that this is a very limiting configuration with respect to some real situations: we use in fact a single signal path very distant from the target point. On the contrary there are now a lot of regions with several available and less distant paths (e.g. the distribution of ground stations in the GPS network in Tuscany,). From our results we can however infer the effect of more measurements on our retrieval procedure, as through Eq. (4.34) we know that measurements impact directly on the precision and accuracy of the retrieved target parameters, through their value and precision (i.e. through their probability density function).

State vector components have been doubled to account for the variables over Siena, and accordingly we have enlarged the prior dataset and related matrixes. The diagonalisation has confirmed the high correlations of parameters also on the horizontal scale[34], part of which is again presumably due to model resolution and approximations.

The main results are summarised in the graphs of Fig. 4.14 and Fig. 4.15. The relevant horizontal correlation make the information contribution from Siena very important also on Florence, both due to ground meteorological measurements and due to GNSS observations. The difference in the observation inclinations gives a less marked contribution, that is a bit more relevant for the precision of the highest levels. However we have to bear in mind that even with the highest slant angle considered, i.e. 75° (equivalent to 15° over the horizon) our hypothetical ray path would pass over Florence at an height greater than 13 km (but this is what we could do with our source dataset). In addition the high horizontal correlation of the source data damps the possibility to differentiate the information from measurements, that all results from interpolations of horizontal features. In

[34]With respect to the 1D case, with two points we have found that the same variance (e.g. the 99%) is explained by about one third more eigenvalues, against the double we should have for horizontally uncorrelated variables.

other words we expect that real measurements, or at least a more detailed prior dataset, could bring to find more relevant contributions to information than what we have found from slanting GNSS paths. Such contribution then is expected to increase when considering several contemporaneous GNSS observations We can then deduce that, according to the adoptable grid resolution, highly slanted path are worth to be processed maintaining their inclination information, while for paths within a cone of a few tens of degrees pre-averaging process can make sense, essentially to save computing time.

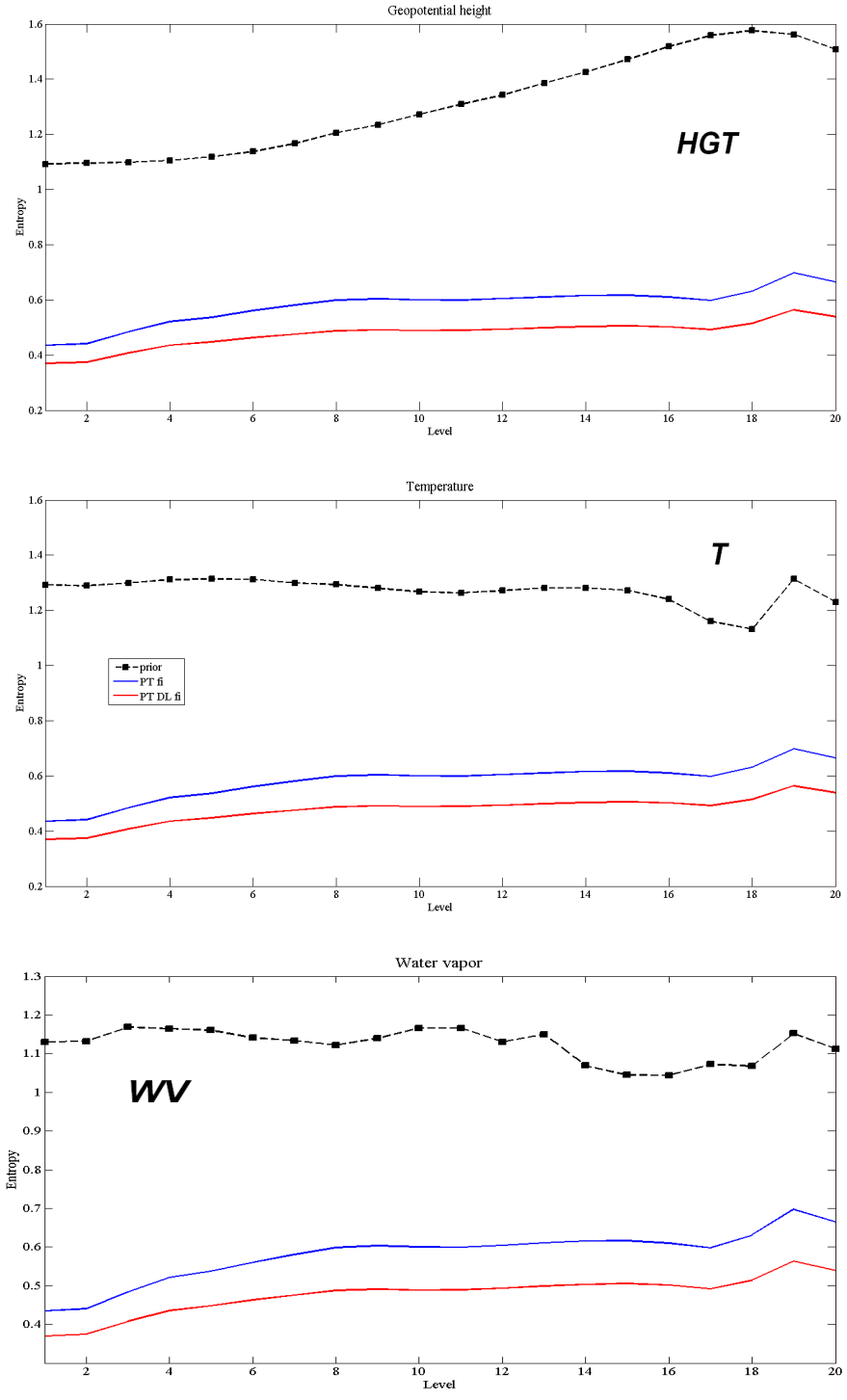


Figure 4.11: Entropies measured for the target parameters above a point corresponding to Florence (fi). The blue line is for retrievals using only ground measurements of pressure P and temperature T; the red one using also the GNSS signal delay (DL). Top black line with squares gives the entropy of the prior dataset, as maximum obtainable value.

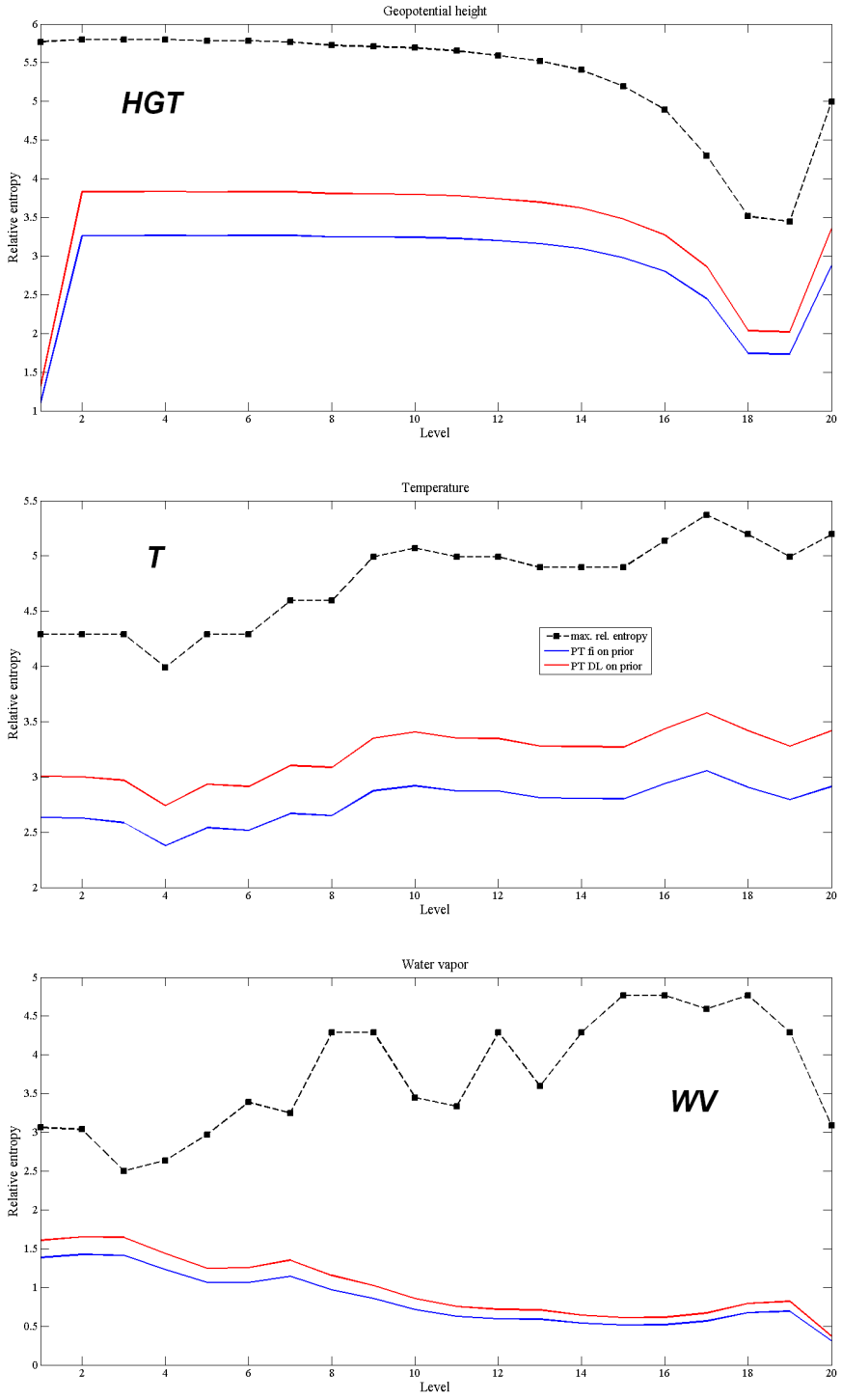


Figure 4.12: Relative entropies measured for the target parameters above a point corresponding to Florence (fi). The blue line is for retrievals using only ground measurements of pressure P and temperature T; the red one using also the GNSS signal delay (DL). Top black line with squares shows the maximum value for the relative entropy.

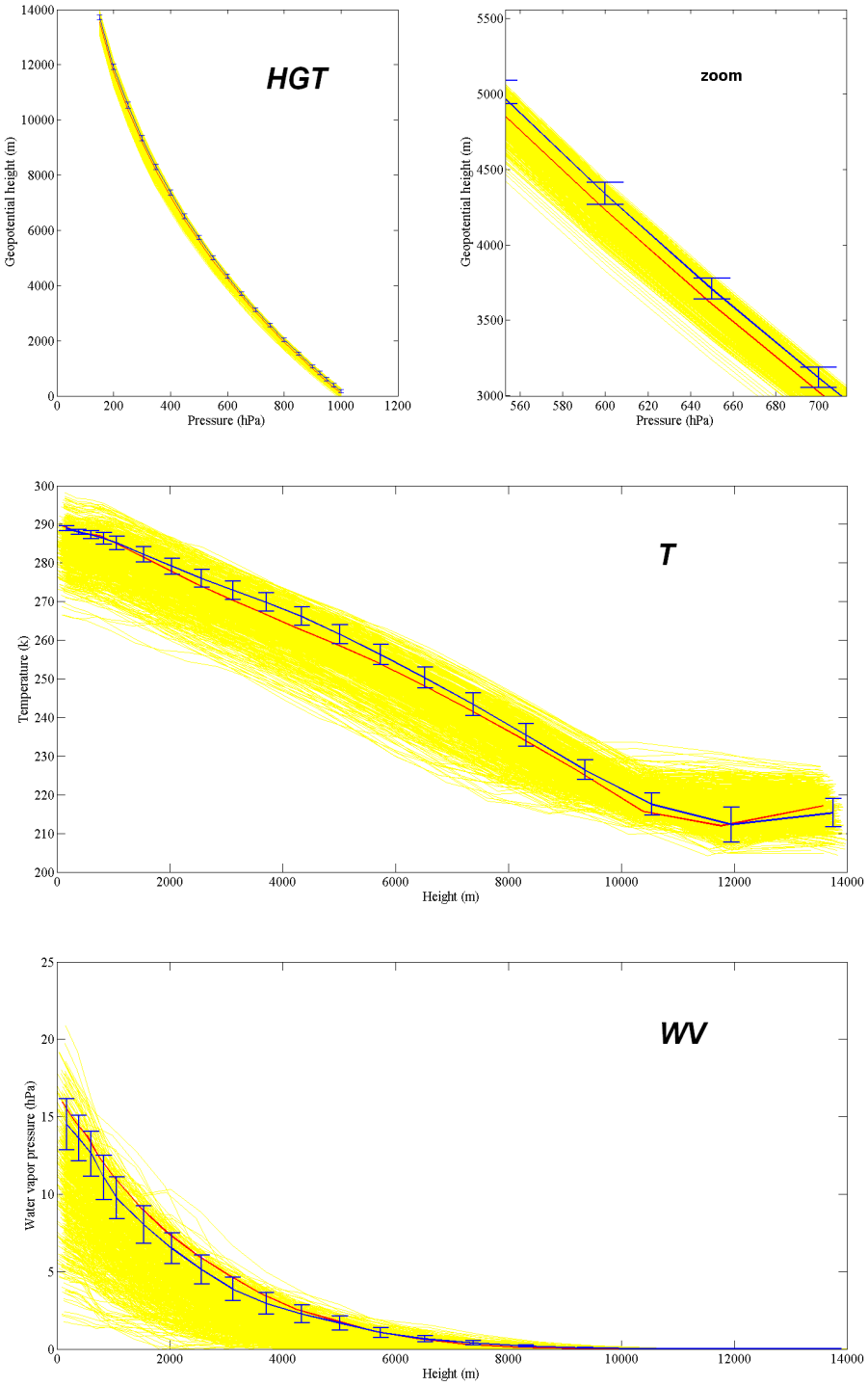


Figure 4.13: Example of a state vector retrieval using ground measurements for pressure, temperature and GNSS signal delay. Red lines are "true" parameter profiles; blue lines the retrieved ones. The yellow background is given by prior states. 1σ error bar are shown too. The geopotential height HGT is given as a function of the pressure P: a zoom is included to better show retrieval results. Temperature T and water vapour partial pressure e instead are given as function of the height (through HGT). Synthetic measurements are generated from the corresponding observables corrupted with errors about half of the measurement uncertainties.

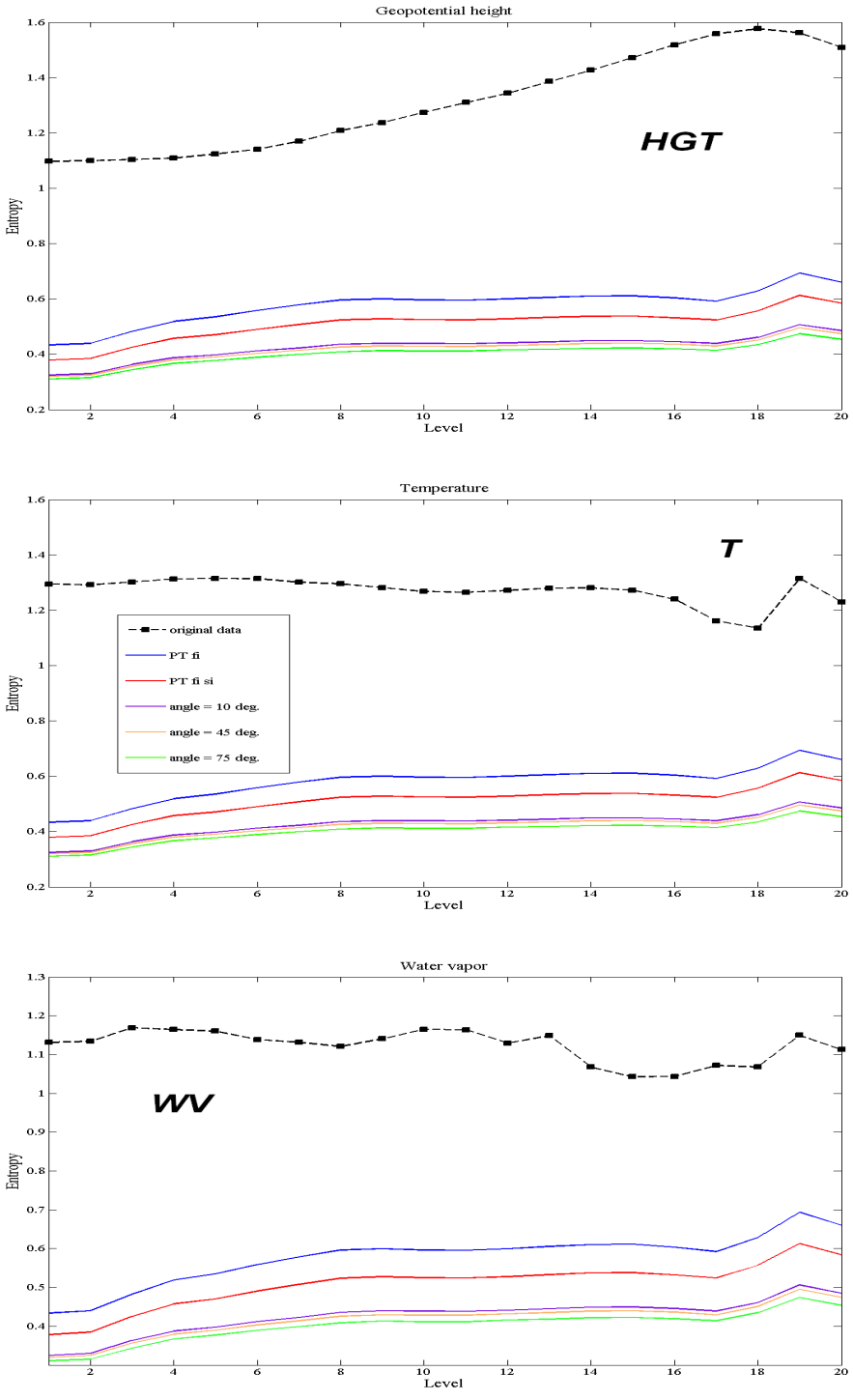


Figure 4.14: Entropies measured for the target parameters above a point corresponding to Florence (fi), but with measurements also in a different point corresponding to Siena (si). The blue line is for retrievals using only ground measurements of pressure P and temperature T in Florence; the red one using also ground P and T measured in Siena; the other coloured lines using in addition the GNSS signal delay (DL) measured in Siena but with different slant angles (with respect to the zenith direction). Top black line with squares shows the entropy of the prior dataset, as maximum obtainable value.

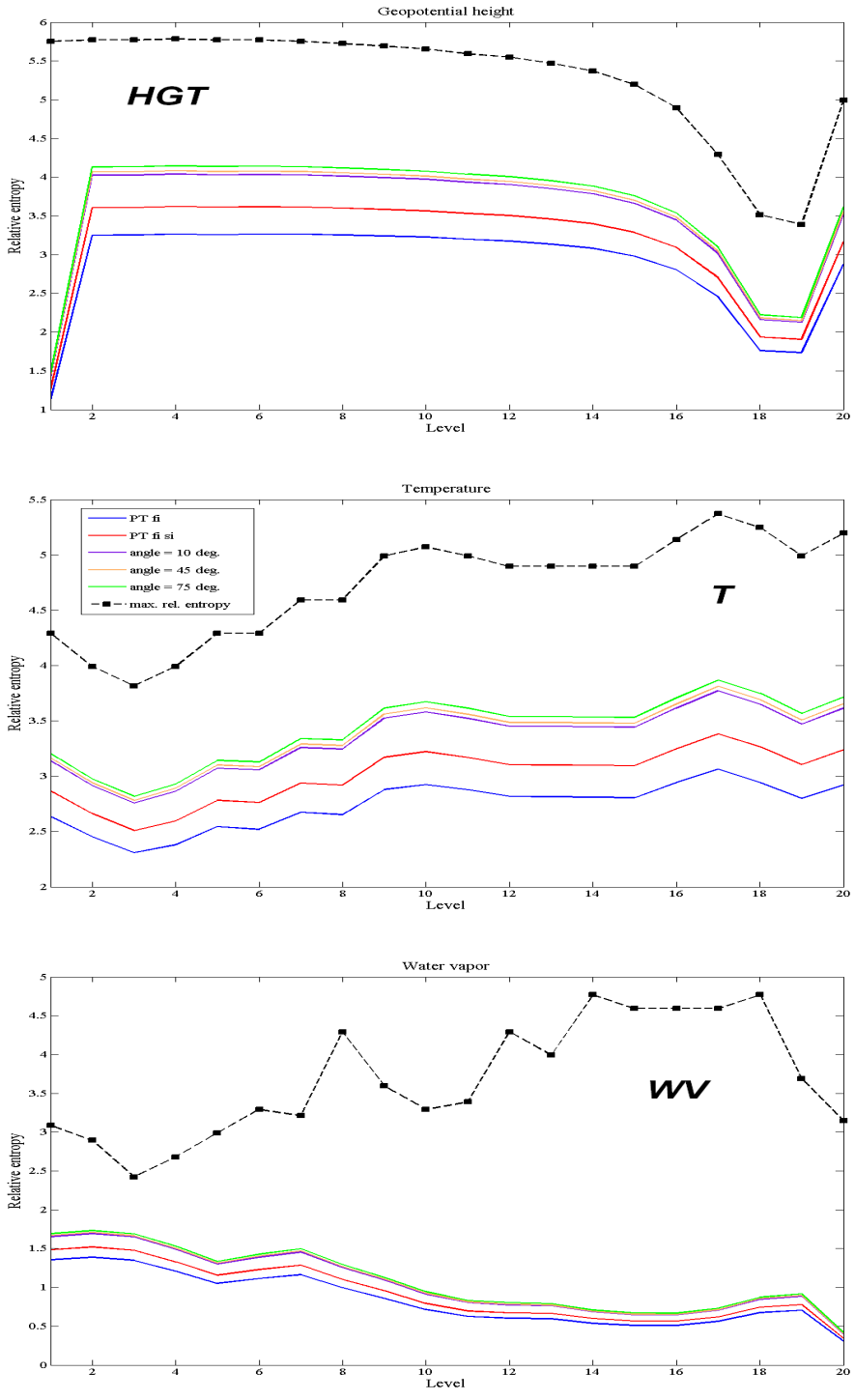


Figure 4.15: Relative entropies measured for the target parameters above a point corresponding to Florence (fi), but with measurements also in a different point corresponding to Siena (si). The blue line is for retrievals using only ground measurements of pressure P and temperature T in Florence; the red one using also ground P and T measured in Siena; the other coloured lines using in addition the GNSS signal delay (DL) measured in Siena but with different slant angles (with respect to the zenith direction). Top black line with squares shows the maximum value for the relative entropy.

Chapter 5

Lines of future development

The are several future developments of the technique described in § 4 and they are analysed in this short but dedicated chapter. Some of them endue not only with a scientific interest opening the possibility for more precise retrieval of atmospheric parameters (as in §§ 5.1, 5.2and 5.4), but also contribute towards an optimisation of the process computing-time and of follow-up applications (§ 5.3), or to extension to additional target parameters potentially of great interest (as in § 5.5).

5.1 Improvement of the *a priori* dataset

In several part of § 4.4 we have stressed the problem of using a prior source dataset made of a relatively small number of states and in addition non from real measurements, but from simulations of a global Numerical Weather Prediction (NWP) model. We have partially solved some of the problems that derive from the poor number of prior states with a generation of synthetic data (as described in § 4.4.2), but with the awareness that any synthetic densification of the prior dataset cannot upgrade its information content. We have already motivated why we believe that these limitations have not jeopardised the main conclusions arising from the test results of § 4.4.4, but for moving from simple synthetic experiments for testing the procedure to reliable assessments of the obtainable precisions, and then to real measurements, we would definitely need a different

a priori.

The ideal solution would be to have a large number of direct measurements, (e.g. radiosondes), but this is not practicable as we want to fill a prior dataset with several years of data, with some measurements per day (say each $3 \div 6\,hr$) for each grid node of our retrieval configuration. An alternative could be to use of NWP model data from Limited Area Model (LAM) simulations, at list to integrate available measurements. As an example, a not too demanding solution for us to be implemented in the future, could be to use the operational simulations run at LaMMA. These are 20 different simulations per day based on the Weather Research and Forecasting (WRF) modelling system, 14 initialised with GFS global model data, and 6 with ECMWF ones, with different resolutions, from 3 to $15\,km$ of horizontal grid resolution, with 35 tropospheric levels.

5.2 Shape of the error probability functions

In § 4.3.1 we have derived a formal expression for the retrieval method we have developed (i.e. Eq. (4.34)). We have also explained that for an easier implementation (and for saving computing time) we have adopted a square shape for the measurement error distributions, defined by a given tolerance value (see Eq. (4.35)), as the choice of a different error shape would not have changed substantially the results.

However, apart from computational aspects, a different choice for the error distribution shape, for instance a truncated Gaussian, could be preferable, because likely more realistic. Using such a distribution would have required a modification of the retrieval implementation. Instead of selecting the states whose related observables are compliant with the measurements within the tolerance ranges, using a "bell-shaped" error distribution function we should also store the probability values associated to the distance between the observables and the available measurements, and then use such information in the histogram generation for the retrieved parameters.

There are two additional aspects in favour of adopting bell-shaped curves as

error distributions.

The first aspect is that a bell-shaped distribution is non zero for a larger domain than a "quasi-equivalent" square one[1]. This means that the final number of selected states will be higher, of course with compliant states non equally probable. In other words with a more realistic error function we would "naturally" have more dense final histograms, with equal prior states, avoiding the loss of part of the information due to a coarse choice of the error distribution shape.

The second aspect is a bit more tricky. Let us assuming to have a number of measurements of the same observables giving exactly identical results, i.e. the same expected values. Errors distribution will be of course the same for all measurements, according to the adopted error model. We could expect that a sequence of identical results should enforce the confidence in the goodness of the measured value, improving its precision or the precision of linked retrievals (i.e. reducing the width of its error distribution or of the error distributions of related retrievals). All bell-shaped distributions for errors confirm this expectation, both for direct measurements and for indirect retrieval as ours, as can be seen from Eq. (4.35). Square shaped errors are an exception, as they state the perfect equivalence of measurements within a range: if results are not identical, then we have a reduction of the error range (or a similar effect on any related retrieval), but in case of identical measurements this cannot happen, and our procedure formal expression or numeric implementation confirms this. As in our numerical experiments we have not supposed to process more measurements of the same observable, this has surely not affected our results, but generally this is a risk to be avoided.

5.3 Transition to continuous variables

We have already said that EOF could be used to reduce the problem dimensionality. An interesting approach, partially alternative, is the use of orthogonal

[1]This "quasi-equivalence" could be assessed by the same standard deviation.

analytical functions. This would need to express the prior states through a given set of these functions, trying to find an acceptable compromise between the error committed in the transition from discrete samples to analytical functions, and the dimension of the basis of functions used to do the job. A proper choice of the basis is essential to be able to represent the large part of the variance through few function elements. Fourier series could result a reasonable choice of an orthogonal basis suitable for representing horizontal patterns in a limited domain. Instead for the vertical domain different bases could be necessary to better reproduce the different patterns of pressure, temperature and water vapour, according to their physics.

Assuming to be able to address this issue, we could proceed with state vectors made by a reduced number of functions coefficients, and the problem dimensionality could be a bit furthermore reduced with an EOF analysis.

The main advantages arising from such an approach would be that:

- we would manage to confine all necessary approximations (e.g. on problem dimensionality and required interpolations) at the initial stage of the retrieval procedure;
- we would efficiently contain the computing time due to:
  - a relevant reduction of the problem dimensionality,
  - the potential to analytically solve at least some parts of the retrieval problem;
- we would produce analytical solutions for the retrieved parameters (i.e. continuous state vectors), with great benefits for possible following applications requiring different scales.

5.4 Upgrading to a 4D processing

The retrieval procedure we have developed has never taken into account the opportunity to exploit the information from a time sequence of measurements.

We have always proceeded considering each measurement as unique, with the only available prior information given by our *a priori* dataset. The reason of this, is that we should primarily state the amount of information we could gain from different basic measurement configuration, independently on the dynamic of the events.

However one of the advantages of GNSS observations is that they are almost continuous, and the refresh time of modern meteorological stations is also very high. In real operational observations we should not waste the information content of previous measurements, that need to be propagated, through a more or less simple evolution model for the target parameters, and than exploited into new retrievals for additionally constraining the prior dataset[2].

This is one of the aspects that could drastically improve the capability of precise retrievals in our problem.

5.5 Growing the state vectors for new measurements

One of the interesting properties of the procedure we have implemented is its flexibility, meaning that it can work independently on number, type, precision and geographical distribution of available measurements. Of course measurement availability and characteristics affect retrieval accuracy, but not the capability to produce a result, that at worst will be no more informative than the prior probability distribution.

Another property is that, in principle, it is easily upgradeable to include additional target parameters, that can be added as new variables in the state vector, becoming new retrievable parameters. Of course the real capability of retrieving a given parameter depends on a number of factors, included the availability of measurements that strongly depend on it. Difference in typical time as well as spatial scales of variation are among the information that can be valuably used,

[2]This roughly summarises the basics of sound techniques based on Kalman filtering (e.g. cf. [34]).

in order to decouple the contribution of different parameters, helping to retrieve their value.

In a core part of a FP7 research proposal, submitted on a call for Galileo scientific applications[3], we have proposed to evaluate the capability of our approach to address the issue of measuring atmospheric target parameters over the sea. The marine environment, including the atmospheric component, is definitely under-sampled, while there is an increasing demand for information related also to potential (commercial) applications. The idea is to test the feasibility of exploiting ships as slow-moving platforms, where "slow" identify platforms whose displacement during the measurement time is much less than the typical spatial scale of variation of the target parameters, and whose velocity fluctuations (e.g. due to waves) occur on time scales much smaller than the measurement one (that in turn will be smaller than the typical time scale of variation to be resolved for the target parameter). Clearly, with respect to the problem of measuring atmospheric delays from fixed stations, we have to manage an additional (slow-changing) variable, the receiver position, which is not known *a priori* with negligible uncertainty and thus become part of the additional parameters in the sate vector. What is to be tested is if the large number of GNSS satellites with similar characteristics, that offshore will be simultaneously in view with Galileo operational in addition to GPS, could allow to effectively track at the same time positions and atmospheric delays, with the necessary accuracy.

The flexibility of the proposed approach could find valuable applications also in very different issues with respect to the one has been developed for. One for all could be the water loading estimation problem. The interest in measuring water masses is obvious and the water loading effect in crustal displacement is

[3]The mentioned proposal is COSMEMOS, that has been submitted to the Call: FP7-GALILEO-2011-GSA-1, Area: Scientific Application; Topic: Galileo and EGNOS for scientific applications and innovative applications in new domains. It has been ranked and it is in the group of proposals selected for funding.

known, and it has already been subject of retrieval through GNSS signal based techniques (e.g. cf. [44]). Variations to be measured are in this case of the order of millimetres, thus comparable with a number of other effects analysed in § 3, that should be carefully modelled and included in the retrieval process. Of course spatial and time scales of interest are in this application completely different from the one addressed in this work and a specific analysis should be performed in order to understand to what extent a technique similar to the one developed could extract useful information and return retrievals with the required precision.

Concluding remarks

The core objective of this work is the evaluation of the potential capabilities of navigation satellite signals to retrieve basic atmospheric parameters. The interest on this subject arises on one side from the increasing demand of meteorological information at more and more detailed spatial and temporal scales, and on the other side from the increasing availability of GNSS measurements, both for the expansion of the ground networks of fixed receivers (mainly settled for precise positioning) and for the programmed launch of new GNSS platforms, *in primis* the European Galileo system.

A number of "receipts" already exists regarding GNSS atmospheric applications, mainly oriented to evaluate zenith integrated water vapour quantities, often expressed in terms of precipitable water. In the present work we have tried an alternative approach to verify on which extent it is possible to retrieve profiles of parameters of interest from one single GNSS observation, integrated with ancillary meteorological ground measurements, in the perspective of using the "forest" of navigation signals that already exist and that is going to densify.

To this purpose we have performed a capillary study of the assumptions more or less explicitly contained in the common processing steps of navigation signals. We have started from the "very beginning", including an analysis of the origin of the atmospheric refractive index at the frequencies of interest (i.e. the GNSS operating ones), in order to identify the dependency on the atmospheric features and to quantify the approximations contained in the geometrical optics scheme, adopted for modelling the navigation signal behaviour.

We have implemented a simulator to evaluate phase delay and bending of

a GNSS signal travelling across a standard atmosphere. Then it has been upgraded to manage general atmospheric scenarios, including the possibility of non monotonic path curvature variations (useful for observations low on the horizon).

With reference to the present GPS satellite constellation, we have accurately analysed the numerous processing steps necessary to obtain that part of the signal containing only tropospheric effects, and the various errors components associated to each step. Such analysis have been integrated by a study of a small set of real data from the ground GPS network in Tuscany[4].

With such acquired background information we have tried to overcome the limits of standard approaches adopted in GNSS meteorological applications, with the aim of measuring at first zenith profiles of all atmospheric features affecting the refractive index (instead of mean water vapour values) and then to extend the method to 3D retrievals. To this aim we have demonstrated the unsuitability of the classical tomography for addressing the 3D problem, due to the varying geometries of GNSS observations from irregularly distributed ground stations, in favour of a flexible probabilistic approach.

A probabilistic procedure has been thus designed for measuring vertical discretised profiles of pressure, temperature and water vapour and their associated errors at each vertical level. The data of input are of course GNSS tropospheric delays and surface measurements of temperature and pressure. The procedure has been upgraded to include simultaneous retrievals from more points, i.e. to potentially perform 3D retrievals. It has been structured to process unevenly distributed measurements and to be able to work independently on the number and precision of observations that are from time to time available.

Numerical experiments on a synthetic dataset have been set up with the main aim of quantifying the information that could be gained from such approach, using entropy and relative entropy as testing parameters. The results demonstrate the potential of GNSS data in providing valuable information for 3D retrieval of

[4]Such network has been set up for supporting precise positioning and consists of heterogeneous receivers that are not born to primarily support GPS meteorological applications.

temperature and water vapour.

However this cannot be considered a conclusive work for two main reasons. The first one could be said "practical", and it is related to the way we have implemented the proposed procedure, which is too computing-time demanding to allow an immediate application to real operational measurements. The second one is more "theoretical" as it refers to the precision we could obtain for the target parameters with, when applying our method to a real measurement situation. The answer to this question can be only partially inferred in this work, as we have not worked with a strongly reliable prior dataset, and (being aware of this) we have not tried retrievals with a synthetic experimental set up similar to a real one (i.e. with several GNSS observations and ground meteo stations).

These problem will be matter of future development, as they have been finally pointed out, with specific indications on how to overcome some of the identified limitations and to develop further improvements towards both scientific and operational advancements. One of this line has been already judged of interest from the refereeing panel of the European FP7 Galileo programme, as it is part of a submitted project proposal, that has been positively ranked and whose activity are going to start in the second half of year 2011.

Bibliography

[1] Air Resources Laboratory (ARL) of NOAA's Office of Oceanic and Atmospheric Research (OAR), 2010. United States Department of Commerce [http://ready.arl.noaa.gov/].

[2] Ashby N., 2003: *Relativity in the Global Positioning System.* Living Rev. Relativity 6, Boulder [http://www.livingreviews.org/lrr-2003-1].

[3] Bao-Yen Tsui, J., 2000: *Fundamental of global positioning system receivers: a software approach.* New York, 2000, John Wiley & Sons Inc., 240 pp.

[4] Bevis, M., S. Businger, T. A. Herring, C. Rocken, R. A. Anthes, and R. H. Ware, 1992: *GPS meteorology: remote sensing of atmospheric water vapor using the global posisioning system.* J. Geoph. Research, 97 (D14), 787-801.

[5] Bilitza, D., 2001: *International Reference Ionosphere 2000.* Radio Sci., 36(2), 261-275.

[6] Brandt, S., 1978: *Statistical and Computational Methods in Data Analysis.* North-Holland Publishing Company, Amsterdam, 418 pp.

[7] Brizard, A. J., 2004: *Introduction to lagrangian and hamiltonian mechanics.* Department of Chemistry and Physics, Saint Michaels College, Colchester, VT, 173 pp.

[8] Buck, A. L., 1981: *New equations for computing vapor pressure and enhancement factor.* J. Appl. Meteorol., 20, 1527-1532.

[9] Chao, C. C., 1972: *A model for tropospheric calibration from daily surface and radiosonde ballon measurements*. Tech. Memo. Calif. Inst. of Technol. Jet Propulsion Lab., 391-350.

[10] Clough, S. A., Y. Beers, G. P. Klein, and L. S. Rothman, 1973: *Dipole moment of water of stark measurements of H_2O*. J. Chem. Phys., 59 (5), 2254-2259.

[11] Dach, R., U. Hugentobler, P. Fridez, and M. Meindl, 2007: *Bernese GPS Software Version 5.0*. Astronomical Institute, University of Bern.

[12] Davidson, J. M., and D. W. Trask, 1985: *Utilization of mobile VLBI for geodetic measurements*. IEEE Trans. Geosci. Remote Sens., GE-23, 426-437.

[13] Davis, J. L., T. A. Herring, I. I. Shapiro, A. E. Rogers, and G. Elgered, 1985: *Geodesy by radio interferometry: Effects of atmospheric modeling errors on estimates of baseline length*. Radio Sci., 20, 1593-1607.

[14] Debye, P., 1929: *Polar Molecules*. Chemical Catalog Company, New York, 172 pp.

[15] Doornbos, E., M. Foster, B. Fritsche, T. van Helleputte, J. van den Ijssel, G. Koppenwallen, H. Lohr, D. Rees, P. Visser, and M. Kern, 2009: *Air density models Derived from multi-satellite drag observations*. ESA's Second Swarm International Science Meeting, 24-26 June GFZ Potsdam, Germany [http://www.congrex.nl/09c24/].

[16] Eckart, C., 1948: *The approximate solution of one dimensional wave equations*. Rev. Mod. Phys., 20, 399-417.

[17] Elgered, G., J. L. Davis, T. A. Herring, and I. I. Shapiro, 1991: *Geodesy by radio interferometry: Water vapor radiometry for estimation the wet delay*. J. Geophys. Res., 96, 6541-6555.

[18] Fermi, E., 1972: *Termodinamica*. Bollati Boringhieri Ed., Torino.

[19] Feynman, R. P., R. B. Leighton, and M. Sands, 1963: *The Feynman Lectures on Physics*. Addison-Wesley Publishing Company, Massachusetts, Vol. 1, 515 pp.

[20] Grupta, A. K., T. F. Mri, and G. J. Szkely, 2000: *How to transform Correlated Randon Variables into Uncorrelated Ones*. Appl. Mathematics Letters, 13, 31-33.

[21] ICD, 1993: *Interface Control Document - Navstar GPS Space Segment / Navigation User Interfaces* (ICD-GPS-200, 1993).

[22] Jackson, J. D., 1962: *Classical electrodynamics*. John Wiley & Sons Inc., New York, 641 pp.

[23] Khinchin, A. I., 1978: *Fondamenti matematici della teoria dell'informazione*. Cremonese Ed., Roma, 129 pp.

[24] Kline, P. A., 1997: *Atomic Clock Augmentation For Receivers Using the Global Positioning System*. PhD Thesis, Ohio University - Avionics Engineering Center [http://scholar.lib.vt.edu/theses/available/etd-112516142975720/].

[25] Klobuchar, J.A., 1987: *Ionospheric time-delay algorithm for single-frequency GPS Users*. IEEE Transactions, AES-23(3), 325-331.

[26] Lomb, N. R., 1976: *Least-squares frequency analysis of unequally spaced data*. Astrophysics and Space Science, 39, 447-462.

[27] Lower, S. K., 1998: *Thermal physics (and some chemistry) of the atmosphere*. J. Chem. Educ., 75 (7), 837-840.

[28] Lowry, A. R., C. Rocken, S. V. Sokolovskiy, and K. D. Anderson, 2002: *Vertical profiling of atmospheric refractivity from ground based GPS*. Radio Sci., 37 (3), 13.1-13.10, doi:10.1029/2000RS002565.

[29] Misra, P., and P. Endge, 2001: *Global Positioning System-Signal, Measurements, and Performance.* Ganga-Jamuna Press, Lincoln, Masssachusetts, 390 pp.

[30] Owens, J. C., 1967: *Optical refractive index of air: Dependence on pressure, temperature and composition.* Appl. Opt., 6, 51-58.

[31] Parkinson, B. W., and J. J. Jr. Spilker, 1996: *Global Positioning System: Theory and Applications, Vols.1 and 2.* American Institute of Aeronautics, 370 LEnfant Promenade, SW, Washington, DC.

[32] Pearson, K., 1901: *On Lines and Planes of Closest Fit to Systems of Points in Space.* Philosophical Magazine, 2 (6), 559-572.

[33] Prasad, R., and M. Ruggieri, 2005: *Applied Satellite Navigation Using GPS, GALILEO, and Augmentation Systems.* Artech House Mobile Comunication Series, Boston, 290 pp.

[34] Rodgers, C., 2000: *Inverse Methods for Atmospheric Sounding: Theory and Practice.* World Scientific Publishing, Series on Atmospheric, Oceanic and Planetary Physics, Vol. 2.

[35] Saastamoinen, J., 1972: *Atmospheric correction for the troposphere and stratosphere in radio ranging of satellites.* in The Use of Artificial Satellites for Geodesy, Geophys. Monogr. Ser., vol. 15, edited by S. W. Henriksen, et al., 247-251.

[36] Shannon, C. E., 1948: *A Mathematical Theory of Communication* The Bell System Technical Journal, Vol. 27, 379423.

[37] Shannon, C. E., 1949: *Communication in the presence of noise.* Proc. Institute of Radio Engineers, 37 (1), 10-21.

[38] Stull, R. B., 2000: *Meteorology for Scientists and Enginieers.* Brooks/Cole Thomson Learning, U.S.A, 502 pp.

[39] Thayer, D., 1974: *An improved equation for the radio refractive index of air.* Radio Sci., 9, 803-807.

[40] Toraldo di Francia, G., and P. Bruscaglioni, 1988: *Onde elettromagnetiche.* Zanichelli Ed., 705 pp.

[41] Tralli, D. M., and S. M. Lichten, 1990: *Stochastic estimation of tropospheric path delays in global positioning system geodetic measurements.* Bull. Geod., 64, 127-159.

[42] Ulaby, T. F., K. R. Morore, and K. A. Fung, 1943: *Microwave remote sensing, active and passive.* Library of Congress Cataloging in Publication Data, Artech-Hause, Londra, Vol.I, 456 pp.

[43] U.S. Standard Atmosphere, 1976. U.S. Government Printing Office, Washington, D.C.

[44] Van Dam, T., Wahr, J., Milly, P. C. D., Shmakin, A. B., Blewitt, G., Lavallèe, D. and Larson, K. M., 2001: *Crustal displacements due to continental water loading* Geophys. Res. Lett., 28, 651-654.

[45] Van Vleck, J. H., and V. F. Weisskopf, 1945: *On the shape of collision-broadened lines.* Rev. Mod. Phys., 17, 227-236.

[46] Vultaggio, M., 2009: *Lezioni di navigazione satellitare.* Università di Napoli Partenophe - Facoltà di Scienze e Tecnologie, Napoli.

[47] Wallace, J., and P. V. Hobbs, 1977: *Atmospheric science an introductory survey.* Academic Press, 467 pp.

[48] Xu Guochang, 2007: *GPS Theory, Algorithms and Applications.* 2nd edition, Berlin Heidelberg 2007, Springer-Verlag, 340 pp.

[49] Zemansky, M. W., 1987: *Calore e Termodinamica.* Zanichelli Ed., Bologna, Vol. I, 255 pp.

Acknowledgements

My first thanks goes to Prof. Rolando Rizzi for his continuous support along the non easy steps of this work, for the fruitful discussions we had together and for his constant transfer of experience, enthusiasm and motivation.

I have in part drawn on the knowledge of my colleagues Andrea Antonini and Riccardo Benedetti. With them I had so many recurrent and intensive arguments to the point that in many aspects we cannot, at present, almost distinguish whom the contributed ideas belong to.

Samantha Melani, Andrea Orlandi and especially Luca Rovai are gratefully acknowledged for their technical support in the development of some aspects of this work; Prof. Kai Borre for his careful advices in the comprehension of some GPS signal criticalities.

I wish to acknowledge in a particular way also Carlo Brandini, Simone Cristofori, Bernardo Gozzini and Chiara Lapucci for their encouragement and for having volunteered to be my back-up in solving daily work issues.

A special thanks goes to my children, Davide and Sofia, who have been understanding and who managed to cope with my absence during the period these pages were written.

Finally, I am grateful to my wife Lucia. For her, there are no words to describe the meaning she has in every single event of my life.

GFS data for this study are from NOAA's National Operational Model Archive and Distribution System (NOMADS).