Pedro Naethe Motta

GRMHD simulations of X-ray binaries in the hard state

Dissertação apresentada ao Departamento de Astronomia do Instituto de Astronomia, Geofísica e Ciências Atmosféricas da Universidade de São Paulo como requisito parcial para a obtenção do título de Mestre em ciências.

Área de Concentração: Astronomia

Orientador(a): Prof. Dr. Rodrigo Nemmen da Silva

Corrected Version: The original is available in IAG-USP

São Paulo

2023

To Flávia and Jorge ...

Acknowledgements

This work was only possible because of the support of my family, friends and my dogs Chloé and Chanel.

I would like to begin by thanking my family, starting with my parents, who have been pillars of strength in my life, Flávia Miller Naethe Motta and Jorge da Motta Ferreira for their support throughout my whole life. I would like to thank my sisters Carolina Naethe Motta and Paula Naethe Galvão for their companionship;

To all of my friends whom I met at Universidade de São Paulo: Pablo, Elismar, Erik, Gê, Douglas, Roberta, Guilherme, Henrique, Marcelo, Fernanda, Gustavo and Thayse. Without you guys, this would have been much harder;

I would also like to thank my advisor Rodrigo Nemmen da Silva for trusting me with a project as exciting as this one;

I thank Professor Alexander Tchekhovskoy and his research group for having me at Northwestern University and for also helping with this project. Special thanks to Peter Anglada for welcoming me and helping me with the bureaucratic part, in addition to familiarizing me with the university;

In addition, I thank Jonatan Jacquemin for collaborating, helping, and teaching me throughout the end of this project. This would not have been able without you as well;

I also would like to express my sincere gratitude to São Paulo Research Foundation (FAPESP), for the financial support, under grant 2021/01563-5 and for the BEPE under grant 2022/07456-9. The opinions, hypotheses, conclusions, or recommendations expressed in this material are the responsibility of the author and do not necessarily reflect the views of FAPESP.

Eu gostaria de expressar a minha sincera gratidão à Fundação de Amparo à Pesquisa

do Estado de São Paulo (FAPESP), pela bolsa sob processo 2021/01563-5 e pela bolsa de estágio de pesquisa no exterior (BEPE) sob processo 2022/07456-9. As opiniões, hipóteses e conclusões ou recomendações expressas neste material são de responsabilidade do autor e não necessariamente refletem a visão da FAPESP;

Finally, I would like to acknowledge LNCC (Laboratório Nacional de Computação Científica) and OLCF (Oak Ridge Leadership Computing Facility) for the computational time given to me on the Santos Dumont and SUMMIT clusters, respectively.

Esta tese/dissertação foi escrita em LaTeX com a classe IAGTESE, para teses e dissertações do IAG.

"Those who died are justified
For wearing the badge, they're the chosen whites
You justify those that died
By wearing the badge, they're the chosen whites
Come on! "

Rage Against the Machine - Killing in the name, 1991

Resumo

As binárias de buracos negros em raios-X (BH XrBs) são laboratórios astrofísicos essenciais para investigar a interação entre discos de acreção e buracos negros. Esses sistemas, compostos por um buraco negro e uma estrela companheira, exibem uma variedade de estados espectrais, oferecendo uma oportunidade única para estudar a dinâmica dos processos de acreção em escalas de tempo mais curtas. Neste trabalho, abordamos o desafio de capturar os processos radiativos que ocorrem no plasma, através da implementação de uma prescrição de radiação de baixo custo computacional no código acelerado por GPU `H-AMR`. Nossa abordagem incorpora uma solução radiativa autoconsistente que leva em conta o bremsstrahlung, a radiação síncrotron, a radiação sincrotron comptonizada e as colisões de coulomb. Por meio de diversos testes, validamos a funcionalidade de nossa metodologia. A nossa implementação fornece uma ferramenta de baixo custo para reproduzir os efeitos radiativos, possibilitando investigar as propriedades e a dinâmica dos discos de acreção em BH XrBs, avançando assim nossa compreensão desses sistemas astrofísicos.

Abstract

Black hole X-ray binaries (XRBs) serve as essential astrophysical laboratories for investigating the physics of the accretion flows and black holes. These systems, composed of a black hole and a companion star, exhibit a variety of spectral states, offering a unique opportunity to study the accretion dynamics on shorter timescales when compared to supermassive black holes (SMBHs). In this work, we address the challenge of capturing the radiation processes within XRBs by presenting an implementation of a radiation prescription, which offers a lower computational cost compared to traditional radiative transfer codes, within the GPU-accelerated code `H-AMR`. Our approach incorporates a radiative solution that accounts for bremsstrahlung, synchrotron radiation, comptonized synchrotron, and coulomb collisions. Through extensive testing, we validate the functionality of our methodology. Our implementation provides a low-cost tool for investigating the properties and dynamics of black hole accretion flows in XRBs.

List of Figures

Contents

Chapter 1

Introduction

Black Holes (BHs) are extremely compact bodies where matter has been compressed tightly resulting in a gravity field so strong that even light can't escape. These bodies arise as one of the most simple mathematical solutions to Einstein's field equations, in a space-time outside a static, spherically symmetric, star. The idea of a body with such immense gravitational pull predates the concept of black holes. In the 1700s, the British physicist John Michell and the French mathematician and physicist Pierre Laplace were already considering the idea of a body so compact, that the escape velocity of a particle exceeds the light speed, resulting in a dark star (Schutz, 2009).

It is possible to characterize the black hole as a supermassive black hole (SMBH) or a stellar-mass black hole . A supermassive black hole is usually found in the center of galaxies (Kormendy and Richstone, 1995). Unlike stellar-mass black holes, the processes of its formation are not so clear and is a subject of investigation. They are usually very massive reaching over millions or billions of solar masses. Stellar-mass black holes are born in the latest phases of a massive star's evolutionary process (Fryer, 1999) and usually have masses on the order of tens of solar masses. In the past few years, mounting evidence has also emerged supporting the presence of intermediate-mass black holes (Baumgardt et al., 2019, and references therein).

Because matter falls toward the black hole with some angular momentum, it will spin around before being accreted forming a disk of matter rotating around the black hole. For matter to spiral inwards, such as for the angular momentum to be carried away, a mechanism is required. It is believed that the mechanism responsible is the magnetorotational instability (MRI)(Balbus and Hawley, 1991), which arises from magnetohydrodynamics (MHD) physics. The gravitational energy lost as the gas gets closer to the black hole is

radiated away by the disk, and provides the main channel for detecting them. In these scenarios, radiative processes play a major role in the dynamic and evolution of the system.

Black holes X-ray binaries (XRBs) are systems composed of a black hole and a companion star and have been identified to have different spectral states throughout their lifetime. When compared to supermassive black holes, the stellar mass ones have a much shorter dynamical timescale, in a way that it's possible to see them alternate states in periods of weeks, which makes them the perfect probe to understand the different dynamics that take place. In order to comprehend the observations of XRBs, there is a need to develop models that capture the disk's physics.

Simulating accretion disks around black holes is a highly challenging task due to the dominant influence of electromagnetic and strong gravitational forces. Capturing the complex physical processes requires the use of a general relativistic magnetohydrodynamic (GRMHD) framework. Numerical codes play a crucial role in evolving these equations and enabling detailed simulations. Several widely used codes have been developed for this purpose, including `H-AMR` (Liska et al., 2022), `HARM`(Gammie et al., 2003), `Cosmos++` (Anninos et al., 2005), `Athena++` (Stone et al., 2020) and `BHAC` (Porth et al., 2017). Although radiation plays a crucial role in high luminosity scenarios, it is often implemented separately as a module or completely omitted due to its computational demands. Therefore, there is a need for the development of low-cost radiation prescriptions that can accurately capture the essential radiative processes in simulations.

Instead of implementing a full radiative transfer module, we incorporate the self-consistent solution proposed by Narayan and Yi (1995); Esin et al. (1996) into the GPU-accelerated code `H-AMR`. This is achieved by storing a lookup table in the GPU's texture memory. By adopting this approach, we can approximately capture the radiative processes, including bremsstrahlung, synchrotron radiation, comptonized synchrotron, and coulomb collisions. In this project, we outline our implementation methodology and present the extensive tests conducted to ensure its proper functionality. Furthermore, we will make this tool publicly available to facilitate future research endeavors.

The structure of this dissertation is as follows: In Chapter 2 we describe the physical processes related to accretion disk dynamics. In the next Chapter 3, we introduce the theoretical background to understand such dynamics. Chapter 4 discusses the implementation of the radiative cooling prescription in the GRMHD code `H-AMR`. To guarantee the

validity of the modifications, we perform a series of tests, described in a concise manner in Chapter 5. Finally, we show the results of the highest resolution run in Chapter 6 and summarize the contents of this work in Chapter 7.

Chapter 2

Accretion disk physics

The accretion disk is formed by matter with angular momentum spiraling, towards the singularity. Many different physical phenomena are involved in the dynamics of this motion. In this Section, we will analyze some of these.

2.1 Magnetorotational instability

Firstly, for the matter to be spiraling, we need a mechanism to dissipate the angular momentum. It is believed that the cause for this transport is the so called magnetorotational instability (MRI). It is sufficient to analyze the theory under magnetohydrodynamics (MHD) scope.

For some time, the nature of this instability was an open question in physics. Until today, it is not 100% clear, but we have clarity that the transport is caused by magnetic forces in the existing plasma. It was first discussed by Chandrasekhar (1961), but the appearance of this instability in the context of accretion disks was only shown by Balbus and Hawley (1991) 30 years later. The effects of this instability can be described by an ad-hoc prescription of an alpha viscosity parameter, which is useful for hydrodynamic simulations. The alpha parameter essentially acts like an internal friction within the accretion disk, due to the differential rotation. This friction heats up the disk as it moves inward, and also causes the disk to gradually lose its angular momentum.

In a qualitative discussion, we'll imagine two fluid elements in close orbits. They're connected by a "spring" that causes a restorative force in them, once there is a perturbation in the orbit of one of these fluid elements, the tension on the spring will force the outer one to accelerate and the inner one to slow down as we can see in the Figure 2.1. The spring is

an analogy to a restorative force caused by the magnetic field. This way, we allow matter to transport angular momentum and fall inward.

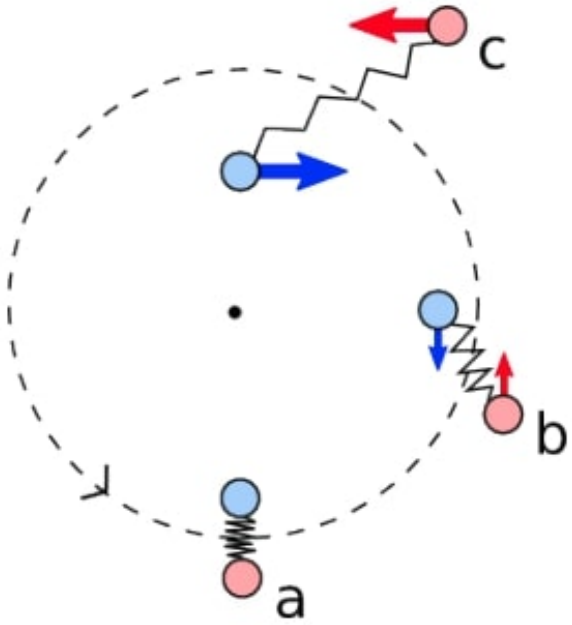

Figure 2.1: The magnetorotational instability as explained by a spring connecting two fluid elements. The outer fluid element is accelerated while the inner one is slowed down, this way, angular momentum is carried out by the accretion disk, from the inner region to the outer region (Murphy, 2014).

2.2 Radiation efficiency and accretion models

Black holes are usually detected due to the radiation emitted by the plasma gravitationally attracted to them. While the matter falls inwards, part of the gravitational energy is transformed into heat by some magnetic turbulence existing in the plasma.

The efficiency at which this heat is radiated away will have a huge impact on the dynamics of the disk. If the gas can radiate away its energy in a timescale shorter than the accretion time, it will cool rapidly, and with decreasing temperature, the pressure will drop resulting in a collapse of the disk into a thin disk. If the gas can't radiate the energy away in time, the gas will become extremely hot, swelling its geometrical form. We can define its energy efficiency by

$$\epsilon = \frac{L}{\dot{M}c^2}, \tag{2.1}$$

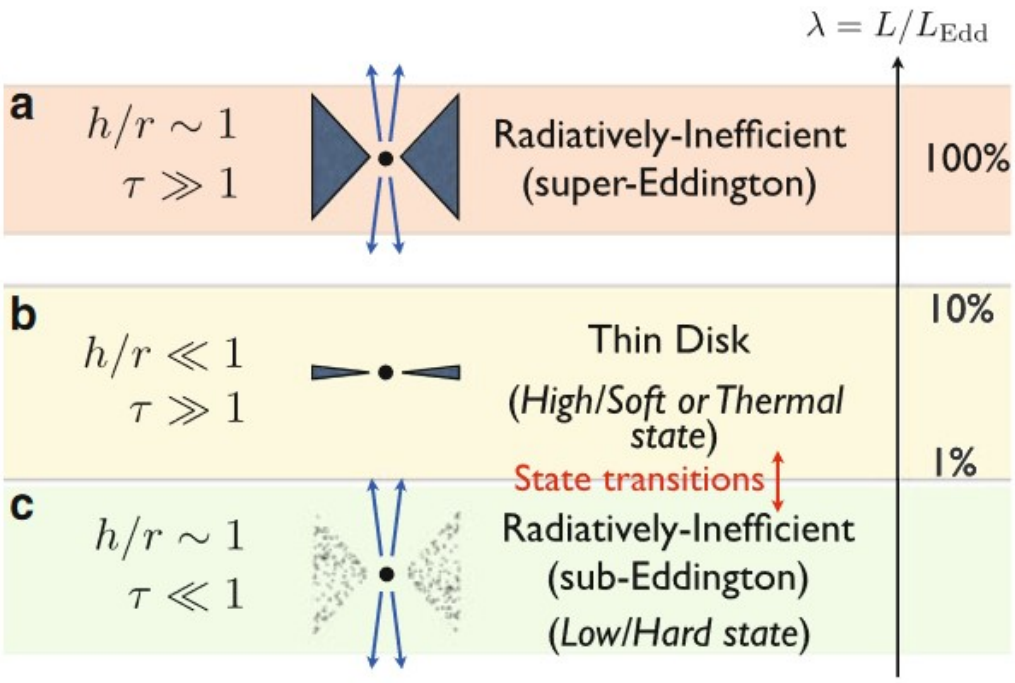

Figure 2.2: Visual representation of accretion flows switching between a RIAF and a thin disk regime as a function of the luminosity (Contopoulos et al., 2015).

where L is the luminosity of the disk, and $\dot{M}c^2$ is the rate at which the rest mass energy is being accreted. Different radiative efficiencies lead to distinct accretion dynamics (Figure 2.2), which we can, basically, characterize in:

- Radiatively inefficient accretion flow (RIAF): The gas can't radiate its energy fast enough ($\epsilon << 0.01$), making it extremely hot and geometrically thick. Most of the energy is then advected into the flow itself, leading to a RIAF (Narayan and Yi, 1994). This happens at low ($\dot{M} \lesssim 0.01\dot{M}_{\rm edd}$) and very high ($\dot{M} \gtrsim 1\dot{M}_{\rm edd}$) accretion rates.

- Thin disk: The energy is radiated away, $\epsilon \approx 0.06$ to 0.4 according to Narayan and Quataert (2005), cooling down the gas and making it collapse into a geometrically thin, optically thick disk. This regime is usually found in moderately high accretion rates ($0.01\dot{M}_{\rm edd} \lesssim \dot{M} < 1\dot{M}_{\rm edd}$). These states are usually modeled as described by Shakura and Sunyaev (1973); Novikov and Thorne (1973).

2.3 Spectral states of black hole x-ray binaries

Historically, the spectral cycle of XRBs consists of five different spectral states (Esin et al., 1997).

1. Quiescent state: In this state, the system presents a clearly non-thermal spectrum with a hard X-ray component (Ferreira, J. et al., 2006), with flux levels lower than the four other states by several orders of magnitude. The accretion rate is very low due to the low density of the disk ($\dot{m} \sim 10^{-9}$). The source remains mostly in the quiescent state, but as the compact object pulls matter from the companion star, it gets slightly brighter, transitioning to the next state.

2. Low/Hard state: The accretion disk density starts to go up, leading to an increase in luminosity. This spectral state presents a power law spectrum with an exponential cutoff around $\sim 100keV$.

3. Intermediate state: Spectrum starts to shift from the low state to the hard state. It usually features both thermal and non-thermal components that compete with each other.

4. High/Soft state: In this state, the spectrum is dominated by a blackbody-like component with characteristic temperature $\sim 1keV$ and total luminosity exceeding the hard state values.

5. Very high state: For some very high luminosity systems, the non-thermal tail of the spectrum becomes comparable with the blackbody-like component. In this state, luminosity seems to be higher than the typical ones seen in the high state and the spectra are also significantly harder.

It is important to mention that despite the success in Esin et al. (1997) model, our knowledge has increased in recent years. Contopoulos et al. (2015) argues that the very high state described in Esin's work is now identified with the intermediate state transition from high to soft and is a state reached before the soft state is entered. Also, the same work discusses the Super-Eddington thick-disk state, which is a RIAF being accreted at very high accretion rates. Due to higher disk density and thickness, the disk becomes optically thick. The time it takes for the photons to escape becomes longer than it takes

for the gas to be accreted, resulting in a small fraction of photons escaping. We are not going to consider this regime in this work.

The low-mass XRBs spend most of their life in the quiescent state, but at times, they produce an outburst that shows the distinct spectral states described above. These systems are perfect to study the different scenarios of accretion physics because they switch between them in a timescale of a few months to a year (Marcel et al., 2018). The difference in the spectra for each state can be seen in the model-generated spectra done by Esin et al. (1997) in Figure 2.3.

One approach to model these different spectral states is modeling the accretion flow as a RIAF truncated at a certain r_{tr} at which it transitions to a thin disk. The RIAF produces non-thermal radiation whereas the thin disk emits thermal radiation locally. We can picture the whole cycle in a BH XrB system as follows: The system is in the quiescent state and, thus, not easily detectable. In this scenario, the truncation radius is very far away ($r_{tr} \sim 10^3 - 10^4 R_g$, where $R_g = GM/c^2$) (Narayan et al., 1996; Hameury et al., 1997), for the thermal component to be important. The first detection is a sudden increase in luminosity and this happens because the rate of mass transfer from the companion star is higher than the accretion rate, so the disk gets denser, increasing the accretion rate. This is the low state at which $r_{tr} = 40 - 100 R_g$ according to the jet emitting disk (JED) model by Ferreira, J. et al. (2006). The disk's temperature increases, and with it, the viscosity has a substantial increase as well, allowing much more transfer of angular momentum outwards and inwards transfer of mass. The thin disk starts getting closer to the black hole, making the thermal component more important. This is the transient state, where both thin disk and RIAF regimes overlap. When the thermal component of the spectrum starts to dominate, we've reached the soft state where the truncation radius is pretty close to the innermost stable circular orbit radius, e.g. $R_{tr} \sim R_{ISCO}$. The accretion rate is higher than the mass transfer from the star, resulting in a gradual decrease in the accretion disk's density. The cooling time starts to increase until it becomes bigger than the viscous time, resulting in the increase of the plasma's temperature, acquiring the toroidal shape again, in a geometrically thick and optically thin environment. The accretion rate decreases as the truncation radius increase, dropping the luminosity, and reaching the hard state again. A more detailed explanation of the spectral state transitions can be found in the work of Fender and Belloni (2012).

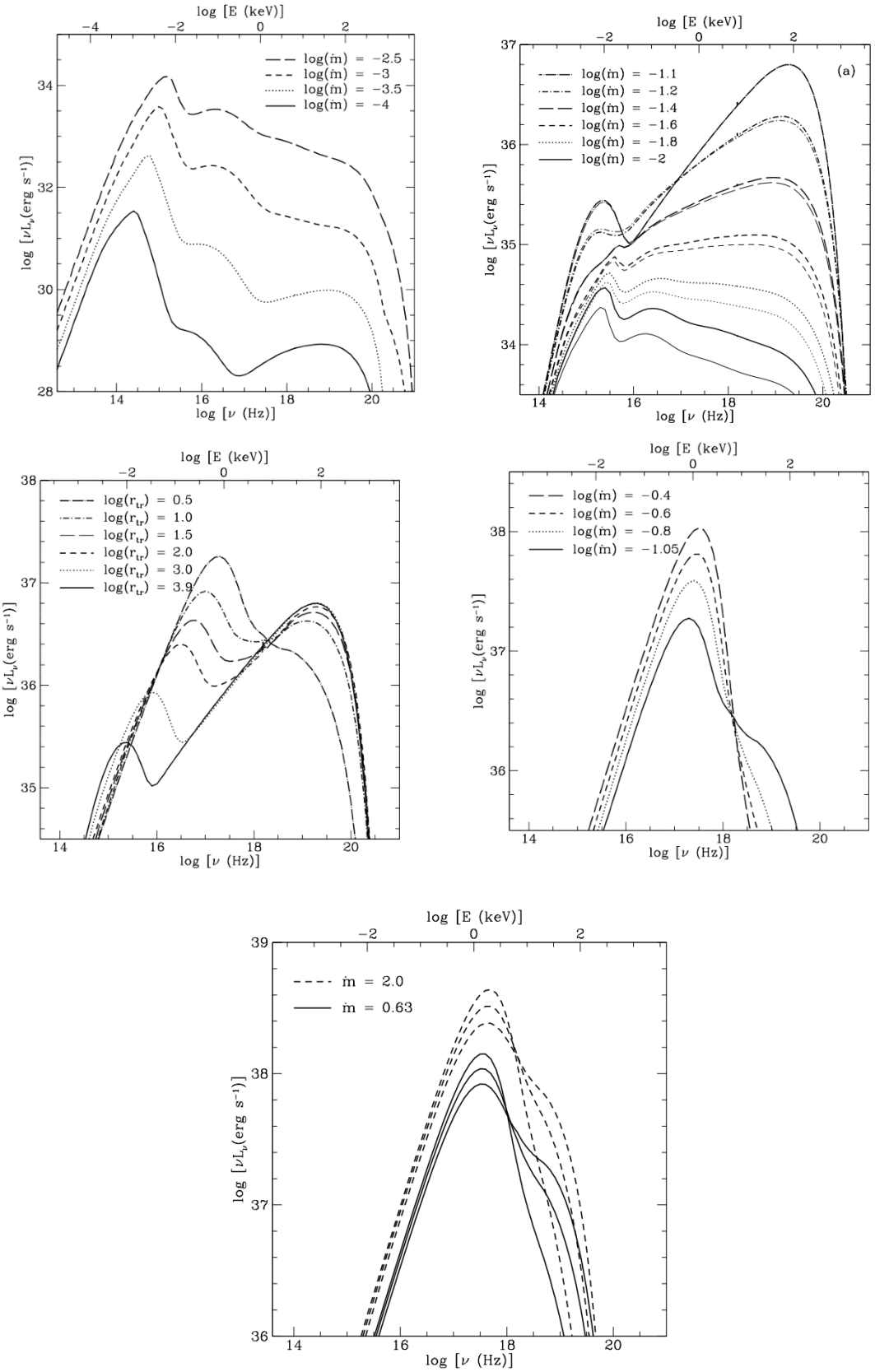

Figure 2.3: Spectra models for different XRB states (Esin et al., 1997). From top to bottom, left to right: Quiescent state, low state, intermediate state, high state, and very high state. The model reproduces observations of the soft X-ray transient Nova Muscae during its 1991 outburst.

Radiation plays a crucial role in regulating the dynamics of these accretion disks, making it essential to incorporate a realistic cooling prescription to reproduce these dynamics. However, accounting for the full radiative effects comes at a significant computational cost, making the theoretical reproduction of different states a challenging task.

Multiple previous simulations have made significant contributions to the study of XRB spectral states. Wu et al. (2016) conducted 2D hydrodynamical simulations with a radiative cooling prescription and consistently observed that the truncated radius moves inward as the accretion rate increases. Dexter et al. (2021) performed 3D radiation GRMHD simulations of the hard state and identified a critical accretion rate of $\dot{m} = 10^{-3}$, indicating the minimum accretion rate at which the disk collapses. Takahashi et al. (2016) conducted 3D GRRMHD simulations and confirmed the inward motion of the truncated disk at $R \approx 30R_g$ for $\dot{m} \approx 0.1$. Liska et al. (2022) employed radiation-transport, two-temperature GRMHD simulations and discovered the formation of a magnetically truncated disk at $R \approx 20R_g$ at $\dot{m} \approx 0.35$, where the inner disk is dominated by magnetic pressure. Dihingia et al. (2022) conducted 2D GRMHD simulations with a radiative cooling prescription to investigate temperature properties on different magnetic topologies such as magnetically arrested disk (MAD) and standard and normal evolution (SANE). However, their focus was on low accretion rates, which resulted in relatively weak radiative cooling processes. As a consequence, they did not observe a collapse of the optically thin disk. Furthermore, Nemmen et al. (2023) performed 3D hydrodynamical simulations with radiative cooling prescription and identified a polynomial power-law relationship between the accretion rate and the truncation radius.

As of 2022, H-AMR has a radiative transfer module based on McKinney et al. (2014) where they implemented radiative transfer equations in a full GRMHD context using an M1 closure scheme Levermore (1984). The M1 closure scheme uses a truncated moment formalism to solve the radiative transfer equations for all-optical depth limits. Although more realistic, this method is more expensive, requiring greater computational power. To account for the radiation effects, we implement the cooling prescription(Narayan and Yi, 1995; Esin et al., 1996) of Bremsstrahlung, synchrotron, and comptonized synchrotron in the 3D GRMHD code H-AMR (Liska et al., 2022) by means of a lookup table using GPU's texture memory. We discuss the implementation in Chapter 4. We plan to leave this useful tool in H-AMR for future use as a computationally faster alternative to the

radiative module.

In the next chapter, we'll discuss the theoretical GRMHD equations as well as some fundamental properties of the code used.

Chapter 3

Theoretical background

This chapter is divided into two parts. In the first, we describe the basics of GRMHD: general relativistic magnetohydrodynamics. In the second part, we describe the radiative cooling prescription used in this work.

3.1 General relativistic magnetohydrodynamics

The accretion disk is a plasma, e.g. a very hot gas with enough charged particles so that its dynamical behavior is dominated by collective effects involving electromagnetic forces. Alongside accounting for electromagnetic effects, the space-time curvature is strong enough to influence the dynamics with general relativistic effects. GRMHD provides a set of equations that accounts for electromagnetic effects in magnetized fluids interacting in curved spacetime.

In this section, we will analyze the equations of GRMHD. We'll use the conventional notation where $c = G = 1$ throughout this description.

3.1.1 Mass density conservation

The first equation describes the conservation of the number of particles, which we can write as

$$(nu^{\mu})_{;\mu} = 0, \tag{3.1}$$

where n is the particle number density, u^{μ} are the components of the 4-velocity vector and the ";" notation characterizes a covariant derivative. We can rewrite this equation on a coordinate basis as

$$\frac{1}{\sqrt{-g}}\partial_{\mu}(\sqrt{-g}\rho u^{\mu}) = 0, \tag{3.2}$$

where $\rho = mn$ is the mass density, m is the mass of the particle and g is the determinant of the metric tensor $g \equiv Det(g_{\mu\nu})$[1].

3.1.2 Conservation of energy and momentum

In general relativity, energy, and momentum are usually condensed into a single mathematical term called stress-energy tensor, which is a rank 2 tensor and is usually written as $T^{\mu\nu}$. This quantity describes the density and flux of momentum and energy in space. We can write the conservation equation as

$$T^{\mu}{}_{\nu;\mu} = 0. \tag{3.3}$$

Rewriting the equation on a coordinate basis

$$\partial_t(\sqrt{-g}T^t{}_\nu) = -\partial_i(\sqrt{-g}T^i{}_\nu) + \sqrt{-g}T^\kappa{}_\lambda\Gamma^\lambda{}_{\nu\kappa}, \tag{3.4}$$

where we have separated the time component and the spatial components represented with the index i. The term $\Gamma^\lambda{}_{\nu\kappa}$ is the well-known Christoffel symbol.

We still need to define what is the stress-energy tensor that describes our situation. We can usually categorize these tensor components as the following:

$$T^{00} = \text{energy density;} \tag{3.5}$$

$$T^{0i} = \text{energy flux across the surface } x^i; \tag{3.6}$$

$$T^{i0} = \text{the } i\text{th momentum density;} \tag{3.7}$$

$$T^{ij} = \text{flux of } i\text{th momentum across the surface } x^j. \tag{3.8}$$

This tensor is defined in such a way as to enforce the above physical meaning of each component.

We're dealing with a highly magnetized fluid, leading us to divide the tensor as a sum of two parts, a fluid one and an electromagnetic term

$$T^{\mu\nu} = T^{\mu\nu}_{fluid} + T^{\mu\nu}_{EM}. \tag{3.9}$$

We'll consider the fluid term describing a perfect fluid. We can write the stress-energy tensor as (Schutz, 2009, section 4.6)

$$T^{\mu\nu}_{fluid} = (\rho + u_e + p)u^\mu u^\nu + pg^{\mu\nu}, \tag{3.10}$$

[1] This passage from equation (3.1) to (3.2) can be found in Schutz (2009), p.153, eq. 6.42.

where u_e is the internal energy and p is the pressure. In general relativity, a perfect fluid is defined as a fluid with no viscosity and no heat conduction in the momentarily co-moving reference frame (MCRF).

We can describe the electromagnetic field as

$$T^{\mu\nu}_{EM} = \begin{pmatrix} \frac{1}{2}(E^2+B^2) & S_x & S_y & S_z \\ S_x & -\sigma_{xx} & -\sigma_{xy} & -\sigma_{xz} \\ S_y & -\sigma_{yx} & -\sigma_{yy} & -\sigma_{yz} \\ S_z & -\sigma_{zx} & -\sigma_{zy} & -\sigma_{zz} \end{pmatrix}. \tag{3.11}$$

Here $\frac{1}{2}(E^2+B^2)$ is the energy density in the electromagnetic field as described by Maxwell's theory, S_i is the poynting vector in the ith direction and σ_{ij} is the Maxwell stress tensor, calculated as

$$\sigma_{ij} = E_i E_j + B_i B_j - \frac{1}{2}(E^2+B^2)\delta_{ij}, \tag{3.12}$$

where δ_{ij} is the Kronecker delta. Rewriting this tensor in terms of the electromagnetic tensor (or Faraday tensor)

$$F^{\mu\nu} = \sqrt{4\pi}\begin{pmatrix} 0 & -E_x & -E_y & -E_z \\ E_x & 0 & -B_z & B_y \\ E_y & B_z & 0 & -B_x \\ E_z & -B_y & B_x & 0 \end{pmatrix}, \tag{3.13}$$

where we have absorbed the factor $\sqrt{4\pi}$ into the definition of the tensor, we arrive at

$$T^{\mu\nu}_{EM} = F^{\mu\alpha}F^{\nu}_{\ \alpha} - \frac{1}{4}g^{\mu\nu}F_{\alpha\beta}F^{\alpha\beta}. \tag{3.14}$$

Now, we adopt the ideal MHD regime, which describes the situation where the electric field vanishes in the fluid frame of the plasma due to the high conductivity, leading us into ignoring the Lorentz force, hence adding the following restriction to the equations $\vec{E} + \vec{v} \times \vec{B} = 0$. We can rewrite this as

$$u_\nu F^{\mu\nu} = 0. \tag{3.15}$$

We can define the magnetic field 4-vector as:

$$b^\mu \equiv \frac{1}{2}\epsilon^{\mu\nu\alpha\beta}u_\nu F_{\beta\alpha} \tag{3.16}$$

where ϵ is the Levi-Cevitta symbol defined as $\epsilon^{\mu\nu\alpha\beta} = -\frac{1}{\sqrt{-g}}[\mu\nu\alpha\beta]$, so that $[\mu\nu\alpha\beta]$ is the anti-symmetric symbol. With the aid of this definition, we'll simplify $T^{\mu\nu}_{EM}$ as

$$T^{\mu\nu}_{EM} = b^2 u^\mu u^\nu + \frac{1}{2} b^2 g^{\mu\nu} - b^\mu b^\nu. \tag{3.17}$$

To sum up, the full stress-energy tensor in MHD can be described as:

$$T^{\mu\nu}_{MHD} = (\rho + u + p + b^2) u^\mu u^\nu + (p + \frac{1}{2} b^2) g^{\mu\nu} - b^\mu b^n u. \tag{3.18}$$

3.1.3 The equations of the magnetic field evolution

The evolution of the magnetic field will be given by the source-free part of Maxwell's equations

$$\nabla \cdot \mathbf{B} = 0; \tag{3.19}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}. \tag{3.20}$$

Using tensors, these equations can be written as[2]

$$F_{\mu\nu,\lambda} + F_{\lambda\mu,\nu} + F_{\nu\lambda,\mu} = 0. \tag{3.21}$$

The other two Maxwell's equations[3]

$$\nabla \cdot \mathbf{E} = \rho; \tag{3.22}$$

$$\nabla \times \mathbf{B} = \mathbf{J} + \frac{\partial \mathbf{E}}{\partial t}, \tag{3.23}$$

will determine the current and charge density. In tensorial notation, this can be written as[4]

$$F^{\mu\nu}{}_{;\nu} = J^\mu. \tag{3.24}$$

By employing the dual tensor, the equation (3.21) can be represented as a conservation equation, meaning that the derivative of a specific quantity is identically zero

$$G^{\mu\nu}{}_{;\nu} = 0. \tag{3.25}$$

[2] See for example exercise 12.53 in Griffiths (1999).

[3] We are setting $\epsilon_0 = \mu_0 = 1$, just for convenience.

[4] Equation 12.126 in Griffiths (1999).

The dual tensor is generally defined as

$$G^{\mu\nu} = \begin{pmatrix} 0 & B_x & B_y & B_z \\ -B_x & 0 & -E_z & E_y \\ -B_y & E_z & 0 & -E_x \\ -B_z & -E_y & E_x & 0 \end{pmatrix}, \tag{3.26}$$

where we have $G^{t\nu} = B^{\nu}$, so we can rewrite the magnetic field four-vector as

$$b^t = B^i u^\mu g_{i_\mu}, \tag{3.27}$$

$$b^i = \frac{(B^i + b^t u^i)}{u^t}. \tag{3.28}$$

We can write the induction equation

$$\partial_t(\sqrt{-g}B^i) = -\partial_j[\sqrt{-g}(b^j u^i - b^i u^j)]; \tag{3.29}$$

$$\frac{1}{\sqrt{-g}}\partial_i(\sqrt{-g}B^i) = 0, \tag{3.30}$$

which is used to evolve the magnetic field in time.

In summary, the full set of GRMHD equations consists of the particle number conservation Equation 3.1, the four energy-momentum Equations 3.4, the MHD stress-energy tensor 3.18 and the induction Equation 3.29 with the constraint 3.30.

3.2 Radiative cooling

We follow the approach done in Esin et al. (1996) and Narayan and Yi (1995), based on the works of Svensson (1982) and Stepney and Guilbert (1983). Using this prescription we will describe the radiative cooling in an optically thin disk and an optically thick disk separately and link these two scenarios through a smooth function that depends on the total opacity of the medium (τ). All the cooling-related quantities depend on at most four parameters: scale height (H), magnetic field (B), electronic density (n_e), and electronic temperature (T_e). We start by describing the optically thin cooling processes.

The total cooling rate for an optically thin disk will be described as a sum of the Bremsstrahlung (Q^-_{brem}) and synchrotron (Q^-_{syn}) process. We also consider the effect of inverse Compton effects hardening the synchrotron radiation photons $\eta(\nu_c)$.

$$Q^-_{thin}(H, B, n_e, T_e) = Q^-_{brem} + \eta(\nu_c)Q^-_{syn}, \tag{3.31}$$

We don't consider the comptonization of bremsstrahlung radiation because (1) it would require solving a more complex problem since it is not as easy as what it is done in the comptonized synchrotron approach and (2) it has been shown by previous works that bremsstrahlung comptonization is not as important because synchrotron overcomes bremsstrahlung at temperatures in which Compton effect is important (Yoon et al., 2020; Fragile and Meier, 2009).

3.2.1 *Bremsstrahlung Radiation*

When electrons interact with the electric field of, for example, a heavy nucleus, they tend to decrease their kinetic energy and change the direction of the trajectory. The energy difference is emitted as electromagnetic radiation called *Bremsstrahlung radiation*. We assume two distinct cases: electron-electron interaction and electron-ion interaction

$$Q^{-}_{brem} = Q^{-}_{ei} + Q^{-}_{ee}. \tag{3.32}$$

For low temperatures ($10^4 K \lesssim T_e \lesssim 10^9 K$), bremsstrahlung dominates the emission.

3.2.1.1 *Electron-proton bremsstrahlung*

Considering a Maxwellian energy distribution, and using a cross-section calculated in Heitler (1984) in Born approximation, we can get the spectral emissivity

$$\frac{dE_{ep}}{dV dt d\omega} = N_p c \int_{1+\omega}^{\infty} \omega \frac{d\sigma}{d\omega} \beta n_e(\gamma) d\gamma, \tag{3.33}$$

where $\omega = h\nu/m_e c^2$ is the dimensionless photon energy, $\theta_e = k_B T_e / m_e c^2$ is the dimensionless temperature, $n_e(\gamma) = n_e \gamma^2 \beta \; exp(-\gamma/\theta_e)/\theta_e K_2(1/\theta_e)$ is the Maxwellian electron energy distribution, K_2 is the modified Bessel function of the second kind and n_e and N_p is the electron and proton number densities, respectively. Integrating (3.33) we get the value of the total emission. Svensson (1982) give us the following fit

$$Q^{-}_{ei} = \frac{dE_{ep}}{dV dt} = n_e N_p \sigma_T c \alpha_f m_e c^2 \begin{cases} 4\left(\dfrac{2\theta_e}{\pi^3}\right)^{1/2} [1 + 1.781 \cdot \theta_e^{1.34}]; & \text{if } \theta_e < 1 \\ \dfrac{9\theta_e}{2\pi} [\ln(2\theta_e \exp(-\gamma_E) + 0.42) + 1.5]; & \text{if } 1 < \theta_e \end{cases}$$

where α_f is the fine structure constant and γ_E is Euler's constant ≈ 0.5772. If we put the known constants in CGS we get

$$Q_{ei}^{-} = 1.48 \times 10^{-22} n_e N_p \begin{cases} 4\left(\frac{2\theta_e}{\pi^3}\right)^{1/2}[1 + 1.781\theta_e^{1.34}]; & \text{if } \theta_e < 1 \\ \frac{9\theta_e}{2\pi}[\ln(1.123\theta_e + 0.48) + 1.5]; & \text{if } 1 < \theta_e \end{cases}$$

giving us the final expression for electron-proton bremsstrahlung.

3.2.1.2 *Electron-electron bremsstrahlung*

The cross-section of this process is way more complex than the previous case. Haug (1975) gives an expression that can be integrated numerically. So the spectrum is given by

$$\frac{dN_{ee}}{dVdtd\omega} = n_e^2 \sigma_T c \alpha_f exp(-x) G(x, \theta_e)/x \tag{3.34}$$

where $x = \omega/\theta_e$. This function $G(x, \theta_e)$ is fitted for 13 different temperatures and interpolated so that, the total emission can be obtained by integrating the fit to the equation (3.34), or by integrating equation (3.20) in Haug (1975). The result is

$$Q_{ee}^{-} = \begin{cases} 2.56 \times 10^{-22} n_e^2 \theta_e^{3/2}(1 + 1.1\theta_e + \theta_e^2 - 1.25\theta_e^{5/2}); & \text{if } \theta_e < 1 \\ 3.42 \times 10^{-22} n_e^2 \theta_e (ln 1.123\theta_e + 1.28). & \text{if } 1 < \theta_e \end{cases}$$

Both Q_{ee}^{-} and Q_{ei}^{-} are given in $ergs/cm^3 \cdot s$.

3.2.2 *Synchrotron Radiation*

Synchrotron radiation is emitted when a relativistic electron under the influence of a magnetic field, accelerates, spinning around a magnetic field line. The magnetic field around the black hole comes from the accretion disk charged particles moving around.

This process will be crucial to describe the energy loss for most regions of inner hot accretion flows. Due to high temperature, electrons are relativistic, making synchrotron dominate over bremsstrahlung.

The equation of the power emitted by synchrotron radiation is found in Bekefi (1966), p.180,

$$\eta_\omega(\omega, \nu, \theta) = \frac{e^2\omega^2}{2\pi c}\left[\sum_{m=1}^{\infty}\left(\frac{cos\theta - \beta_\parallel}{sin\theta}\right)^2 J_m^2(x) + \beta_\perp^2 J_m'^2(x)\right]\delta(y). \tag{3.35}$$

Here, $\beta = v/c$, $\omega = m\omega_0/(1 - \beta_{\|}cos\theta)$, $\omega_0 = -eB_0/(\gamma m_0)$ is the synchrotron frequency, $J_m(x)$ is a second kind Bessel function, $J'_m(x)$ is the derivative of the second kind Bessel function and $\delta(y)$ is Dirac's delta function. The coordinates x and y are defined as

$$x = \frac{\omega}{\omega_0}\beta_{\perp}sin(\theta); \tag{3.36}$$

$$y = m\omega_0 - \omega(1 - \beta_{\|}cos\theta), \tag{3.37}$$

$\beta_{\perp}$ and $\beta_{\|}$ are the velocities perpendicular and parallel to the magnetic field line, normalized by the speed of light. A representation of the process can be seen in Figure 3.1.

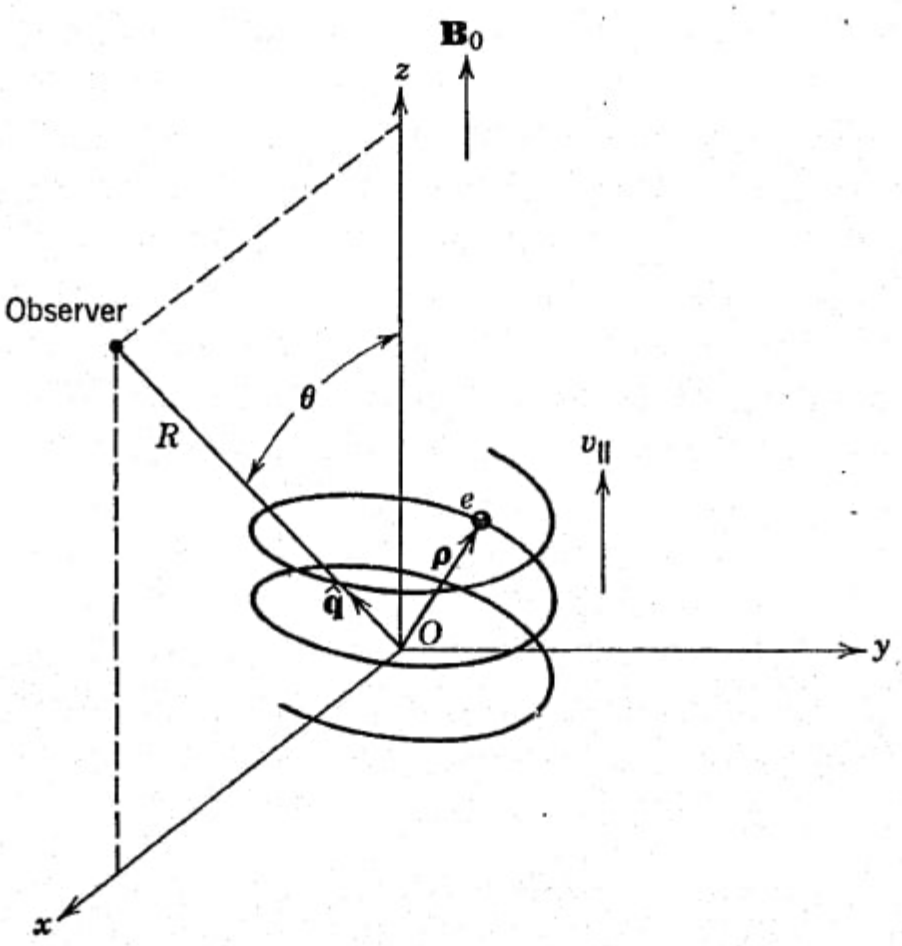

Fig. 6.1 Vector diagram for an electron in helical motion in a uniform magnetic field.

Figure 3.1: An electron spiraling around a uniform magnetic field line, emitting synchrotron radiation. Image taken from p.178 of Bekefi (1966).

If we want to assume a thermal distribution of electrons, we need to integrate the emissivity over a Maxwellian distribution of electrons $N(E) = N_0E^2exp(-E/k_BT_e)$. Pacholczyk

(1970) has calculated this result to be

$$\epsilon_s d\nu = 4.43 \times 10^{-30} \frac{4\pi\nu n_e}{K_2(1/\theta_e)} I\left(\frac{x_M}{sin\phi}\right) d\nu \; ergs \; cm^{-3} \; s^{-1} \tag{3.38}$$

where

$$x_M = \frac{2\nu}{3\nu_0\theta_e^2}, \tag{3.39}$$

$$\nu_0 = \frac{eB}{2\pi m_e c}. \tag{3.40}$$

ϕ is the angle between the velocity vector of the electrons and the direction of the magnetic field locally. $I(x)$ is a tabulated function. If we assume an isotropic velocity distribution, we can average over ϕ, resulting in this new function fitted by Mahadevan et al. (1996)

$$I'(x_M) = \frac{4.0505}{x_M^{1/6}}\left(1 + \frac{0.4}{x_M^{1/4}} + \frac{0.5316}{x_M^{1/2}}\right) exp(-1.8899 x_M^{1/3}). \tag{3.41}$$

We substitute this function for $I(x_M/sin\phi)$ in equation (3.38). These equations are valid only for optically thin emission, but, below a critical frequency ν_c, the emission becomes self-absorbed. This can be estimated as the frequency at which the synchrotron emission is equal to the blackbody emission in the Rayleigh-Jeans limit. If we consider a thin annulus of height $2H$, radius R, and thickness ΔR and make the blackbody condition happen in the upper and lower surface of the annulus, we'll get to the equation

$$\boxed{2H(2\pi R\Delta R)} \times \epsilon_s d\nu = \boxed{2} \times \boxed{(2\pi R\Delta R)} \times \boxed{\frac{2\pi\nu_c^2 k_B T_e}{c^2}} d\nu \tag{3.42}$$

- Volume of the annulus;
- Upper and Lower annulus contribution;
- Surface area of the annulus;
- Rayleigh-Jeans blackbody emission,

Finally, we can divide the synchrotron emission into two regimes. Below the critical frequency, the emission is self-absorbed completely, approximating the volume emissivity by a blackbody emission of the disk, divided by the disk volume. Above ν_c, the emission is optically thin and we can use the results from (3.38) averaged over ϕ

$$Q^-_{syn} = \frac{2\pi R^2}{2H\pi R^2} \int_0^{\nu_c} 2\pi \frac{\nu^2}{c^2} k_B T_e d\nu + \int_{\nu_c}^{\infty} \epsilon_s(\nu) d\nu. \tag{3.43}$$

Integrating over the emissivity, we get

$$Q^{-}_{syn} = \frac{2\pi k_B T_e \nu_c^3}{3Hc^2} + (6.76\times10^{-28})\frac{n_e}{K_2(1/\theta_e)a_1^{1/6}} \times \left[\frac{1}{a_4^{11/2}}\Gamma\left(\frac{11}{2}, a_4\nu_c^{1/3}\right) + \frac{a_2}{a_4^{19/4}}\Gamma\left(\frac{19}{4}, a_4\nu_c^{1/3}\right) + \frac{a_3}{a_4^4}(a_4^3\nu_c + 3a_4^2\nu_c^{2/3} + 6a_4\nu_c^{1/3} + 6)e^{-a_4\nu_c^{1/3}}\right], \tag{3.44}$$

where we define the parameters a_1, a_2, a_3 and a_4 as

$$a_1 = \frac{2}{3\nu_0\theta_e^2}, \qquad a_2 = \frac{0.4}{a_1^{1/4}}, \qquad a_3 = \frac{0.5316}{a_1^{1/2}}, \qquad a_4 = 1.8899a_1^{1/3}, \tag{3.45}$$

and

$$\Gamma(a,x) = \int_x^\infty t^{a-1}e^{-t}dt, \tag{3.46}$$

is the lower incomplete gamma function.

We encounter a problem when calculating the synchrotron emission rate for low temperatures ($T_e \lesssim 10^6\ K$) because the modified Bessel function of the second kind approaches zero $K_2(1/\theta_e)$. This is not a problem for Esin et al. (1996) because her regime considers $T_e \sim 10^9 - 10^{10.7}K$. Following Fragile and Meier (2009), $K_2(1/\theta_e)$ is replaced by $2\theta_e^2$ for low temperatures. It is important to mention that for $\theta_e < 1$, the synchrotron radiation turns into cyclotron radiation.

3.2.3 Comptonization of synchrotron radiation

Compton scattering occurs when a photon interacts with a charged particle, usually an electron. If the photon's energy is higher than the electron's, the photon will transfer energy to the electrons and this is called the Compton effect. If the photon's energy is lower, the electron will transfer energy to the photon and this is called the inverse Compton effect. Usually, what happens in the accretion disk, is relativistic electrons in the surrounding corona transfer energy to photons emitted by the disk, cooling the electron.

We treat the Compton cooling rate using a Comptonized energy enhancement factor η as done by Dermer et al. (1991). We define this factor as the average of the energy change of a photon since the injection of the photon in the medium until it escapes, divided by the initial energy

$$\eta = \frac{\langle\Delta E\rangle}{E_{\text{init}}}. \tag{3.47}$$

We begin a probabilistic approach where η is given by

$$\eta = \sum_{j=0}^{\infty} \mathcal{P}_j \mathcal{A}_j. \tag{3.48}$$

where $\mathcal{P}_j$ is the probability of an electron be scattered j times and then escape, while $\mathcal{A}_j$ is the average energy amplification of an electron scattered j times. Clearly, $\mathcal{A}_0 = 1$.

We'll consider we're inside the Thompson scattering regime, where the photon's energy is way smaller than the electron's rest mass energy, e.g., $\nu \ll m_e c^2/h$, so that we do not saturate to the Wien regime. This way, we can approximate $\mathcal{A}_j = A^j$, where $A \simeq 1 + 4\theta_e + 16\theta_e^2$ as described by Svensson (1984). If we consider the average number of scattering for a photon to saturate to the Wien regime as $j_{\max}$, where it can be approximately given by the relation $3\theta_e = E_{init} A^{j_{max}}$ so that $j_{\max} = \ln(3\theta_e/E_{\rm init})/\ln A$, then

$$\eta = \sum_{j=0}^{j_{\max}} \mathcal{P}_j A^j + A^{j_{\max}} \sum_{j=j_{\max}}^{\infty} \mathcal{P}_j. \tag{3.49}$$

We're considering that electrons that have entered Wien's regime gain no further energy on average, regardless of the number of subsequent scatterings.

Given the probability of a photon that travels a distance r without scattering being $\exp(-r/\lambda)$, where $\lambda = 1/n_e\sigma_T$ is the Thomson scattering mean free path and considering the scatterings as independent[5], from Lightman and Band (1981), we make a rough approximation for the escape probability of an electron that has scattered j times, such that

$$\mathcal{P}_j = (1 - e^{-\tau})^j e^{-\tau} = (1 - P)P^j \tag{3.50}$$

where $\tau = n_e \sigma_T H$, n_e is the electronic density, σ_T is Thomson's cross-section and H is a characteristic scale height of the scattering region. We define P^j as the probability of the photon scattering j times[6] and the term $(1 - P)$ represents the probability of escaping before scattering again. If we substitute Equation (3.50) in (3.49), using the geometric series

$$(1-u)^{-1} = \sum_{j=0}^{\infty} u^j, \qquad \frac{u^m}{1-u} = \sum_{j=m}^{\infty} = u^j \tag{3.51}$$

[5] We can consider them independent if the successive scattering events are not strongly dependent on the location of the previous scattering event, which happens for optically thin disks.

[6] Note that $\mathcal{P}_j$ is the probability of scattering j times and *then escaping*, while P^j is the probability of scattering j times.

and the value of j_{max}, we finally get to the equation

$$\eta = 1 + \frac{P(A-1)}{1-PA}\left[1 - \left(\frac{E_{init}}{3\theta}\right)^{1-lnP/lnA}\right]; \tag{3.52}$$

$$P = 1 - e^{-\tau}; \tag{3.53}$$

$$A = 1 + 4\theta_e + 16\theta_e^2, \tag{3.54}$$

where E_{init} can be calculated using

$$E_{init} = \frac{h\nu}{m_e c^2}. \tag{3.55}$$

So the comptonization factor depends on the frequency of the initial radiation. Because synchrotron peaks around the critical frequency, we consider the comptonization at this frequency, such that

$$\eta(\nu) = \eta(\nu_c). \tag{3.56}$$

3.2.4 Optically thick cooling

As this work aims to understand and implement the effect of cooling in the intermediate state where both thin and thick disks coexist, it is important to take the consequences of the denser (optically thick) and colder regions into consideration. Thin disks radiate as a modified multi-color blackbody.

Following Narayan and Yi (1995), we can assume the effective surface flux of a thin disk as

$$Q^-_{\rm Total} = \frac{4\sigma_T T_e^4}{H}\frac{1}{3\tau/2 + \sqrt{3} + 1/\tau_{abs}} \tag{3.57}$$

where $\tau = \tau_{abs} + \tau_{sca}$ is the total optical depth, τ_{abs} is the absorption optical depth and τ_{sca} is the scattered optical depth. We can approximate τ_{abs} as described in Narayan and Yi (1995)

$$\tau_{abs} = \frac{H}{4\sigma_T T_e^4}(Q^-_{thin}). \tag{3.58}$$

Notice that (3.57) is valid for both regimes. When $\tau \gg 1$, e.g, an optically thick regime, the expression can be approximated to a blackbody radiation limit

$$Q_{\rm Total} = Q_{\rm thick} = \frac{8\sigma_T T_e^4}{3H\tau}, \tag{3.59}$$

while when $\tau \ll 1$, e.g, an optically thin regime, (3.57) turns to (3.31).

The total cooling ends up depending on the four parameters described at the beginning of the section. The only parameter that is not local is the scale height of the disk, which is a global quantity and will depend on the radial coordinate. We will deal with this in an upcoming section, but for now, we manage to describe non-local processes such as inverse Compton effects into a local approximation.

3.3 Coulomb collisions

In the denser regions, electrons and ions collide transferring kinetic energy between one another. It becomes a very important mechanism to describe adequately the thermodynamics of denser environments, such as the optically thick disk, where these collisions will happen more often. To account for this, we follow the approach in Sądowski et al. (2016)

$$Q_{cc} = \frac{3}{2}\frac{m_e}{m_i}\frac{\rho}{m_p} n_e \log(\Lambda) \frac{c k_b \sigma_T (T_i - T_e)}{K_2(1/\theta_i) K_2(1/\theta_e)} \times \left[\frac{2(\theta_e + \theta_i)^2 + 1}{\theta_e + \theta_i} K_1(1/\theta_m) + 2K_0(1\theta_m) \right], \tag{3.60}$$

where the coulomb logarithm is considered to be $\Lambda \approx 20$, K_2, K_1 and K_0 are different kinds of modified Bessel functions, $\theta_m = (1/\theta_e + 1/\theta_i)^{-1}$. As you may notice, coulomb collisions will take into account one more parameter, which is the ion temperature (T_i), and will be implemented differently than the other cooling processes but we deal with that in the next chapter.

As a consequence of this effect Liska et al. (2022) found that in the optically thick disk, Coulomb collisions were sufficiently strong to thermalize the ion and electron temperatures ($T_i = T_e$), whereas in the optically thin flow, a two-temperature plasma is observable.

All the processes described in this section are treated locally.

3.4 The GPU-accelerated GRMHD code H-AMR

`H-AMR` is a GRMHD code derived from the `HARMPI` code (Tchekhovskoy, 2019), which, in turn, is derived from the `HARM2D` code (Gammie et al., 2003). `H-AMR` is a code written in C that has three levels of parallelization: (i) CUDA or OpenCL handles computation on GPUs or CPUs, (ii) OpenMP handles communication and gridding and (iii) Non-blocking MPI takes care of the transfer of boundary cells across nodes. Regarding the transfer of information across GPUs, NVLINK is used for intra-node transfers and GPU-DIRECT

for MPI transfer across different nodes. H-AMR is currently not an open-source code. It utilizes a finite volume method (FVM), treating the partial differential equations as integral conservation laws, ensuring conservation properties within each control volume. The finite volume method focuses on the conservation of quantities within the control volume and accurately captures the flow dynamics by considering the fluxes of these quantities across the control volume interfaces. To handle fluids with shock waves, H-AMR employs a shock-capturing technique known as the Godunov-based HLLE scheme. This scheme computes shock waves or discontinuities as part of the solution, without requiring additional techniques to treat these characteristics.

Godunov's scheme is a conservative numerical scheme for solving partial differential equations solving Riemann problems at each cell boundary. Riemann problems are specific initial value problems that arise when solving conservation equations together with constant initial data that has a discontinuity in the domain. These problems appear naturally in FVM due to the discreteness of the grid.

When the energy of a flow is predominantly kinetic, the use of a Godunov-type conservative numerical scheme with linearization of the Riemann problem can lead to the prediction of non-physical states, such as negative density or internal energy. For instance, in cases where the dominant energy of the flow is kinetic, the linearization of the Riemann problem may indicate a negative internal energy. To address this issue, a new method called the Harten-Lax-van Leer-Einfeldt (HLLE) scheme was proposed (Einfeldt, 1988). This scheme ensures positive conservation of internal energy and density, effectively avoiding the occurrence of negative values. The HLLE scheme incorporates specific stability bounds that depend on the absolute values of the maximal and minimal wave speeds.

H-AMR expresses the equations of motion in a conservative form

$$\frac{\partial \mathbf{U}(\mathbf{p})}{\partial t} = -\frac{\partial \mathbf{F_1}(\mathbf{p})}{\partial x^1} - \frac{\partial \mathbf{F_2}(\mathbf{p})}{\partial x^2} - \frac{\partial \mathbf{F_3}(\mathbf{p})}{\partial x^3} + \mathbf{S}(\mathbf{p}) \tag{3.61}$$

where $\mathbf{U}$ is the conservative quantities vector, such as particle number density, energy density, momentum density, $\mathbf{F}_i$ is the vector of fluxes in the i-th direction of the corresponding conserved quantity, $\mathbf{p}$ is the vector of primitive quantities such as particle density, fluid internal energy, velocity. Finally, $\mathbf{S}(\mathbf{p})$ accounts for the warping of the space-time grid and may include physical processes, such as nuclear heating, gas-radiation interactions, etc. The transformation of conserved variables to primitive ones is performed using the

well-known Newton-Raphson root-finding method or the Aitken acceleration scheme.

`H-AMR` uses an adaptive mesh refinement (AMR) mechanism which allows focusing the resolution on regions of interest. Although this is not very advantageous for RIAF, which spans most of the computational domain, it is very useful for simulating small-scale features, such as thin accretion disks, which are expected in this work.

Regarding the coordinate system used, spherical grids are the most natural and efficient ones. Among other reasons, they support a logarithmic spacing in the radial coordinate which provides higher resolution for regions close to the black hole and lowers the resolution progressively farther away. This is good because closer regions tend to have shorter timescales.

A major problem is dealing with polar coordinate singularity. They have implemented transmissive boundary conditions across the singularity to minimize dissipation. Also, spherical grids tend to squeeze the cells near the pole which causes the Courant condition (Courant et al., 1928) to limit the global timestep more than for a Cartesian grid. The Courant condition states that the distance that any information travels during the time step length within the mesh must be lower than the distance between mesh elements. If this condition is not achieved, instabilities are amplified and may cause divergence of the simulation, where the solution or behavior of the system becomes increasingly different or deviates significantly from the expected result. There are two main approaches to deal with this problem. One is the use of Cartesian grids and the other is to use static mesh refinement (SMR) to derefine the grid near the pole. The code combines the best of both approaches to deal with the squeezing of cells in the ϕ-direction, for a more detailed discussion see Liska et al. (2022).

`H-AMR` has yielded notable results so far, including the discovery that warped accretion disks drive both vertical and radial structural oscillations (Kaaz et al., 2022), contributions to the study of truncated disks (Liska et al., 2022), the understanding of black hole spin-down through magnetically arrested disks (Lowell et al., 2023), and the achievement of the highest resolution 3D GRMHD simulations for short-duration gamma-ray burst (sGRB) jets (Gottlieb et al., 2022).

Chapter 4

Radiative cooling implementation

In this section, I describe how the equations shown in Chapter 3 were implemented into the code and the difficulties involved in the process. As described by McKinney et al. (2014), we can write the evolution of the energy-momentum equation as

$$\nabla_\mu T^\mu_{\ \nu} = G_\nu, \tag{4.1}$$

where G_ν is the external 4-force density. When $G_\nu = 0$, the energy and momentum are conserved. It makes sense to implement the cooling inside this 4-force density vector. One can describe a radiation stress-energy tensor as

$$R^\mu_{\ \nu} = \begin{pmatrix} \hat{E} & \hat{F}^i \\ \hat{F}^j & \hat{P}^{ij} \end{pmatrix}, \tag{4.2}$$

where $\hat{E}$ is the radiation energy density, $\hat{F}$ is the radiation flux vector and $\hat{P}$ is the radiation pressure tensor. The radiation stress-energy tensor allows obtaining G^μ, which in the orthonormal fluid frame becomes

$$G^\mu = \begin{pmatrix} \tau_{abs}\hat{E} - \lambda \\ \tau_{tot}\hat{F}^i \end{pmatrix}, \tag{4.3}$$

as described by Mihalas and Mihalas (1984). τ is the optical depth and λ is the gas-fluid frame energy density emission rate of the gas. Boosting from the lab frame to fluid orthonormal frame, the covariant 4-force can be written

$$G^\mu = -(\tau_{abs} R^\mu_{\ \alpha} u^\alpha + \lambda u^\mu) - \tau_{es}(R^\mu_{\ \alpha} u^\alpha + R^\alpha_{\ \beta} u_\alpha u^\beta u^\mu), \tag{4.4}$$

where u^α is the 4-velocity vector.

We will not consider the addition of pressure and flux of the radiation in the surrounding matter, we'll only account for the energy emission of the plasma. So in our approach, we

only consider the energy depletion in each cell block of the grid. Because of this, we can consider that the 4-force density can be written as

$$G^{\mu} = -\lambda u^{\mu}, \tag{4.5}$$

where $\lambda = Q^{-}_{total}$ is the energy emission rate due to radiation.

The cooling function, called `cooling_function`, is a device function in the code, which means it runs on the GPU and is used to retrieve the cooling value from texture memory. This function is called within the source_texture_cooling function, which calculates the Equation 4.1. The `source_texture_cooling` function is invoked within the code's `source` function, which resides in the GPU_program1.cu file.

4.1 Calculating cooling in H-AMR

The most forward approach to deal with this problem would be to implement the equations described directly into the code and calculate it at every cell center at each timestep. This method by itself would be cheaper than the full radiative transfer with M1 closure implemented in H-AMR. Past results from our group Nemmen et al. (2023) have shown that the usage of lookup tables is also an alternative, faster method to implement cooling in simulations. A lookup table is a data structure used to store and retrieve pre-computed values. This way, we would calculate all the cooling before running the simulation for a range of each of the four parameters and retrieve the closest value from this table while running the simulation. This is an even faster way than computing the equations at each timestep.

Typically, the cooling values and parameters are passed to the code as an array, so at every timestep, the code analyzes the values of scale height (H), magnetic field $|B|$, electronic density (n_e) and electronic temperature (T_e) inside each cell and retrieves the cooling value that matches these parameters. Since we are using a GPU-accelerated code that is coded in CUDA/C, we decided to pass the cooling table as an array but allocate it inside the texture memory of the GPU. Previous works have done this kind of implementation in CUDA and found that the usage of texture memory is a significant way to improve the performance of the code. Schneider and Robertson (2017) allocated a lookup table using texture memory in the code CHOLLA (Schneider and Robertson, 2015) for simulating cooling in galactic winds.

To calculate the cooling table, we have independently developed a code that combines OpenMP and MPI, separate from `H-AMR`, that divides the calculation of the cooling equations into multiple threads and processes, which you can see the efficiency in Figure 4.1. In some simulations, we used tables with 100 million values, so establishing an efficient way to calculate the tables was extremely necessary. The very large tables were calculated using OLCF's supercomputer SUMMIT. The output files were stored in binary format to optimize the writing and reading processes, resulting in faster performance and reduced storage requirements. We also developed a code apart that stores the table in the texture memory of the GPU and allows the user to test texture memory interpolated values without necessarily running `H-AMR`. All the documentation and code are available at our Github. In the "H-AMR" git repository maintained by USP's black hole group, you can find a dedicated branch called "Cooling_pedro", which houses my modified version of the code. This branch incorporates detailed documentation of the newly implemented functions and variables.

Since `H-AMR` is not publicly available yet, these documents are also not publicly available yet. We also note that binary files must be handled carefully. Reading and writing binary data across different operating systems is not advised because of potential differences in endianness (ordering of bytes within a multi-byte data type) and data representation, which can lead to compatibility issues. This way, it is advised to generate the tables within the operating system that `H-AMR` will be run.

4.1.1 Quality Factor

Many physical processes take place inside the accretion disk and one must be able to resolve them in order to ensure the reliability of the results. Take local stress as an illustrative example. This quantity depends on many different processes that take place inside the accretion disk, such as the amplification of small magnetic fluctuations due to MRI, couplings between different wave modes, and the dissipation of the turbulence at small scales. To account for all these effects, the simulation must resolve the fastest-growing linear modes of MRI, the shortest length scale of the coupled wave modes, and the turbulence dissipation's small scales. If these processes are not resolved, the MRI is not able to transfer angular momentum outwards which stops the gas from spiraling inwards and induces the formation of blobs.

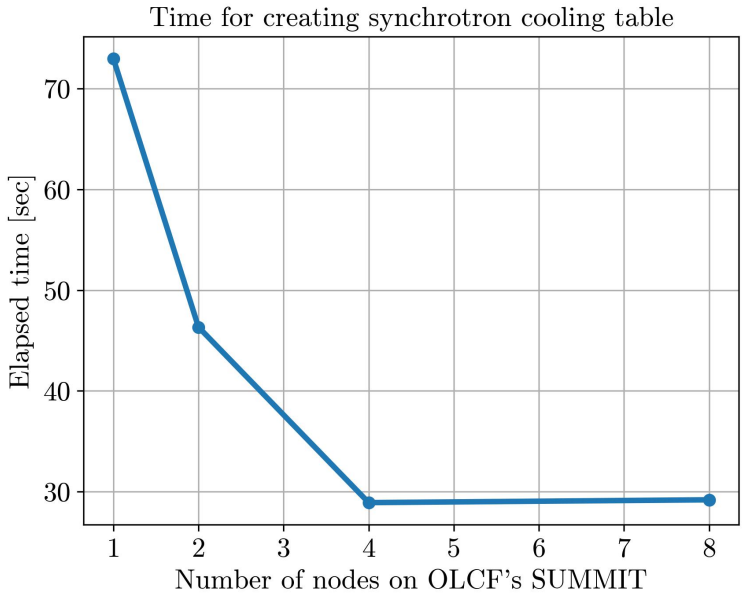

Figure 4.1: Time elapsed on my OpenMP+MPI table generation code per number of MPI ranks. This is the generation of the bremsstrahlung values in a 32^4 values table using OLCF's supercomputer SUMMIT. Each compute node on Summit contains two IBM POWER9 processors summing up to 168 OpenMP threads per node.

Usually when simulating geometrically thick disks (RIAFs), these problems are more difficult to arise due to the higher quantity of cells inside the disk, but when dealing with geometrically thin disks, the scales become smaller and one must have a quantitative parameter to ensure that the results are trustworthy. Hawley et al. (2011) analyzes different stratified shearing boxes and global simulations to characterize "quality factors"which are defined as the number of cells available for resolving the fastest-growing MRI mode

$$Q_i = \frac{2\pi v_A}{\Omega_{rot} dx^i}, \tag{4.6}$$

where $v_A = \sqrt{b_\mu b^\mu}/(\rho + B^2 + \gamma\epsilon)$ is the typical Alfvén speed, $\Omega_{rot} = \sqrt{u^\mu u^\nu g_{\mu\nu}}\delta^\mu_\nu / u^0$ is the rotational frequency and dx^i is the coordinate interval in each direction.

According to Hawley et al. (2011), the established value to solve for these modes is in the range $10 - 20$. We aim to stay inside this range for the whole simulation time. To tackle this, we set a target scale height $(H/R)_{\text{target}}$ following Noble et al. (2009). The temperature required to achieve a desired aspect ratio H/R in Newtonian gravity is

$$T_{\text{desired}} = \frac{\pi}{2}\left[\left(\frac{H}{R}\right)_{\text{target}} \Omega R\right]^2 \tag{4.7}$$

where $\Omega = 1/(r^{3/2} + a_{spin})$ is the relativistic orbital frequency in code units ($G = c = M = 1$). This way, we set a floor for each cell, if the temperature of the gas is below the

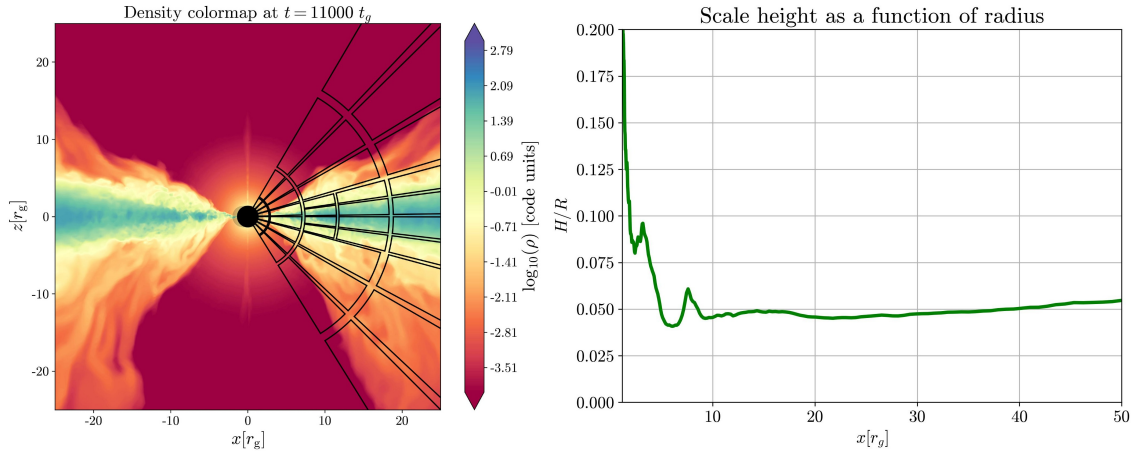

Figure 4.2: Density scale height plotted for a simulation with $(H/R)_{\text{target}} = 0.05$, with 1 level of SMR targeting the blocks near the midplane. Our base grid resolution is $(794 \times 384 \times 640)$ reaching $(1588 \times 768 \times 1240)$ in the refined region.

desired temperature, the cooling goes to zero in that cell. By using this, we can control the temperature floor and avoid the runaway cooling effect (Liska et al., 2022) also controlling the necessary resolution to solve the MRI within the geometrically thin disk as it is possible to see in Figure 4.2.

4.1.2 Necessary resolution to resolve MRI processes

Once we control the scale height of the disk, we can calculate the necessary resolution to resolve the MRI process. We do this by considering that very low values of H/R can be approximated as an angle. Ideally, you should have $16 - 32$ cells per $\theta = H/R$, however, 10 cells seem to be enough to resolve the MRI. Since θ ranges from 0 to π, it is easy to calculate the amount of H/R that fits inside this range, therefore calculate the total number of cells in the θ dimension. In `H-AMR`, it's very important to keep the aspect ratio of each cell as close to a cube as possible, so that $\Delta R = \Delta\theta = \Delta\phi$, allowing the calculation of the necessary resolution for the whole grid. We target a cube cell mainly due to numerical stability, finite volume methods are designed to work most stably and accurately on grids with similar spacing in all three spatial dimensions. We created a google sheets document where you can check if all the conditions are satisfied, as well as the necessary resolution for a certain scale height.

4.1.3 Implicit values for Cooling

The radiative 4-force G_μ can become stiff when its magnitude is significantly larger than the conserved quantities. To deal with this, we level the upper limit of the cooling function by the internal energy of the gas. If the cooling meets the criterion $u_t dt Q_{\rm tot} > 0.3|u_g|$, we adapt it as

$$Q_{\rm tot} \equiv \frac{0.3 u_g Q_{\rm tot}}{dt|u_t Q_{\rm tot}|}. \tag{4.8}$$

We choose the value of 30% of the internal energy to regulate the value of the cooling.

A realistic way to deal with this would be to utilize an implicit solver. H-AMR benefits from an implicit solver in the M1 closure radiative transfer mode, which uses an implicit-explicit (IMEX) Runge–Kutta scheme (Pareschi and Russo, 2005; McKinney et al., 2014).

An implicit solver is a numerical method used to solve time-dependent equations in computational simulations. It is particularly useful when dealing with stiff systems of equations, where the time scales of different processes vary significantly. In contrast to explicit solvers, which update the solution at a new time step using only information from the previous time step, implicit solvers consider both current and future time steps simultaneously.

In an implicit solver, the equations are formulated in such a way that they involve the unknowns at the current time step and the next time step. This creates a system of equations that needs to be solved simultaneously. The advantage of this approach is that implicit solvers are more stable and can use larger time steps, making them more efficient for stiff systems, which tend to be computationally challenging for explicit methods.

We also adapted the implicit solver from the code to work with my prescription. One can switch easily between the two approaches and can decide which one fits their problem best. Using our cooling prescription within the M1 closure scheme seems to be faster than calculating individual opacities for different radiative processes, but it slows down the code considerably compared to the internal energy approach.

4.1.4 Parameters

As described before, radiative cooling depends on four parameters and this subsection indicates how we calculate them in H-AMR.

RIAFs are characterized by a two-temperature plasma approach. This happens because

electrons cool down way more efficiently than ions due to the large mass difference between these particles. Following Liska et al. (2022), a two-temperature thermodynamics system was implemented, where they evolve the ion and electron tracers separately.

$$\kappa = p_{e,i}/\rho^{\gamma-1}, \tag{4.9}$$

which allows us to write

$$T_e = \frac{p_e m_H}{k_B}; \tag{4.10}$$

$$T_i = \frac{p_i m_H}{k_B}, \tag{4.11}$$

where we considered the mean molecular weight as $\mu = 1$.

The electron number density is calculated considering an ionized hydrogen torus, so that

$$\rho = n_e m_e + n_H m_H, \tag{4.12}$$

considering $n_e = n_H$ and $m_e \ll m_H$,

$$n_e \sim \frac{\rho}{m_H}. \tag{4.13}$$

The magnetic field value calculated at every cell is given by

$$B = \sqrt{B_\mu B^\mu}. \tag{4.14}$$

The remaining parameter necessary for calculating the cooling value is the scale height, which is a global quantity. We could approximate this quantity to a local temperature scaleheight defined as $H_T = T_e^4/\nabla(T_e^4)$, which was done for previous works (Yoon et al., 2020; Fragile and Meier, 2009). We tried to use this, but the result was very noisy and did not represent physical scale height values. We took advantage of our floor temperature and approximated the value of the scale height as $(H/R)_{\text{target}} \times R_{\text{cyl}}$, where $(H/R)_{\text{target}}$ is the targeted scale height of our prescription and R_{cyl} is the radial cylindrical coordinate. It is possible to check how H behaves in Figure 4.3. By doing so, we approximated the scale height to a realistic value.

4.1.5 Texture memory

One important aspect of computational performance is the efficient utilization of a computer's memory. To optimize memory access and accelerate computations, modern

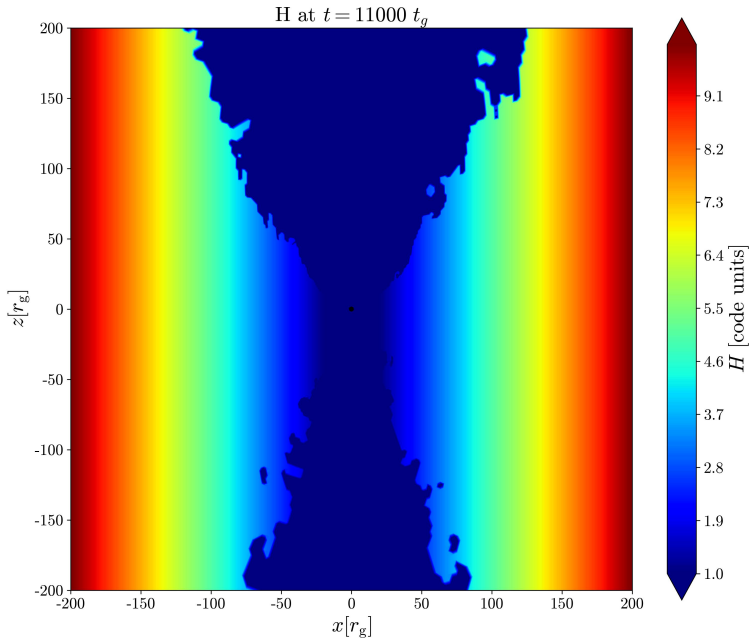

Figure 4.3: Scale height parameter calculated as $(H/R)_{\text{target}} \times R_{\text{cyl}}$ in code units.

graphics processing units (GPUs) provide a specialized type of memory called texture memory.

Texture memory is a high-speed memory space located on the GPU that is specifically designed to store and access 2D or 3D data, commonly known as textures. Traditionally, texture memory has been primarily used in computer graphics for rendering realistic images and textures on screen. However, its capabilities have been extended to other domains, including scientific computing, such as astrophysical simulations.

The key advantage of texture memory lies in its optimized memory access patterns. It is designed to exploit spatial locality, which refers to the tendency of data elements to be accessed in close proximity to each other. In the context of astrophysical simulations, texture memory allows efficient access to large arrays or grids of data representing physical properties such as radiative cooling.

Texture memory also offers additional features that can enhance computation. These include built-in interpolation capabilities, which allow for smooth and accurate sampling of data between discrete grid points. This will be useful for computing in-between values in our radiative cooling table.

Texture memory can be conceptualized as a grid-like structure where each cell stores a value at its center, as illustrated in Figure 4.4 for a 2D texture. When accessing the

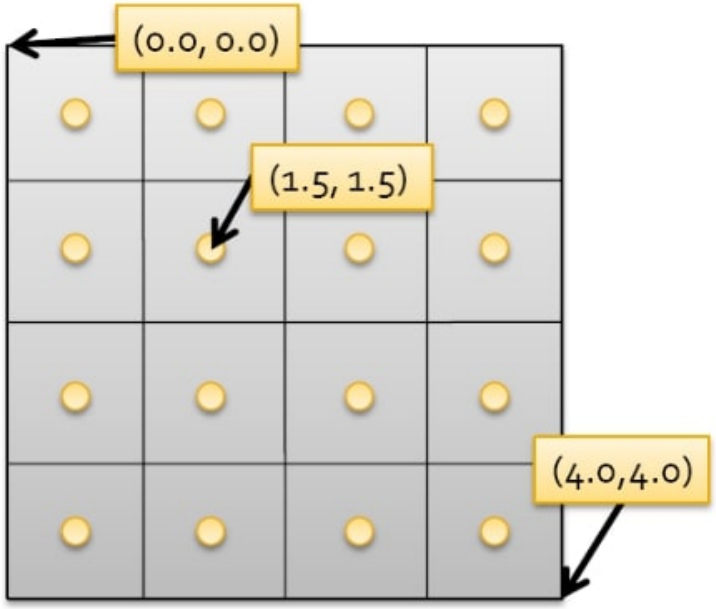

Figure 4.4: This is an example of a 2D texture grid, where the value of the table is stored at the yellow dots in the center of each cell. The coordinates indicate the position in the grid where you want to fetch the data (Moshovos, 2009).

value in a cell, you must specify the desired coordinates. Each coordinate is related to a parameter, in our case, the cooling value is determined by four parameters (H, B, n_e, T_e). To accommodate these four parameters, a 4D texture is needed. Since this feature was implemented for graphical rendering, 4D textures are not available. We employed 3D textures, where two of the parameters are flattened into a single dimension. We choose to flatten the dimensions of n_e and $T1_e$ into a single texture coordinate.

Handling the boundaries of the table is a crucial aspect, and CUDA offers various methods to address this challenge. In this case, the clamp mode was employed, which selects the closest cell to the specified value. For example, if the values fall below the minimum thresholds, such as $H < H_{\min}$, $B < B_{\min}$, $T_e < T_e^{\min}$, and $n_e < n_e^{\min}$, the code will retrieve the closest cell containing the value $Q_{\text{total}}(H_{\min}, B_{\min}, T_e^{\min}, n_e^{\min})$. The CUDA architecture also provides alternative methods like the wrap mode. The distinction between these methods is depicted in Figure 4.5.

To retrieve the appropriate data for the cooling calculations, it is necessary to navigate the texture grid. Let's consider the texture coordinates in this context. We define N_H, N_B, N_{ne}, and N_{Te} as the number of scale height, magnetic field, electronic density, and electronic temperature data points utilized, each equally spaced on a logarithmic scale. Additionally, we can denote C_H, C_B, and C_{ne+Te} as the respective coordinates within the

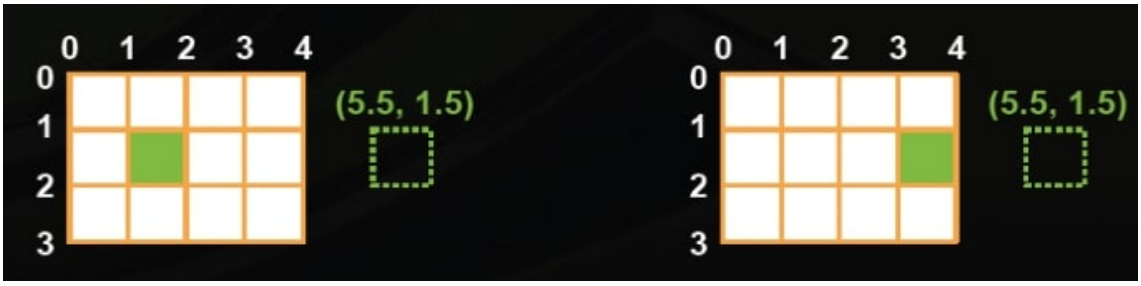

Figure 4.5: This picture portrays the main difference between the clamp mode and the wrap mode. Left: Wrap mode fetches the corresponding coordinate, repeating the grid in the out-of-bounds space. Right: The clamp mode will take the nearest value to the chosen coordinates, this is the mode used in this work (Moshovos, 2009).

grid. With that in mind, the mapping of each parameter into non-normalized coordinates follows

$$C_H = \left(\frac{\log_{10}(H/H_{\min})(N_H - 1)}{\log_{10}(H_{\max}/H_{\min})} \right) + 0.5; \tag{4.15}$$

$$C_B = \left(\frac{\log_{10}(B/B_{\min})(N_B - 1)}{\log_{10}(B_{\max}/B_{\min})} \right) + 0.5; \tag{4.16}$$

We make use of the `floor` which returns the largest integer less than or equal to a given number. For more details, refer to CUDA's documentation and programming guide.

In the $n_e + T_e$ dimension, we perform manual interpolation using a linear interpolation formula, similar to how it is done in texture memory. The interpolation in texture memory occurs when the index is not an integer. Let's consider a generic example with two neighboring indices, i and $i + 1$, representing the electron number density (n_e) values in our table.

When we choose a specific n_e value from the table, say $n_e = x$, the corresponding texture coordinate index will be an integer if x matches one of the n_e values in the table (i.e., x is equal to either n_e at index i or n_e at index $i + 1$). However, if we choose a value of x that lies between n_e at index i and n_e at index $i + 1$, then the index will not be an integer, and it will have a fractional part.

This fractional part indicates how close the chosen n_e value (x) is to index i or index $i + 1$, and the interpolation process will then calculate the intermediate value based on this proximity. In other words, the interpolation takes into account the fractional part to estimate the value between n_e at index i and n_e at index $i + 1$, allowing us to approximate the table value for the specific n_e in our calculations.

We opt for manual interpolation to have precise control over the interpolation process,

considering the fractional parts of both n_e and T_e individually. This approach enables us to achieve a highly accurate interpolation. By carefully handling the fractional components, we can ensure that the interpolated values align closely with the original data points, resulting in more reliable and precise calculations.

In our interpolation process, we also account for the specific point where we apply the approximation $K_2(1/\theta_e) \to 2\theta_e^2$ for the synchrotron process. This approximation represents a transition between two regimes with different variation rates. Failing to consider this transition properly during interpolation could potentially lead to inaccurate results.

To address this, we introduce conditional checks in the interpolation code to identify whether the current values of n_e and T_e lie within the region where the approximation is applied. If they do, we apply the appropriate approximation accordingly. However, for values outside this specific region, we ensure that the regular interpolation method is employed to maintain accuracy. The complete interpolation calculation can be found in our GitHub repository.

We compare the usage of global memory and texture memory to store table values and access data for different numbers of parameters. In the case of global memory, we employ a binary search algorithm. Binary search has a time complexity of $\mathbb{O}(\log n)$, where n represents the number of elements in the sorted list. This logarithmic time complexity significantly improves search speed compared to linear search, which has a time complexity of $\mathbb{O}(n)$ in the worst case. While we do not employ interpolation for the global array, we do utilize it for texture memory usage.

Once we retrieve the appropriate value from the table in both cases, we store it in a pointer. We select the parameters to be searched (H, B, n_e, and T_e) randomly, taking into account the total number of combinations: $130^4, 150^4, 180^4, 200^4$, and 230^4.

The comparison between global memory and texture memory usage is illustrated in Figure 4.6. We observe that for a small number of calculations, linear memory exhibits faster performance. However, as the quantity of requests increases, the time demand for texture memory decreases. Around 150^4 calculations, they both converge to a similar value. We anticipate that using a larger number of blocks and threads will further amplify the differences between the two memory types. In the comparison, we focus solely on the calculation time and do not consider the time required for creating the texture object or reading the table values from the binary file. This allows us to specifically assess the

performance differences based on the calculations themselves.

In our study, we conducted a comparison between the time it takes to retrieve table values from texture memory and the time required to calculate the cooling equations directly. This analysis allowed us to determine how much faster it is to compute values from a pre-calculated cooling table rather than letting the code perform calculations in real time. For this specific test, we conducted the cooling calculations for different numbers of iterations, specifically for $5^4, 10^4, 15^4, 20^4$, and 50^4 iterations.

The findings of this comparison are presented in Figure 4.7. Notably, we observed that as the number of calculations increases, the time difference between using texture memory and direct calculations remains relatively stable, consistently showing that texture memory usage is approximately 1000 times faster. However, as the number of calculations becomes even larger, this time difference seems to further widen, with the texture memory approach becoming nearly 10000 times faster compared to direct calculations.

These results demonstrate the significant advantage of utilizing pre-computed cooling tables in terms of computational efficiency. By employing texture memory to store and retrieve pre-calculated values, we can dramatically accelerate the cooling process and achieve substantial time savings, especially for scenarios involving a large number of calculations. Such optimization can significantly enhance the overall performance of the code.

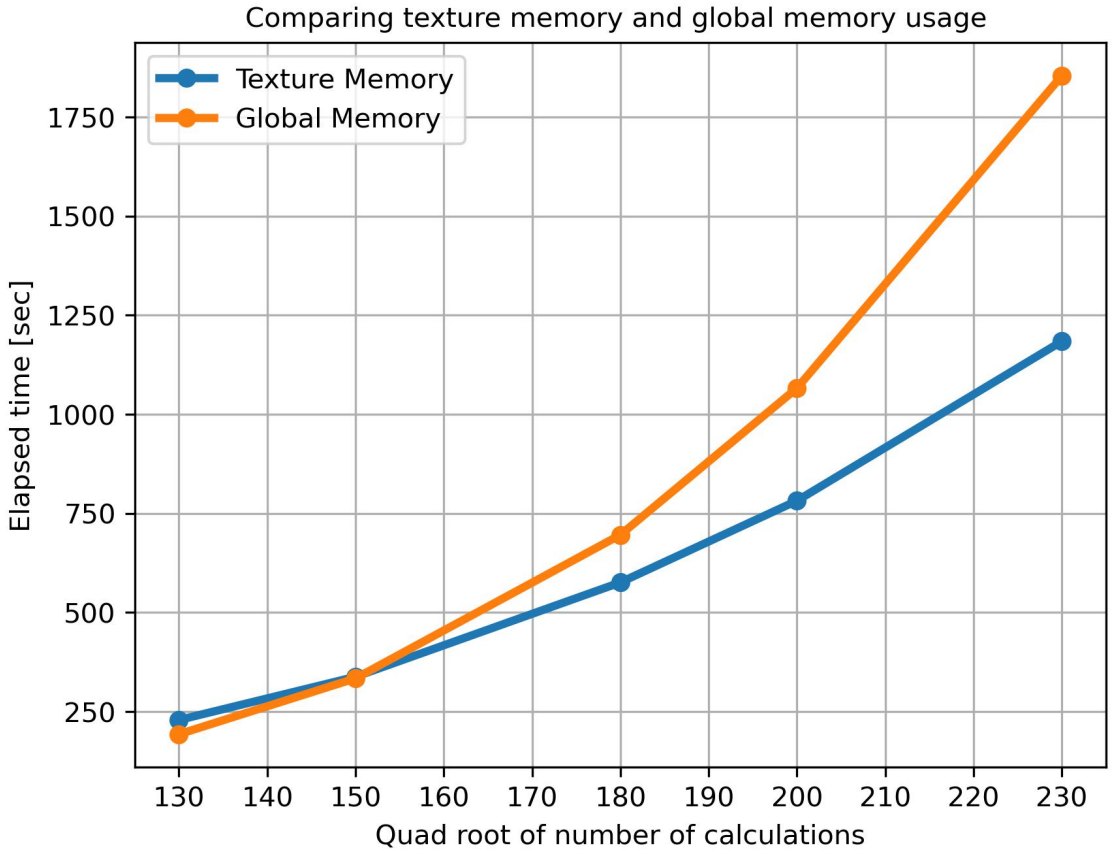

Figure 4.6: Comparison between texture memory and global memory in accessing table values. For this test, we make use of a single GPU GTX 1660S using 1 block and 1 thread of the GPU.

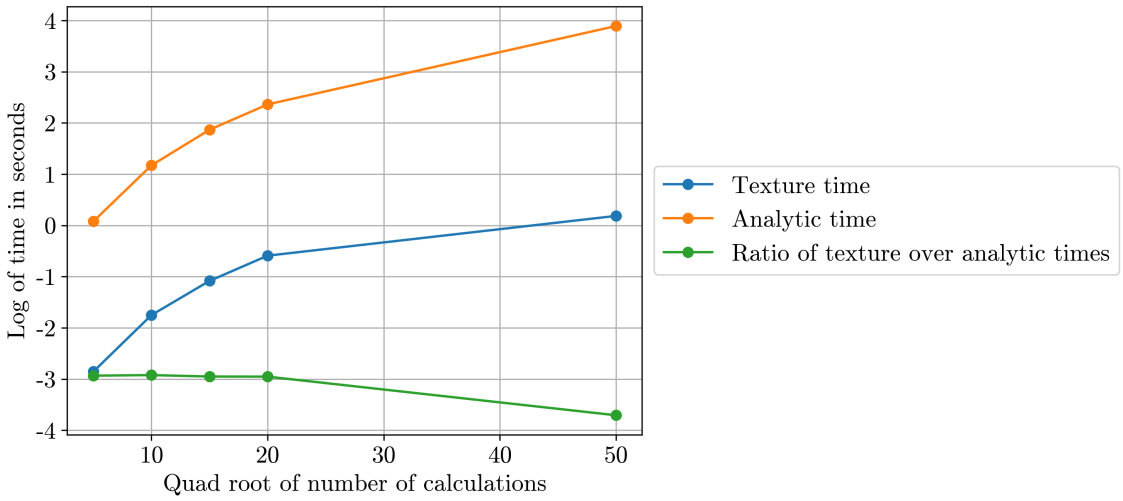

Figure 4.7: Comparison between texture memory and analytical equations calculation. For this test, we make use of a single GPU GTX 1660S using 1 block and 1 thread of the GPU.

Chapter 5

Validation of the cooling implementation

We performed a range of tests in order to test the validity of our cooling equations. In order to achieve that we performed low-resolution 2D simulations of an initial torus (Fishbone and Moncrief, 1976) with $r_{in} = 12R_g$ and $r_{max} = 25R_g$ in a grid with $r_{out} = 10^3 R_g$. We let the simulation run without cooling until $4000R_g/c$, reaching a stable accretion rate. We then turn on the cooling prescription.

To achieve the comparison between the different types of cooling, we implemented distinct texture objects referent to bremsstrahlung, black body approximation, comptonized synchrotron, Compton enhancement factor, and absorption opacity values alongside the existent coulomb collision and total cooling values.

We considered an accretion rate of $\approx 0.02\ \dot{M}_{edd}$ to test the results in optically thick and optically thin regimes separately since at this accretion rate, it allows us to have both regions. The resolution of these runs was (264×128) which is sufficient for the tests since we only tried to observe the type of emission from each optical limit.

In Figure 5.1 is possible to see the general properties of the gas. It is clear that the gas shows the formation of a "blob". This is an outcome due to not resolving the MRI in these regions. As a consequence, the gas doesn't spiral inwards as expected which leads to an accumulation of the gas in that region. This blob does not appear in higher-resolution simulations. Since we are only interested in addressing the different cooling processes, this is not a concern for the test run.

In Figure 5.2, we can see the different cooling processes in code units. To convert from code units to CGS, one needs to simply:

$$Q_{\rm cgs} = Q_{\rm code} \frac{\rho_{scale} c^2}{GM/c^3}, \tag{5.1}$$

where $\rho_{scale} = 1 \times 10^{-5}$ for this simulation, $M = 10M_{\odot}$, c is the speed of light and G is

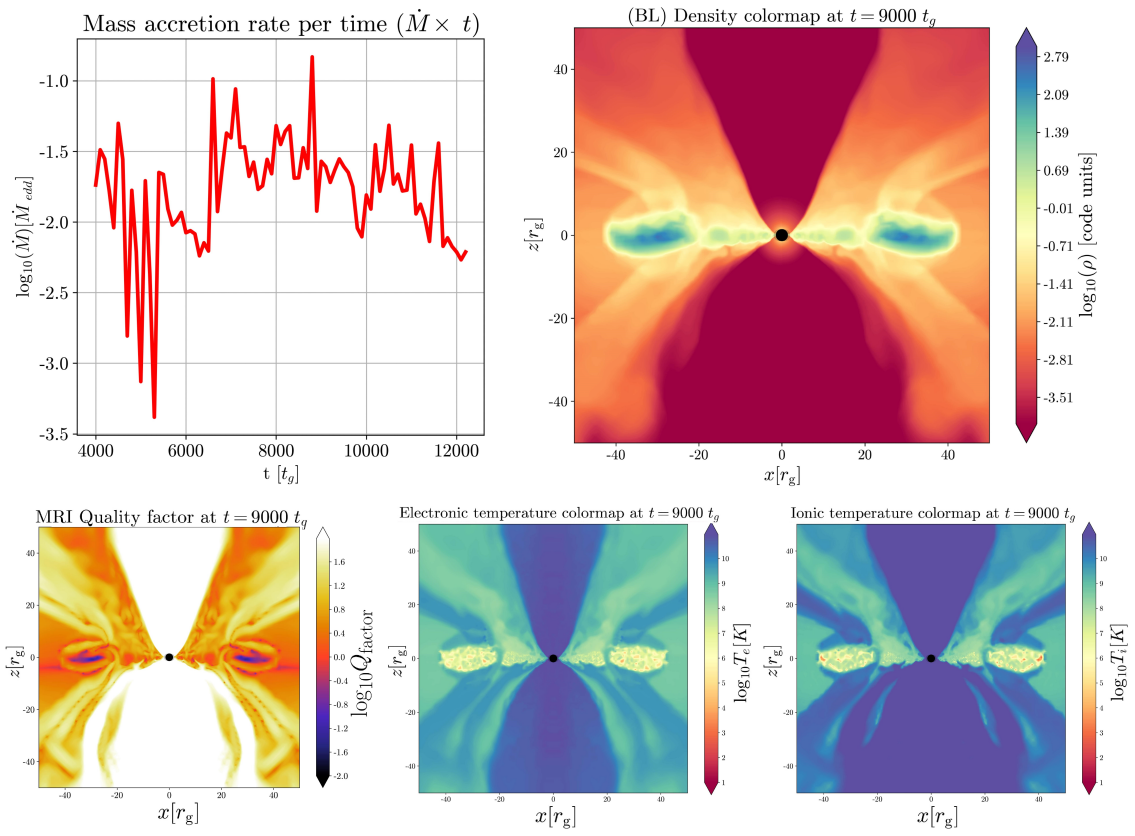

Figure 5.1: Top: (Left) accretion rate per time, (Right) density. Bottom:(Left) MRI quality factor, (Middle) electronic temperature, and (Right) ion temperature.

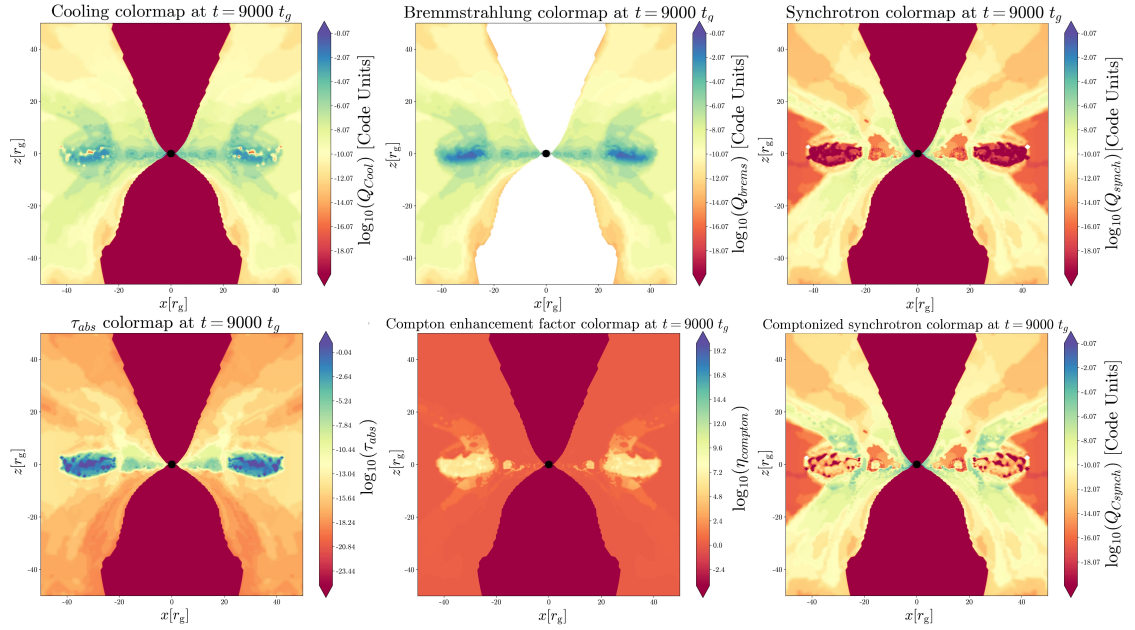

Figure 5.2: Top: (Left) Total cooling rate, (Middle) bremsstrahlung cooling rate, (Right) synchrotron cooling rate. Bottom:(Left) Absorption opacity, (Middle) Compton enhancement factor and (Right) comptonized synchrotron.

the gravitational constant.

5.1 Optically thick and optically thin regime tests

We expect the optically thick disk to radiate as a black body and the corona to radiate as $Q_{brems} + \eta(\nu_c)Q_{synch}$. The comptonization must happen primarily in the corona, where the hot electrons will scatter upcoming photons coming from the disk and enhance their energy. We can see from Equation 5.2 that the Compton enhancement factor is higher in the optically thick disk, instead of the corona. This is an expected outcome as higher density will allow more up-scattering of the photons and compensates for the lower temperature. However, this is misleading since our total cooling function for the optically thick limit (eq. (3.57)) does not consider optically thin emission where $\tau \gg 1$. In the high-density areas closer to the mid-plane, we have a higher Compton enhancement factor. We expect that in a resolved disk, we will be able to see an increase in the Compton enhancement area due to the formation of a corona.

Because equation (3.57) is an adaptation to meet both optically thick and optically thin regimes, we want to make sure that it is valid in our simulations. We compute the following quantities in order to analyze how the total cooling is correlated to the two optical limits

$$\delta_{\text{thick}} = \frac{|Q_{bbody} - Q_{tot}|}{Q_{tot}}; \tag{5.2}$$

$$\delta_{\text{thin}} = \frac{|(Q_{brems} + \eta(\nu_c)Q_{syn}) - Q_{tot}|}{Q_{tot}}. \tag{5.3}$$

We expect δ_{thin} to be zero above and below the thick disk and tend to higher values as we approach the mid-plane, since in this region, the blackbody cooling is stronger than Q_{thin}. As for δ_{thick}, we expect it to be close to zero in the mid-plane and much greater than one in the rest, since our blackbody cooling approximation is generally stronger than the optically thin emission. Both parameters can be seen in Figures 5.3 and 5.4. We also compared the zeroth and first-order expansion taken from equation (3.57) to both optical limits. It is clear that in both cases, the first-order approximation better describes the regions where $\tau \gtrsim 1$, as expected. We have done this test again for a higher resolution run and have encountered the same results, with the absence of the blob. We tackle this in Chapter **??**.

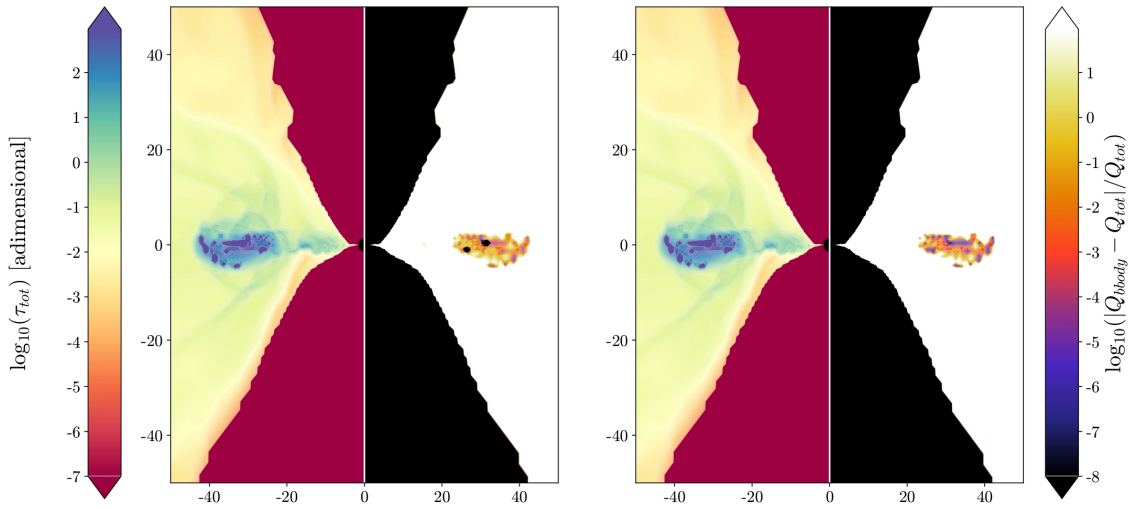

Figure 5.3: The images are split into two halves: the left side shows the distribution of the total optical depth (τ_{tot}), while the right side displays the (δ_{thick}) parameter. We compared the zeroth-order (left) and first-order (right) approximations of Equation (3.57) for the case where τ is much greater than 1. It is possible to see that first-order approximation better describes the region where $\tau \gtrsim 1$.

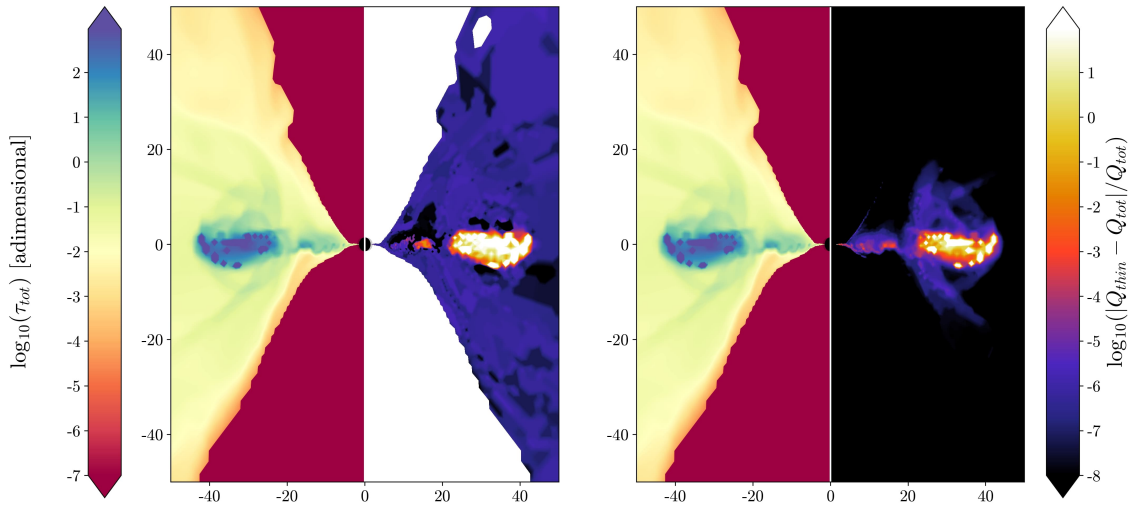

Figure 5.4: The images are split into two halves: the left side shows the distribution of the total optical depth (τ_{tot}), while the right side displays the (δ_{thin}) parameter. We compared the zeroth-order (left) and first-order (right) approximations of Equation (3.57) for the case where τ is much greater than 1. It is possible to see that first-order approximation better describes the region where $\tau \gtrsim 1$.

5.2 Analytical comparison to cooling table values

In order to validate the accuracy of our simulation results, we compared the calculated cooling values obtained from our lookup table with the analytical cooling solutions presented in Section 3.2.

Firstly, we captured a snapshot of the simulation at a particular time. From this snapshot, we extracted the four essential parameters required to calculate the cooling within the analytical equations. These parameters included the scale height, magnetic field value, electronic density, and electronic temperature associated with each cell in the simulation grid. Next, using the extracted parameters, we applied the analytical equations to calculate the cooling value corresponding to each individual cell in the grid. This allowed us to determine the expected analytical cooling value for every cell in our simulation.

Subsequently, we compared the calculated cooling values from our simulation's look-up table with the corresponding analytical values. Upon analyzing the results, we found that there was a maximum error of 45.72% between these two sets of values. We calculate the average and median of the error to be around $\sim 2.07\%$ and $\sim 0.2\%$, respectively. We attribute this maximum error to the interpolation function employed in the texture memory. The interpolation function utilizes a linear interpolation scheme for each dimension, and it introduces an interpolation error that is of the same order as the cooling values themselves. As a result, this interpolation-induced error can contribute significantly to the overall discrepancy observed between the simulation's lookup table values and the analytical cooling solutions. The overall color map showing the error can be seen in Figure 5.5.

5.3 Comparison with other methods

We also compared the values of Q_{thin} (3.31) and Q_{tot} (3.57) to previous results obtained by Marcel et al. (2019) who used code `BELM` (Belmont et al., 2008; Belmont, 2009) to take into account radiative processes in their two-temperature plasma code. `BELM` solves the time-dependent kinetic equations for isotropic and homogeneous distributions of electrons, positrons and photons, considering the effects of synchrotron self-absorbed radiation, Compton scattering, pair production/annihilation, Coulomb collisions, electron-proton bremsstrahlung radiation and some prescriptions for additional particle heating and acceleration. The comparison can be seen in Figure 5.6. The values are very similar in

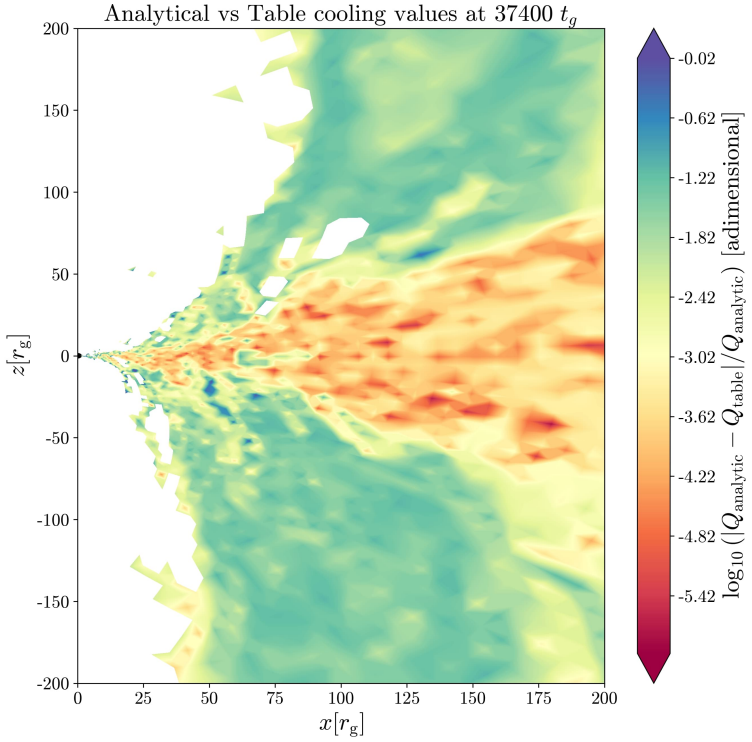

Figure 5.5: Comparison between the analytical cooling values and the lookup table values in logarithm scale. For this comparison, we used a 100^4 cooling table.

general but disagree mostly at $\tau > 1$ and $T_e < 10^8$. In this region, we expect to be in an optically thick regime and be dominated by blackbody cooling. To compare our values with theirs, we compute that $P_{tot} = n_e k_B T_e$ and get an averaged value for H, being $H/R \sim 0.1$ and $R \sim 30 r_g$. We use their definition of the magnetization parameter ($\mu = 0.1$) to calculate B as a function of the electron temperature (T_e) and the Thompson optical depth (τ_T)

$$B = \sqrt{4\pi\mu n_e k_B T_e}. \tag{5.4}$$

The comparison can be seen in Figure 5.6 where we compare the table values for Q_{tot} and Q_{thin} to the cooling values taken from `BELM` done by Marcel et al. (2019). It is evident that Q_{tot} provides a closer match to the values observed in the optically thick regime, characterized by $\tau > 1$ and $T \lesssim 10^8$. This can be attributed to the fact that our solutions for Q_{thin} are applicable only in the optically thin regime. To incorporate the optically thick behavior, we include a blackbody-like cooling limit within the expression for Q_{tot}.

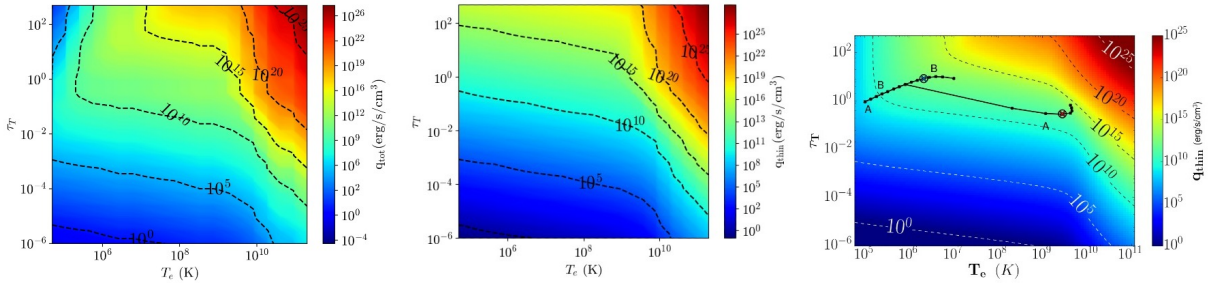

Figure 5.6: Comparison of the table cooling values Q_{tot} (3.57) (left) and Q_{thin} (3.31) (middle) to the interpolated values taken from BELM (right) (Marcel et al., 2019). They consider a magnetization parameter $\mu = B^2/\mu_0/P_{tot} = 0.1$, 20 values of optical depth τ_T in the range $[10^{-6}, 5 \times 10^2]$, and 20 values of electron temperature Te within $[5 \times 10^4, 2 \times 10^{11}]$K. We can clearly see that our Q_{tot} plot is more assertive in the values for low temperature and high τ. This is because Q_{tot} includes the blackbody treatment given to optically thick regimes.

5.4 Cooling of the wind

In our cooling prescription, we primarily utilize local quantities, such as magnetic field strength (B), electron number density (n_e), and electron temperature (T_e), which are all variables stored within each cell of the simulation. However, there is one exception - the scale height of the disk (H). To handle this, we adopt an approximation method: we determine an approximated value for the scale height by taking the targeted ratio of scale height to radius (H/R)targeted from the floor we impose and then multiply it by the cylindrical radius ($R_{\rm cyl}$) of the respective cell. This approximation works well within the disk region, where the $(H/R)_{\rm targeted}$ is part of the disk's structure. However, it may not accurately represent the scale height in other regions, such as the winds, where the physical conditions differ significantly from the disk.

To ensure numerical stability and avoid floor instabilities, we choose to set our cooling rate to zero in the jet region, defined as $B/\rho > 1$. However, we continue to apply cooling to the wind region. The areas that are being cooled are depicted in Figure 5.7 We want to assess whether the scale height approximation poses any issues for the wind's physics. To do so, we look at some conserved quantities in the winds.

In MHD theory, we have some conserved quantities along the field lines of the magnetic field. The Bernoulli integral can be written in MHD plasma as (Ferreira, J., 2002)

$$E = \frac{u^2}{2} + h + \Phi_g - \Omega_* \frac{r B_\phi}{\eta} - \int \frac{Q(s')}{\rho(s') u_p(s')} ds'. \tag{5.5}$$

where each term is, respectively, kinetic energy, enthalpy, gravitational potential, magnetic

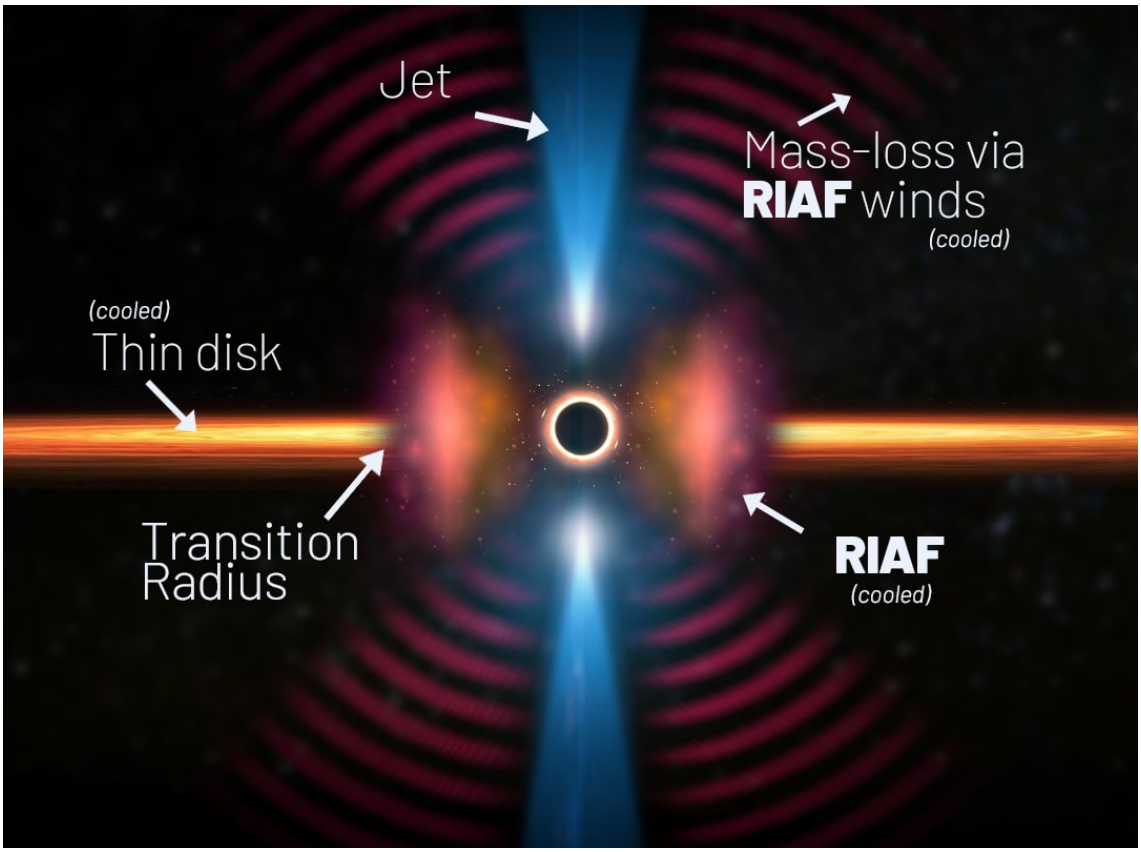

Figure 5.7: Image depicting black hole accretion flow regions that are being cooled. The image also shows a hybred disk.

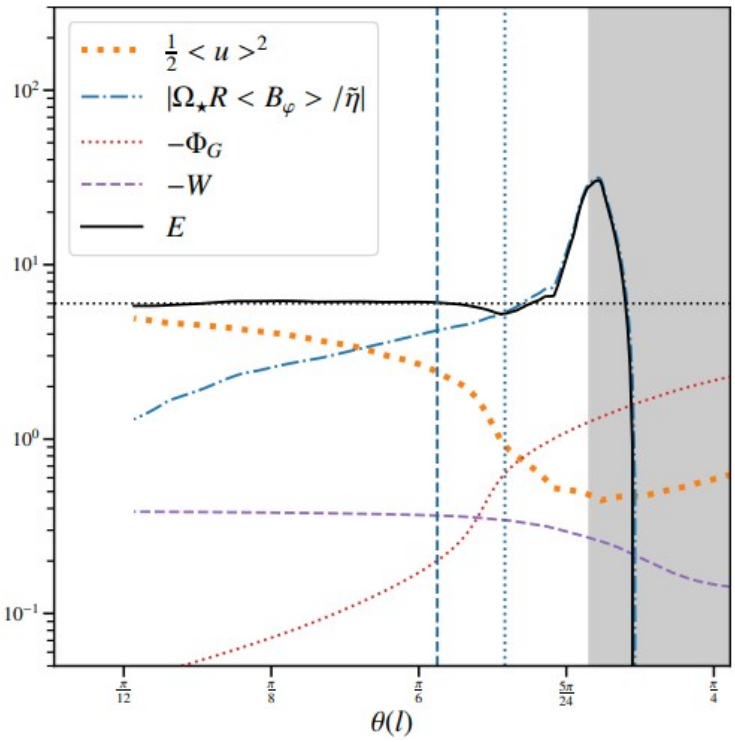

Figure 5.8: Bernoulli invariant and its component from MHD simulations done by Jacquemin-Ide (2021). The square-dotted orange line corresponds to the kinetic energy, the dashed-dotted blue line represents the magnetic torque, the dotted red line indicates the gravitational potential, the dashed purple line depicts the enthalpy, and finally, the solid black line represents the total quantity. The shaded grey area represents the turbulent atmosphere, while the vertical lines denote the various critical surfaces. The vertical dashed line indicates the location of the fast magneto-sonic surface, and the vertical dotted line represents the Alfvénic surface. The x-axis shows the theta angle between the polar axis and the magnetic field line. The y-axis represents the value of each quantity normalized.

torque, and cooling integral along the streamline. We then compare the value of enthalpy to the value of the cooling integral along the field line to determine whether the cooling is impacting the conserved quantities and if the scale height approximation imposes a problem on the wind. We choose the enthalpy because it is the smallest component of the Bernoulli integral near the disk. The impact of each term can be seen in Figure 5.8.

For this test, we performed a 3D simulation with resolution $(600 \times 336 \times 320)$ setting our cooling floor to $H/R_{\rm target} = 0.1$. We start the simulation with the standard Fishbone and Moncrief (1976) solution, using $r_{\rm in} = 20r_g$, $r_{\rm max} = 41r_g$ and normalize the magnetization parameter $\beta = 100$. We use the following vector potential

$$A_\phi = \frac{\rho}{\rho_{\rm max}} \left(\frac{r \sin\theta}{20.} \right)^3 e^{\frac{-r}{400}} - 0.2, \tag{5.6}$$

in order to achieve a MAD saturation regime. We check the magnetic flux in Figure 5.9. We let the simulation run up to $29600r_g/c$ before initializing the cooling which acts until

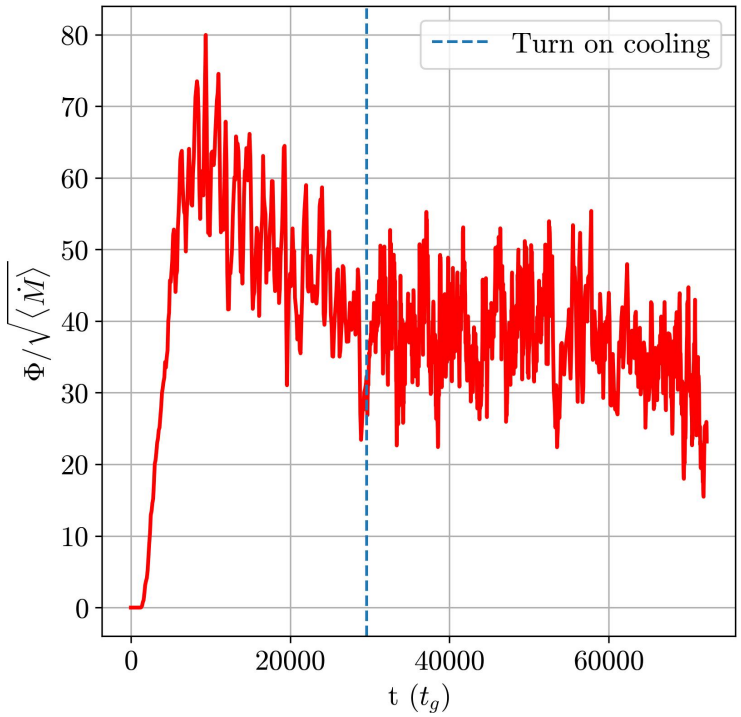

Figure 5.9: Magnetic flux normalized by Mass accretion rate. We expect a MAD with magnetic saturation around ~ 50 (Tchekhovskoy and McKinney, 2012) for a non-cooled disk. With cooling, we expect a MAD saturation around ~ 30.

we finish the simulation at $72327 r_g/c$. Our analysis of the winds consists of analyzing the time-averaged quantities over the range of $67327 - 72327 r_g/c$.

Firstly, we extract the field lines directly from A_ϕ and plot over the color map of enthalpy in Figure 5.10. We can see that the field lines follow the contours of the enthalpy, which is already a good indication that the conditions in the wind are being preserved far from the disk. We also plot the streamlines over the density color map in Figure 5.11. We observe that the contour that represents twice the pressure scale height matches better with the frontier between the wind and the disk, while the density scale height defines the denser areas of the optically thick disk.

To perform the analysis, we select some of the field lines that do not cross the event horizon and are not entangled within the disk structure. We then extract these lines' coordinates and plot the enthalpy along the field line, Figure 5.12. It is possible to see that the cooling affects the parts of the wind that are very close to the disk, but overall it reaches an almost constant value away from the disk.

Finally, we compute the importance of the cooling term along the field lines when compared to the enthalpy in Figure 5.13. So our cooling is affecting the parts of the

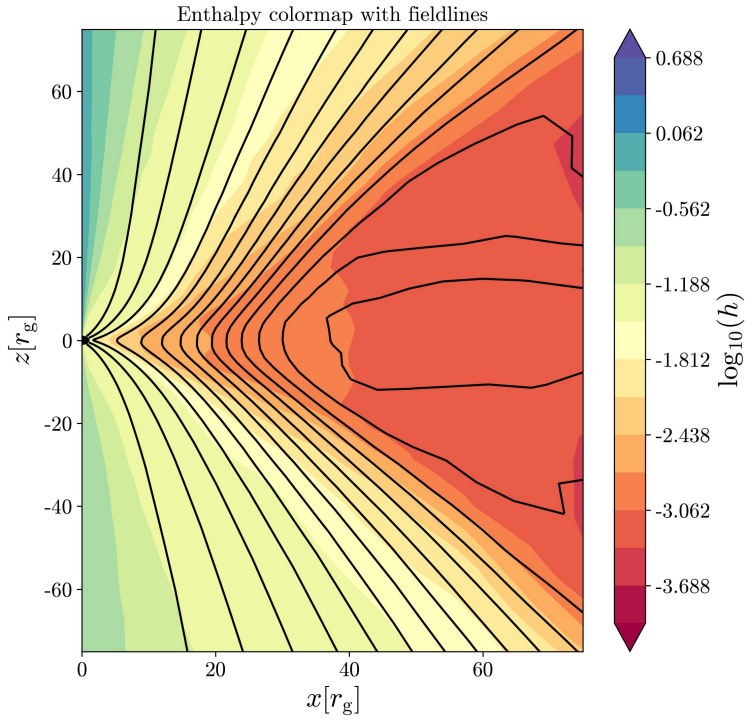

Figure 5.10: Representation of the field lines over the color map of enthalpy in code units averaged over the last $5000r_g/c$.

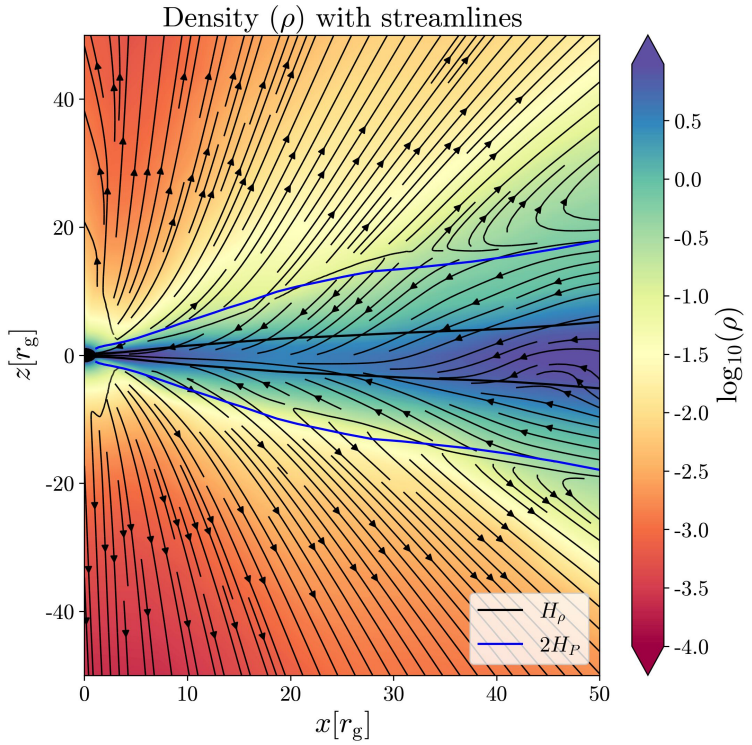

Figure 5.11: Representation of the streamlines over the color map of density in code units averaged over the last $5000r_g/c$.

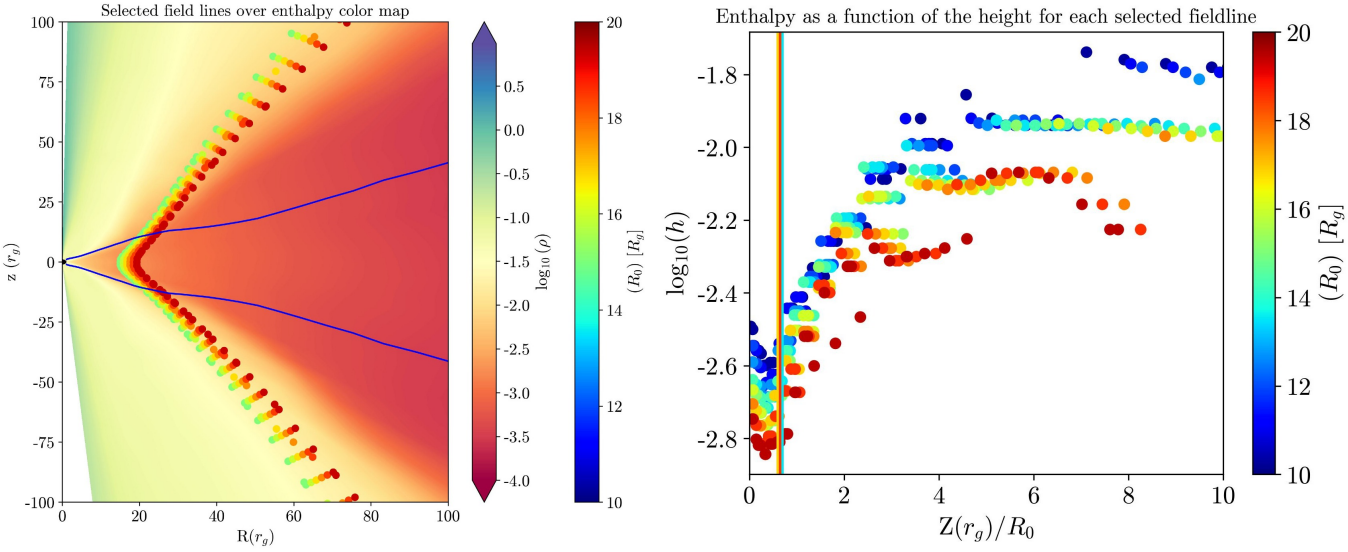

Figure 5.12: (Left) Representation of some selected field lines, each color represents a different line, and the blue line represent twice the pressure scale height. The enthalpy is plotted in the back as a color map. (Right) We plot the enthalpy along each field line. Each color represents a different field line with different starting positions (R_0). The vertical lines indicate the height at which each field line reaches twice the pressure scale height.

wind that are relatively close to the disk up to twice its initial radius R_0, e.g since our field lines start at around $\sim 10 - 20r_g$, our cooling is having an impact on them up to $\sim 20 - 40r_g$. Despite this occurrence, it is evident from Figure 5.8 that enthalpy exhibits the smallest magnitude near the disk. Consequently, one or more of the remaining terms in the Bernoulli invariant might be stronger than the cooling effect.

5.5 *Validation of the coulomb collisions implementation*

We also performed tests to guarantee the validity of the Coulomb prescription presented in Equation (3.60). The coulomb collisions account for the transfer of energy between ions and electrons, which is important in a two-temperature simulation to adequately represent the thermodynamics of the system. For this, we ran low resolution (264×112), 2D simulations of a torus (Fishbone and Moncrief, 1976) with $r_{in} = 12R_g$ and $r_{max} = 25R_g$ in a grid extended to $r_{out} = 10^3 R_g$. There is no presence of magnetic fields in this simulation, which is done to avoid the evolution of the disk.

We start the simulation with active coulomb collisions but without cooling effects. We set the ion temperature to be $T_i = (\Gamma - 1)u_g/\rho^\Gamma$ and $T_e = 0.01T_I$, where $\Gamma = 5/3$ is the adiabatic index, u_g is the internal energy of the gas and ρ is the mass density. We set an

Importance of cooling compared to enthalpy: cooled averaged 5000 t_g, each color represents a field line

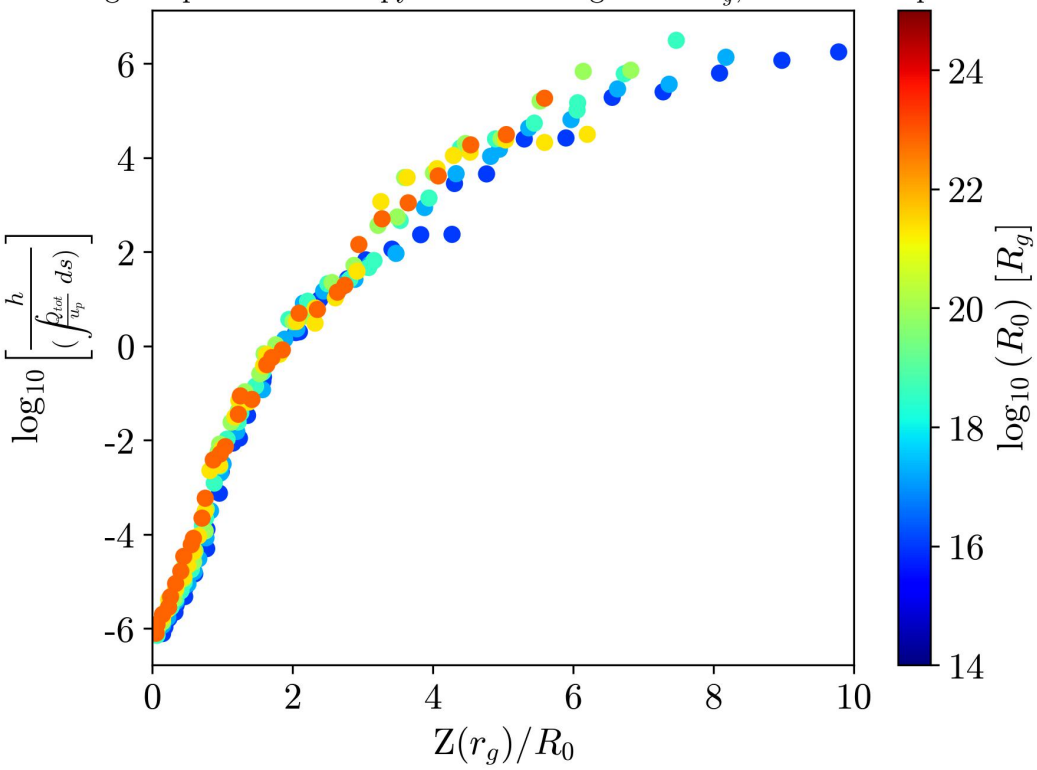

Figure 5.13: Ratio of enthalpy and line integral of the cooling function for each selected field line, with proper dimensions.

upper energy transfer limit of 30% of the internal energy of the gas in the cell. We see that after $t \approx 5000 t_g$, the two temperatures seem to be very close (Figure 5.14). Since the coulomb prescription depends on the difference in temperature, the thermalization becomes slower as the temperatures approach each other.

We solve the equation analytically in order to compare the simulation results. The variation of the temperature over time due to the coulomb process is described in equations (60) and (61) of Sądowski et al. (2016). Without any cooling process, the equations can be written as:

$$\frac{dT_i}{dt} = -\frac{m_p(\Gamma - 1)}{k_B \rho} Q_c; \tag{5.7}$$

$$\frac{dT_e}{dt} = +\frac{m_p(\Gamma - 1)}{k_B \rho} Q_c, \tag{5.8}$$

where m_p is the proton mass, k_B is the Boltzmann constant and Q_c is the value of the energy transfer rate via coulomb interaction. Since the gas is static, ρ can be treated as a constant in time, allowing us to integrate this equation via an iteration method that can be described as

$$T_i^{n+1} = T_i^n - \frac{m_p(\Gamma - 1)}{k_B \rho} Q_c^n dt; \tag{5.9}$$

$$T_e^{n+1} = T_e^n + \frac{m_p(\Gamma - 1)}{k_B \rho} Q_c^n dt, \tag{5.10}$$

where n is our iteration index and dt is the size of the timestep.

We analyze the evolution of the temperature for a single cell in the midplane of the gas near $r = 50 r_g$. This cell in particular has an averaged electronic density of $n_e = 1.39 \times 10^{17}$ g/cm^3. We also average the size of the timestep to a value of 1.06×10^{-6} s, assuming it doesn't change much throughout the simulation. With these values, we get to the comparison depicted below (Figure 5.15).

The slight difference in values between the simulation and the analytical prediction is expected due to the approximations made. Although we treated the electronic density as a constant, it varies approximately 0.1% around the averaged value during the simulations. The coulomb prescription in the code is subjected to the upper limit value of 30% of the energy of the cell. The timestep varies within 10% of the average value. This simulation was configured to dump data files at every $5 t_g$ between $t = 0 t_g$ and $t = 400 t_g$ and at every $25 t_g$ later on.

A larger difference arises when the temperatures get too close together. This is a limitation of dealing with a lookup table prescription. As $T_i/T_e \to 1$, $Q_c \to 0$. The interpolation made by texture memory seems to decay faster than the analytical equation, which slows down the thermalization for the simulation. We believe that a larger, i.e. more detailed table can decrease this error.

The coupling of the electronic temperature and ionic temperature is not expected to happen within the optically thin regime and the error begins to be visible when $T_i/T_e \approx 1.1$, so we don't think this error will have a high impact on the physics of the disk.

In summary, we conducted several tests to evaluate the accuracy and reliability of our cooling and Coulomb collision functions. We compared the interpolated values with analytical solutions and examined the behavior of our approximated scale height in the wind region. Additionally, we benchmarked our cooling prescription against previous works(Marcel et al., 2019) to ensure consistent and comparable results.

The outcomes of all these tests indicate that our cooling prescription functions effectively within the code. The interpolated values closely match the analytical ones, and the approximated scale height shows minimal impact on the wind dynamics. Furthermore, the comparison with previous works demonstrates that our cooling processes yield results in good agreement.

Overall, the successful outcome of these tests confirms that our cooling prescription is robust and reliable, making it an efficient and practical method for handling cooling dynamics within our simulations.

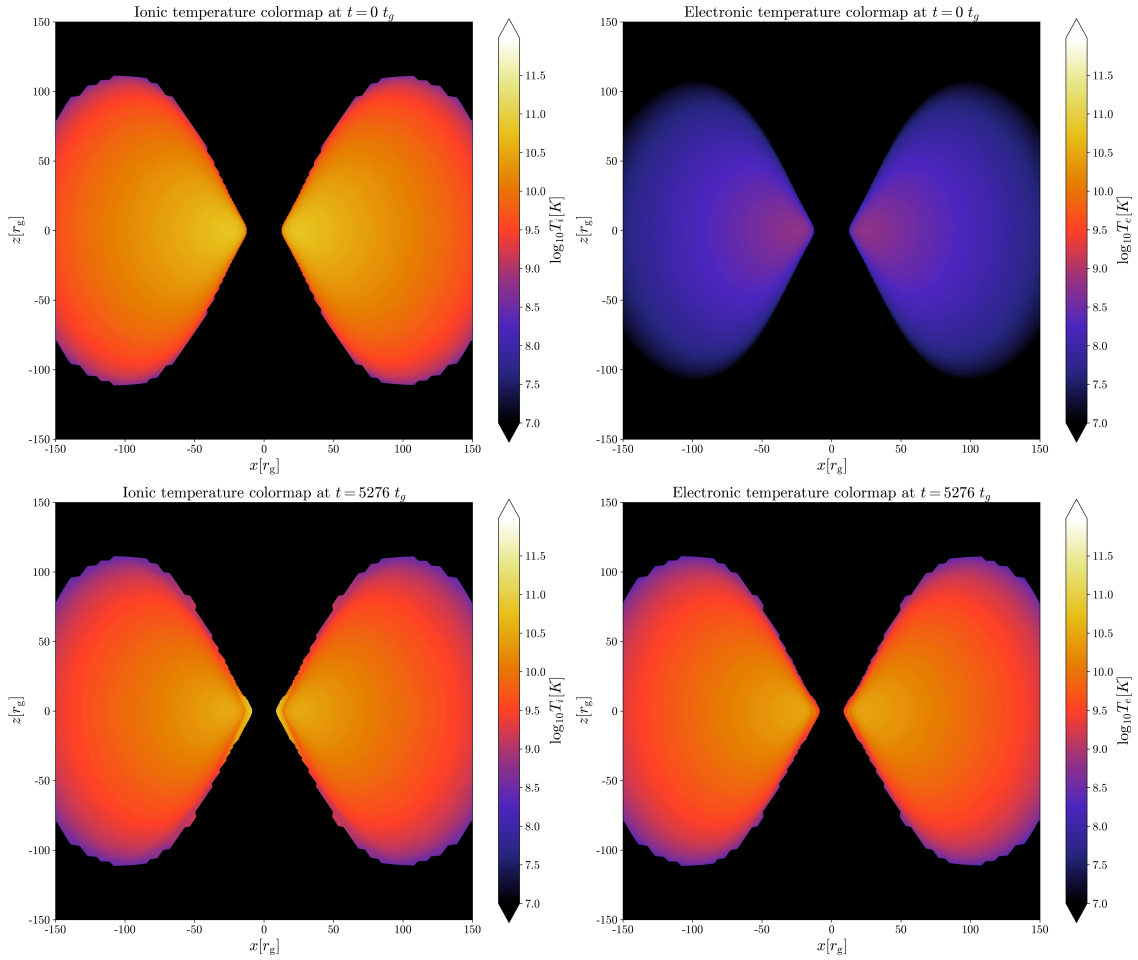

Figure 5.14: Electron (right) and ion (left) temperatures for different simulation times, $t = 0t_g$ and $t = 5276t_g$

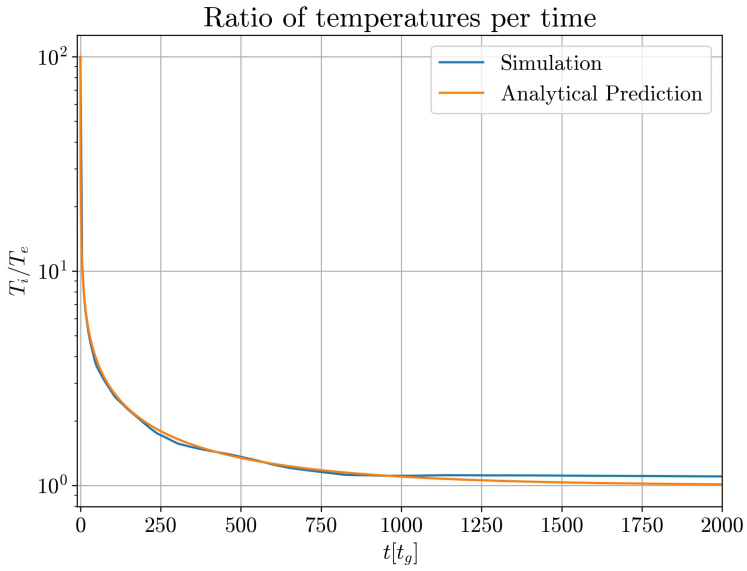

Figure 5.15: Comparison between analytical and simulation results of the ratio of ionic and electronic temperature per time. The orange line is the result of iterations of Equation (5.10), and the blue line is the result of the simulation.

Chapter 6

Results

Many different simulations were carried out during the testing analysis. Here, we demonstrate how the cooling affected the following physical quantities: Density, ion temperature and electron temperature. We also compare the importance of each cooling component for total cooling and check how the cooling behaves in a 3D simulation.

For this purpose, we set up a high-resolution 3D model. We assumed an initial torus following Fishbone and Moncrief (1976), with $r_{\rm in} = 12R_g$ and $r_{\rm max} = 25R_g$ in a grid that extends to $r_{\rm out} = 2000R_g$. We initiate the cooling process from the very beginning, in order to achieve a MAD regime, following Liska et al. (2019, 2022), which sets

$$A_\phi = \left(\frac{\rho}{\rho_{\rm max}} - 0.05\right)^2 r^2. \tag{6.1}$$

In this scenario, they chose $(H/R)_{\rm target} = 0.03$, for which we decided to choose $(H/R)_{\rm target} = 0.05$, to decrease the computational cost of the simulation. We used a base resolution $(794 \times 384 \times 640)$ reaching $(1588 \times 768 \times 1240)$ with one level of static mesh refinement (SMR) refining the closest upper and lower midplane blocks from $r = 3R_g$ to $r \approx 100R_g$.

We ran the simulation for $11000R_g/c$, using 96 GPUs. The main parameters are displayed in Figure 6.1.

Due to coulomb collisions, we can see that the collapsed parts, e.g, the denser regions are able to reach a thermal equilibrium where $T_i = T_e$, while the less dense gas presents a two-temperature medium. The mass accretion rate of this run had an average of $\langle \dot{M} \rangle = 0.35$ and a magnetic flux saturation around $\Phi \sim 20$.

After $11000R_g/c$, we lower the $(H/R)_{\rm target}$ to 0.01 and let it run for $\sim 900R_g/c$. Although the resolution is not high enough to resolve the MRI in the denser regions, we can compare and discuss some of the results. From Figure 6.2, it is possible to see that

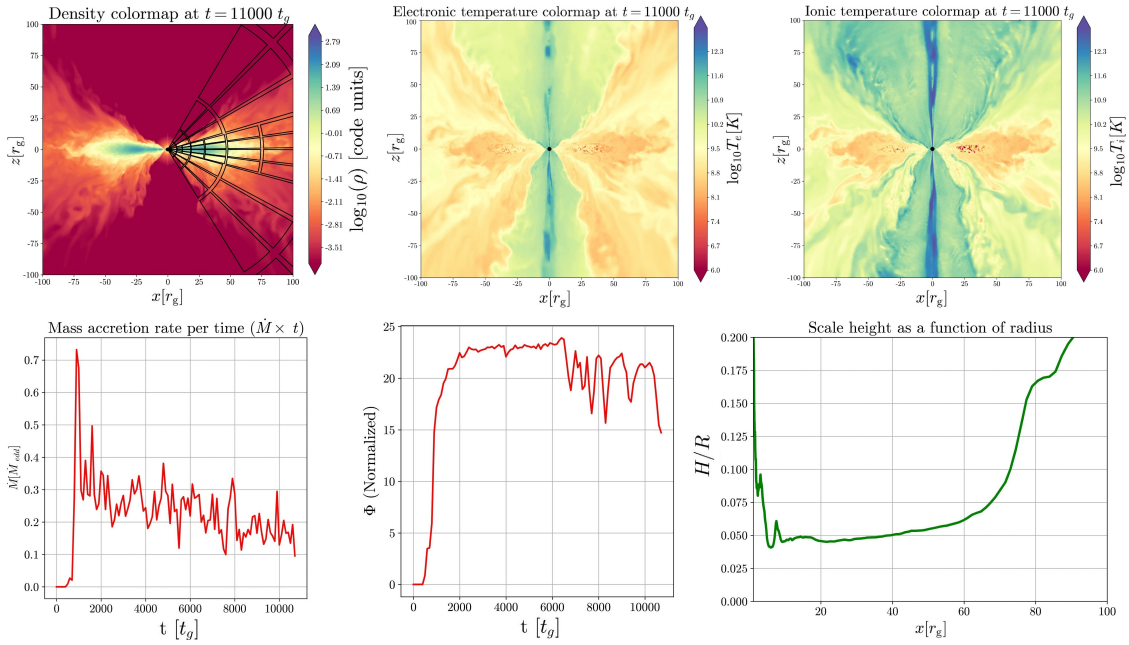

Figure 6.1: Top: (Left) Density color map in code units, (Middle) Electronic temperature in kelvin, (Right) Ionic temperature in kelvin. Bottom: (Left) Mass accretion rate, (Middle) Magnetic flux normalized, (Right) Density scale height over the radius. The block division in the density color map demonstrates the grid's SMR near the mid-plane. The color maps and the scale height were calculated at $t = 11000 R_g/c$.

allowing a lower floor drops the temperature of the disk significantly. The MRI quality factor clearly shows that this resolution fails to resolve the denser regions of the geometrically thin disk as expected.

Figure 6.3 shows values for each component of the cooling function. In this scenario, we can clearly see the Compton enhancement factor happening primarily in the corona very close to the black hole, as expected.

In Figure 6.4, we do the same comparison as the one in Section 5.1. We get the same results for high resolution, where we get a cooling function with optically thin cooling in the corona and optically thick cooling near the midplane.

Finally, for comparison, we generate cooling color maps using two different cooling tables: one with 33^4 values and another with 100^4 values. The higher-resolution table offers significantly finer detail in the resulting image as seen by Figure 6.5, especially in the area where cooling varies more. However, due to its binary storage format, the file size reaches a maximum of approximately $\sim 700MB$. It's worth noting that each GPU has a limited amount of memory available for texture operations. While we attempted to implement all seven tables with 100^4 values, the $V100$ GPUs on the SUMMIT system were unable to handle the computational load.

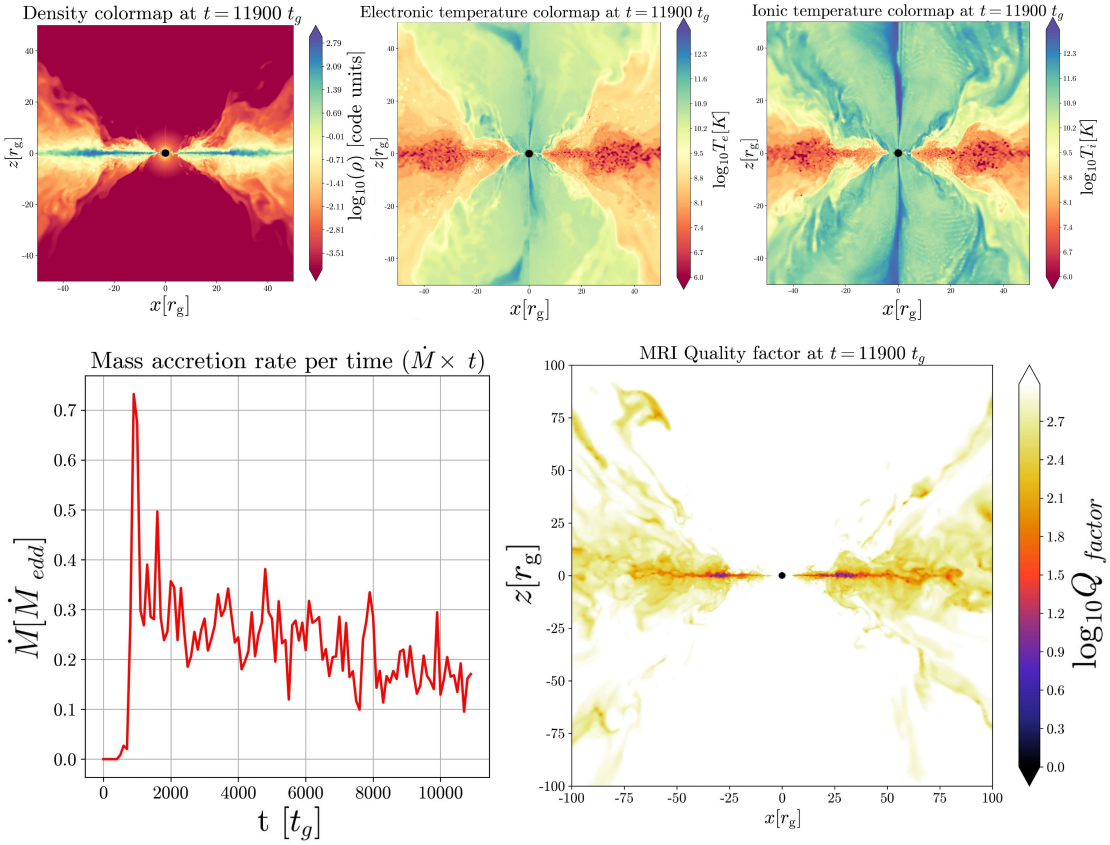

Figure 6.2: Snapshot at $11900 R_g/c$. For this snapshot, we lowed the floor to $(H/R)_{\text{target}} = 0.01$. Top: (Left) Density color map, (Middle) electronic temperature, (Right) ionic temperature. Bottom: (Left) Mass accretion rate and (Right) MRI quality factor.

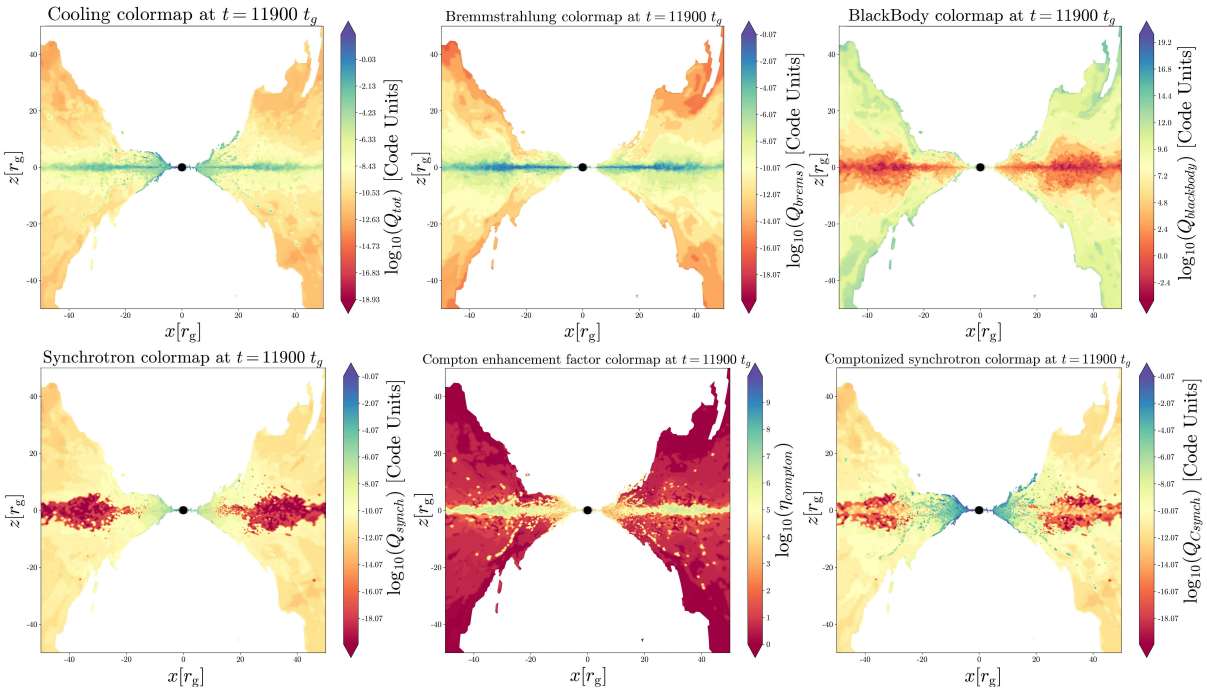

Figure 6.3: Cooling function components snapshot at $11900 R_g/c$. For this snapshot, we lowed the floor to $(H/R)_{\text{target}} = 0.01$. Top: (Left) Bremsstrahlung radiation, (Middle) blackbody radiation, (Right) total cooling values. Bottom: (Left) Synchrotron radiation, (Middle) Compton enhancement parameter and (Right) Comptonized synchrotron radiation.

During the phase with $(H/R)_{\text{targeted}} = 0.1$, we did not observe the formation of a hybrid disk. Previous studies on thin disks suggest that the ratio (H/R) typically falls within the range of 10^{-3} to 0.01 (Piran et al., 2015; Frank et al., 2002). Additionally, our simulation duration was relatively shorter, running for approximately $10^3 t_g$ compared to the typical $10^4 t_g$ used in previous simulations of truncated accretion disks with the `H-AMR` code. In the future, we plan to run a simulation with a lower floor to explore the potential truncation of the disk.

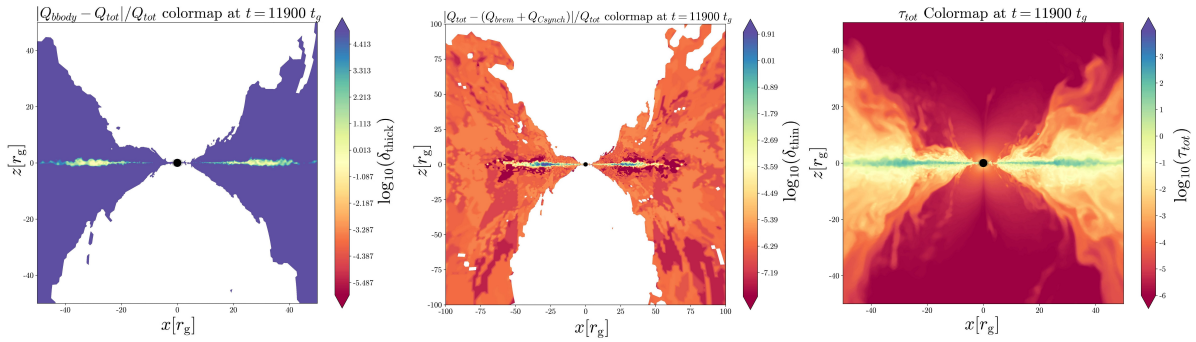

Figure 6.4: Snapshot at $11900 R_g/c$. For this snapshot, we lowed the floor to $(H/R)_{\text{target}} = 0.01$. Parameter δ_{thick} and δ_{thin} defined in equations (5.2) and (5.3).

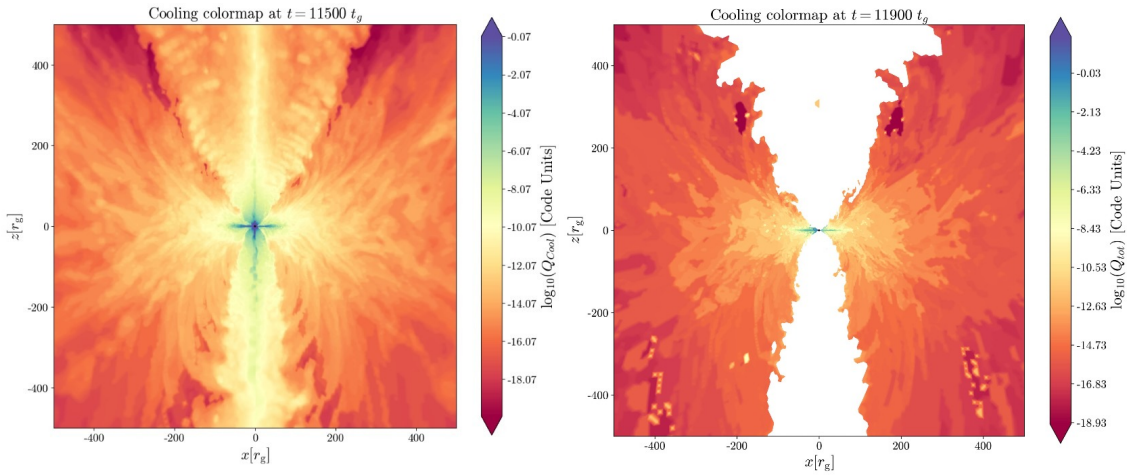

Figure 6.5: (Left) Cooling values color map using 100^4 table. (Right) Cooling values color map using 33^4 table.

Chapter 7

Conclusion

In conclusion, we successfully implemented a realistic cooling prescription into `H-AMR`, a state-of-the-art 3D GRMHD code, utilizing texture memory. The implementation of this cooling prescription opens up new possibilities for future simulations that need radiative cooling effects without the need of expensive computational resources.

Although this work has made progress in addressing the cooling aspect, further advancements are still necessary. One important area to focus on is the implementation of radiation pressure. This is particularly crucial for simulating scenarios with high accretion rates, where radiation pressure dominates over magnetic and thermal pressure. Building on the work by Esin et al. (1996), we can calculate radiation pressure as follows:

$$P_{\rm rad} = \frac{Q_{\rm tot} H}{2c} \left(\tau + \frac{2}{\sqrt{3}} \right). \tag{7.1}$$

while in the radiation-transport code, it is represented as a tensor within radiation stress-energy tensor calculated as $P^{ij} = \int \hat{I}_\nu d\nu d\Omega N^i N^j$. This quantity can be treated as a local value, assuming the approximations made for the scale height. Consequently, it wouldn't require any additional calculations, and its impact on performance should be negligible. Nonetheless, special attention must be paid to accurately implement and handle pressure-related effects to ensure the reliability of the simulations.

Another area we would like to explore is the implementation of Comptonized bremsstrahlung. While previous studies (Yoon et al., 2020; Fragile and Meier, 2009) have suggested that Comptonized bremsstrahlung may not be significant, it's possible that this conclusion was influenced by their specific simulation setup and regime. Therefore, it is worth investigating this phenomenon further in our research.

Our future objective is to replicate the simulation conducted in Liska et al. (2022), which utilized a highly realistic but computationally expensive approach by incorporating

radiative transport with M1 closure (McKinney et al., 2014). By attempting to achieve comparable results using a more cost-effective method, we aim to demonstrate the efficacy and efficiency of our approach.

7.1 Future perspectives

While the main focus of the dissertation was to examine the radiative cooling prescription in the context of simulating black hole x-ray binaries, the implementation and testing of our modifications to the `H-AMR` framework proved to be more time-consuming than initially anticipated. However, the testing phase was crucial to ensure the robustness and reliability of our tool. To share the methodologies developed in this study, we plan to release a computational methods paper.

Additionally, we encountered delays due to the resolution issue specific to general relativistic magnetohydrodynamics (GRMHD) codes. GRMHD simulations require adequate resolution to accurately capture the behavior of the magnetorotational instability (MRI), a challenge that our research group had not encountered in previous α-prescription hydrodynamical simulations (Nemmen et al., 2023). To deal with this, we needed access to supercomputers such as LNCC's (Laboratório Nacional de Computação Científica) Santos Dumont and OLCF's SUMMIT. Using the tool developed in this work, we intend to simulate different spectral states of black hole X-ray binaries in the near future.

Bibliography

Anninos P., Fragile P. C., Salmonson J. D., Cosmos++: Relativistic Magnetohydrodynamics on Unstructured Grids with Local Adaptive Refinement, ApJ, 2005, vol. 635, p. 723

Balbus S. A., Hawley J. F., A Powerful Local Shear Instability in Weakly Magnetized Disks. I. Linear Analysis, ApJ, 1991, vol. 376, p. 214

Baumgardt H., He C., Sweet S. M., Drinkwater M., Sollima A., Hurley J., Usher C., Kamann S., Dalgleish H., Dreizler S., Husser T.-O., No evidence for intermediate-mass black holes in the globular clusters ω Cen and NGC 6624, Monthly Notices of the Royal Astronomical Society, 2019, vol. 488, p. 5340

Bekefi G., Radiation Processes in Plasmas. Radiation Processes in Plasmas, Wiley, 1966

Belmont R., Numerical computation of isotropic Compton scattering, A&A, 2009, vol. 506, p. 589

Belmont R., Malzac J., Marcowith A., Simulating radiation and kinetic processes in relativistic plasmas, A&A, 2008, vol. 491, p. 617

Chandrasekhar S., Hydrodynamic and Hydromagnetic Stability. International series of monographs on physics, Clarendon Press, 1961

Contopoulos I., Gabuzda D., Kylafis N., The Formation and Disruption of Black Hole Jets. Springer Cham, 2015

Courant R., Friedrichs K., Lewy H., Über die partiellen Differenzengleichungen der mathematischen Physik, Mathematische Annalen, 1928, vol. 100, p. 32

Dermer C. D., Liang E. P., Canfield E., Luminosity Enhancement Factor for Thermal Comptonization and the Electron Energy Balance, ApJ, 1991, vol. 369, p. 410

Dexter J., Scepi N., Begelman M. C., Radiation GRMHD Simulations of the Hard State of Black Hole X-Ray Binaries and the Collapse of a Hot Accretion Flow, The Astrophysical Journal Letters, 2021, vol. 919, p. L20

Dihingia I. K., Mizuno Y., Fromm C. M., Rezzolla L., Temperature properties in magnetized and radiatively cooled two-temperature accretion flows on to a black hole, Monthly Notices of the Royal Astronomical Society, 2022, vol. 518, p. 405

Einfeldt B., On Godunov-Type Methods for Gas Dynamics, SIAM Journal on Numerical Analysis, 1988, vol. 25, p. 294

Esin A. A., McClintock J. E., Narayan R., Advection-Dominated Accretion and the Spectral States of Black Hole X-Ray Binaries: Application to Nova Muscae 1991, ApJ, 1997, vol. 489, p. 865

Esin A. A., Narayan R., Ostriker E., Yi I., Hot One-Temperature Accretion Flows around Black Holes, apj, 1996, vol. 465, p. 312

Fender R., Belloni T., Stellar-Mass Black Holes and Ultraluminous X-ray Sources, Science, 2012, vol. 337, p. 540

Ferreira, J. Theory of magnetized accretion discs driving jets, EAS Publications Series, 2002, vol. 3, p. 229

Ferreira, J. Petrucci, P.-O. Henri, G. Saugé, L. Pelletier, G. A unified accretion-ejection paradigm for black hole X-ray binaries - I. The dynamical constituents, A&A, 2006, vol. 447, p. 813

Fishbone L. G., Moncrief V., Relativistic fluid disks in orbit around Kerr black holes., ApJ, 1976, vol. 207, p. 962

Fragile P. C., Meier D. L., GENERAL RELATIVISTIC MAGNETOHYDRODYNAMIC SIMULATIONS OF THE HARD STATE AS A MAGNETICALLY DOMINATED ACCRETION FLOW, The Astrophysical Journal, 2009, vol. 693, p. 771

Fryer C. L., Mass Limits For Black Hole Formation, The Astrophysical Journal, 1999, vol. 522, p. 413

Gammie C. F., McKinney J. C., Tóth G., HARM: A Numerical Scheme for General Relativistic Magnetohydrodynamics, ApJ, 2003, vol. 589, p. 444

Gottlieb O., Moseley S., Ramirez-Aguilar T., Murguia-Berthier A., Liska M., Tchekhovskoy A., On the Jet–Ejecta Interaction in 3D GRMHD Simulations of a Binary Neutron Star Merger Aftermath, The Astrophysical Journal Letters, 2022, vol. 933, p. L2

Griffiths D. J., Introduction to electrodynamics; 3rd ed.. Prentice-Hall Upper Saddle River, NJ, 1999

Hameury J.-M., Lasota J.-P., McClintock J. E., Narayan R., Advection-dominated Flows around Black Holes and the X-Ray Delay in the Outburst of GRO J1655–40, The Astrophysical Journal, 1997, vol. 489, p. 234

Haug E., Bremsstrahlung and pair production in the field of free electrons., Zeitschrift Naturforschung Teil A, 1975, vol. 30, p. 1099

Hawley J. F., Guan X., Krolik J. H., ASSESSING QUANTITATIVE RESULTS IN ACCRETION SIMULATIONS: FROM LOCAL TO GLOBAL, The Astrophysical Journal, 2011, vol. 738, p. 84

Heitler W., The Quantum Theory of Radiation. Dover Books on Physics Series, Dover Publications, 1984

Jacquemin-Ide J., , 2021 Magnetically driven jets and winds launched from turbulent accretion disks PhD thesis, Université Grenoble Alpes

Kaaz N., Liska M. T. P., Jacquemin-Ide J., Andalman Z. L., Musoke G., Tchekhovskoy A., Porth O., , 2022 Nozzle Shocks, Disk Tearing and Streamers Drive Rapid Accretion in 3D GRMHD Simulations of Warped Thin Disks

Kormendy J., Richstone D., Inward Bound—The Search For Supermassive Black Holes In Galactic Nuclei, ARA&A, 1995, vol. 33, p. 581

Levermore C. D., Relating Eddington factors to flux limiters., J. Quant. Spec. Radiat. Transf., 1984, vol. 31, p. 149

Lightman A. P., Band D. L., Relativistic thermal plasmas - Radiation mechanisms, apj, 1981, vol. 251, p. 713

Liska M., Tchekhovskoy A., Ingram A., van der Klis M., Bardeen–Petterson alignment, jets, and magnetic truncation in GRMHD simulations of tilted thin accretion discs, Monthly Notices of the Royal Astronomical Society, 2019, vol. 487, p. 550

Liska M. T. P., Chatterjee K., Issa D., Yoon D., Kaaz N., Tchekhovskoy A., van Eijnatten D., Musoke G., Hesp C., Rohoza V., Markoff S., Ingram A., van der Klis M., H-AMR: A New GPU-accelerated GRMHD Code for Exascale Computing with 3D Adaptive Mesh Refinement and Local Adaptive Time Stepping, The Astrophysical Journal Supplement Series, 2022, vol. 263, p. 26

Liska M. T. P., Musoke G., Tchekhovskoy A., Porth O., Beloborodov A. M., Formation of Magnetically Truncated Accretion Disks in 3D Radiation-transport Two-temperature GRMHD Simulations, The Astrophysical Journal Letters, 2022, vol. 935, p. L1

Lowell B., Jacquemin-Ide J., Tchekhovskoy A., Duncan A., , 2023 Rapid Black Hole Spin-down by Thick Magnetically Arrested Disks

McKinney J. C., Tchekhovskoy A., Sadowski A., Narayan R., Three-dimensional general relativistic radiation magnetohydrodynamical simulation of super-Eddington accretion, using a new code harmrad with M1 closure, Monthly Notices of the Royal Astronomical Society, 2014, vol. 441, p. 3177

Mahadevan R., Narayan R., Yi I., Harmony in Electrons: Cyclotron and Synchrotron Emission by Thermal Electrons in a Magnetic Field, ApJ, 1996, vol. 465, p. 327

Marcel G., Ferreira J., Clavel M., Petrucci P. O., Malzac J., Corbel S., Rodriguez J., Belmont R., Coriat M., Henri G., Cangemi F., A unified accretion-ejection paradigm for black hole X-ray binaries. IV. Replication of the 2010-2011 activity cycle of GX 339-4, A&A, 2019, vol. 626, p. A115

Marcel G., Ferreira J., Petrucci P.-O., Henri G., Belmont R., Clavel M., Malzac J., Coriat M., Corbel S., Rodriguez J., Loh A., Chakravorty S., Drappeau S., A unified accretion-ejection paradigm for black hole X-ray binaries, Astronomy & Astrophysics, 2018, vol. 615, p. A57

Mihalas D., Mihalas B. W., Foundations of radiation hydrodynamics. Dover, 1984

Moshovos A., , 2009 Textures Slides

Murphy N., , 2014 The Magnetorotational Instability https://lweb.cfa.harvard.edu/~namurphy/Lectures/Ay253_07_MHDinstabilities2.pdf

Narayan R., McClintock J. E., Yi I., A New Model for Black Hole Soft X-Ray Transients in Quiescence, The Astrophysical Journal, 1996, vol. 457, p. 821

Narayan R., Quataert E., Black Hole Accretion, Science, 2005, vol. 307, p. 77

Narayan R., Yi I., Advection-dominated Accretion: A Self-similar Solution, ApJ, 1994, vol. 428, p. L13

Narayan R., Yi I., Advection-dominated Accretion: Underfed Black Holes and Neutron Stars, ApJ, 1995, vol. 452, p. 710

Nemmen R., Vemado A., Almeida I., Garcia J., Motta P., Emergence of hot corona and truncated disk in simulations of accreting stellar mass black holes, 2023

Noble S. C., Krolik J. H., Hawley J. F., Direct Calculation of the Radiative Efficiency of an Accretion Disk Around a Black Hole, ApJ, 2009, vol. 692, p. 411

Novikov I. D., Thorne K. S., Astrophysics of black holes.. In Black Holes (Les Astres Occlus) , 1973, p. 343

Pacholczyk A. G., Radio astrophysics. Nonthermal processes in galactic and extragalactic sources. W. H. Freeman, 1970

Pareschi L., Russo G., Implicit–Explicit Runge–Kutta Schemes and Applications to Hyperbolic Systems with Relaxation, Journal of Scientific Computing, 2005, vol. 25, p. 129

Porth O., Olivares H., Mizuno Y., Younsi Z., Rezzolla L., Moscibrodzka M., Falcke H., Kramer M., The black hole accretion code, Computational Astrophysics and Cosmology, 2017, vol. 4, p. 1

Schneider E. E., Robertson B. E., CHOLLA: A NEW MASSIVELY PARALLEL HYDRODYNAMICS CODE FOR ASTROPHYSICAL SIMULATION, The Astrophysical Journal Supplement Series, 2015, vol. 217, p. 24

Schneider E. E., Robertson B. E., HYDRODYNAMICAL COUPLING OF MASS AND MOMENTUM IN MULTIPHASE GALACTIC WINDS, The Astrophysical Journal, 2017, vol. 834, p. 144

Schutz B., A First Course in General Relativity 2 edn. Cambridge University Press, 2009, 281 p.

Shakura N. I., Sunyaev R. A., Black holes in binary systems. Observational appearance., A&A, 1973, vol. 24, p. 337

Stepney S., Guilbert P. W., Numerical fits to important rates in high temperature astrophysical plasmas., MNRAS, 1983, vol. 204, p. 1269

Stone J. M., Tomida K., White C. J., Felker K. G., The Athena++ Adaptive Mesh Refinement Framework: Design and Magnetohydrodynamic Solvers, The Astrophysical Journal Supplement Series, 2020, vol. 249, p. 4

Svensson R., Electron-Positron Pair Equilibria in Relativistic Plasmas, ApJ, 1982, vol. 258, p. 335

Svensson R., Steady mildly relativistic thermal plasmas - Processes and properties, MNRAS, 1984, vol. 209, p. 175

Sądowski A., Wielgus M., Narayan R., Abarca D., McKinney J. C., Chael A., Radiative, two-temperature simulations of low-luminosity black hole accretion flows in general relativity, Monthly Notices of the Royal Astronomical Society, 2016, vol. 466, p. 705

Takahashi H. R., Ohsuga K., Kawashima T., Sekiguchi Y., Formation of Overheated Regions and Truncated Disks around Black Holes: Three-dimensional General Relativistic Radiation-magnetohydrodynamics Simulations, ApJ, 2016, vol. 826, p. 23

Tchekhovskoy A., , 2019 HARMPI: 3D massively parallel general relativictic MHD code Astrophysics Source Code Library, record ascl:1912.014

Tchekhovskoy A., McKinney J. C., Prograde and retrograde black holes: whose jet is more powerful?, Monthly Notices of the Royal Astronomical Society: Letters, 2012, vol. 423, p. L55

Wu M.-C., Xie F.-G., Yuan Y.-F., Gan Z., Hot accretion flow with radiative cooling: state transitions in black hole X-ray binaries, Monthly Notices of the Royal Astronomical Society, 2016, vol. 459, p. 1543

Yoon D., Chatterjee K., Markoff S. B., van Eijnatten D., Younsi Z., Liska M., Tchekhovskoy A., Spectral and imaging properties of Sgr A* from high-resolution 3D GRMHD simulations with radiative cooling, MNRAS, 2020, vol. 499, p. 3178

Appendix

Appendix A

Computational Infrastructure

A.1 Codes

Throughout this work, we used the code `H-AMR` which is not publicly available yet. Besides using `H-AMR`, we developed an OpenMP+MPI code to handle the creation of the cooling table and to do some tests to ensure the results we were getting were correct. This code was written in `C`. The table creation and tests are publicly available in Github repository, which has a `README` file describing how to generate the table and run the tests.

The implementation of our texture memory follows a straightforward process. At the start of the code, the cooling tables are read from a binary file and loaded into a 3D texture object. During this phase, status messages "Reading cooling table from the binary file" and "Loading cooling into texture object" are displayed to indicate progress.

When the code starts the calculations, a function called `source_texture_cooling` is invoked for each cell. Within this function, equation 4.5 is computed:

$$G^{\mu} = -Q_{\rm total} u^{\mu}, \tag{A.1}$$

where $Q_{\rm total}$ is calculated by calling the `cooling_function`.

The activation of cooling involves setting six switches in the `config.h` file. These switches control various aspects:

1. Activating the cooling process.
2. Activating the usage of cooling components.
3. Setting the usage of the floor by setting a targeted scale height ratio $(H/R)_{\rm targeted}$
4. Specifying the floor's targeted ratio of scale height to radius $(H/R)_{\rm targeted}$.

5. Using a local thermal scale height, as done in Fragile and Meier (2009).

6. Setting the maximum percentage of internal energy that cooling can extract from each cell.

This configuration allows us to efficiently manage the cooling process and use it according to specific requirements.

A.2 Clusters

A.2.1 SUMMIT

Summit is an IBM system[1] located at the Oak Ridge Leadership Computing Facility. Each of the approximately 4600 compute nodes on Summit contains two IBM POWER9 processors and six NVIDIA Tesla V100 accelerators and provides a theoretical double-precision capability of approximately 40 TF. Each POWER9 processor is connected via dual NVLINK bricks, each capable of a 25GB/s transfer rate in each direction. For more information, refer to SUMMIT's user guide.

A.2.2 Sdumont

We made use of the sequana_gpu_shared nodes, which consist of 94 computational nodes called Bull Sequana X1120 (GPU). Each node contains two Intel Xeon Skylake 6252 processors, providing a total of 48 cores (24 cores per CPU). These nodes are equipped with 384GB of RAM to handle memory-intensive tasks. Additionally, each node is enhanced with 4 NVIDIA Volta V100 GPUs, which help speed up specialized tasks that require intense parallel processing.

[1] IBM stands for International Business Machines Corporation. It is a multinational technology company that is well-known for manufacturing and selling computer hardware, software, and various other technology-related products and services.